# Topic Models with Topic Ordering Regularities for Topic Segmentation

Lan Du
Department of computing
Macquarie University
Sydney, NSW 2109, Australia
Email: lan.du@mq.edu.au

John K Pate
Department of Computing
Macquarie University
Sydney, NSW 2109, Australia
Email: john.pate@mq.edu.au

Mark Johnson
Department of Computing
Macquarie University
Sydney, NSW 2109, Australia
Email: mark.johnson@mq.edu.au

***Abstract*—Documents from the same domain usually discuss similar topics in a similar order. In this paper we present new ordering-based topic models that use generalised Mallows models to capture this regularity to constrain topic assignments. Specifically, these new models assume that there is a canonical topic ordering shared amongst documents from the same domain, and each document-specific topic ordering is allowed to vary from the canonical topic ordering. Instead of full orderings over a set of all possible topics covered by a domain, we make use of top-$t$ orderings via a multistage ranking process. We show how to reformulate the new models so that a point-wise sampling algorithm from the Bayesian word segmentation literature can be used for posterior inference. Experimental results on several document collections with different properties show that our model performs much better than the other topic ordering-based models, and competitively with other state-of-the-art topic segmentation models.**

***Index Terms*—topic model; topic segmentation; top-$t$ ordering;**

## I. Introduction

Probabilistic topics models, such as Latent Dirichlet allocation (LDA) [1], have shown remarkable success in modelling the semantic aspects of natural text. These models usually assume that documents are *bags-of-words*, ignoring any ordering regularities, e.g., word order and topic structure. Recently, a great amount of effort from Machine Learning, NLP and Data Mining communities has focused on developing more comprehensive topic models that can go beyond the bag-of-words assumption and can handle different features of natural text. They have typically been created by taking some standard topic model, like LDA, as a fundamental building block, and then incorporating additional information from the text either by introducing new model variables or modifying the priors. This information includes, for example, authorship [2], document links [3], [4], temporal information [5], and tags or labels [6]. In this paper we are interested in models of topic structure, which assume that a document consists of a sequence of topically-coherent segments, such as a section or a chapter. These models maintain a bag-of-words assumption within each segment, but can learn structured relationships between segments. In particular, we are interested in learning typical topic orderings that are especially robust among documents from the same domain.

Consider the three English Wikipedia articles about Bejing, Shanghai and Guangzhou, the three biggest cities in China. These articles describe the three cities in terms of *History*, *Geography*, *Politics*, *Economy*, *Culture*, *Education*, *Transportation*, etc. Those topics are not discussed in an arbitrary order, but are instead addressed in a largely shared order. For instance, *History* and *Geography* usually appear first; *Geography* is discussed after *History* and is often followed by *Politics*; *Economy* is always discussed before both *Education* and *Transportation*; and *Media* and *Sports* usually appear at the end. We noticed that each topic is discussed in only one section, and paragraphs in a section share the same topic. In addition, there is some variation in the topic ordering, such as the reversal in the *Education* and *Transportation* sections in the Shanghai and Guangzhou articles, and the different positions of the *Culture* section in the three articles.

Models that capture these kinds of domain-specific topic ordering regularities may be more accurate. Inspired by the idea of the Global Model (GM) [7], we introduce an ordering-based topic model by incorporating a multistage ranking model, known as the generalised Mallows model (GMM) [8], [9], in a simple generative process. Briefly, our model represents the topic structure of each document as an ordering of the latent topics, and generates from each topic, in order, a contiguous sequence of paragraphs or sentences. All topic orderings are distributed according to the GMM. As in the GM, this simple generative process guarantees that each topic is discussed in at most one section or location of a document.

Moreover, we observed that both the number and the topics discussed in each document can vary across different documents from even the same domain. This variation means that a document-specific topic ordering should be a partial ordering of a subset of the domain's topics, rather than a total ordering of all of the topics. In order to capture this variation, we make use of partial orderings, also known as top-$t$ orderings, where $t$ is the number of topics addressed in a particular document [10], [11]. The use of top-$t$ topic orderings implies that the most common and important topics are usually chosen at the early stages of a ranking process, i.e., the first $t$ stages, and the topics selected in later stages are more likely to be dropped than earlier ones. This partialness not only differentiates our models from the GM but also introduces the challenge of

learning a topic model that exploits top-$t$ orderings. We show how this problem can be overcome by adapting the idea of multistage ranking models [9], [12] to generate partially-ordered topics. The experimental results demonstrate that our models have better topic segmentation performance than other approaches. The rest of the paper is organised as follows. Section II briefly reviews related work. We then present our ordering-based topic models in Section III, and elaborate the derivation of a point-wise sampling algorithm in Section IV. Experimental results are reported in Section V. Section VI concludes the paper with future work.

## II. Related Work

We are particularly interested in generative models that can incorporate information about topic structure, and how to apply them to, for example, the topic segmentation task. In this section we briefly review some related work.

**Models of Topic Structure**: Most existing models of topic structure have assumed an HMM-structure for topic sequences. They model the transition between topics or topic distributions with a first order Markov chain. These models include the aspect model [13], the content models [14], [15], the hidden topic Markov model [16], sequential LDA [17], the structured topic model [18], etc. Without using a Markov chain, the segmented topic model [19] and the adaptive topic model [20] use a simple tree structure and a DAG structure respectively to capture the topical structure. However, those models are not appropriate for modelling the topic orderings that we are interested in, where there is a canonical ordering and any given topic appears at most once.

**Models of Orderings**: Recently the statistical ranking methods, such as the GMM and the Luce model (see [21] for more details), have gained increasing usage in topic modelling. The topical ranking of a document given by the Topical Pagerank algorithm [22] has been used to guide topic assignments [23]. A loss function based on the Plackett-Luce model has been used together with PLSI in multi-document summarisation [24]. The most similar work to ours is the GM [7] and the model introduced by Frermann et al. [25]. Our model may be regarded as a novel variant of the GM. The main differences reside in the use of top-$t$ topic orderings and the corresponding posterior inference algorithm. These differences are also the two main contributions of this paper.

**Models of Topic Segmentation**: Our work is also related to topic segmentation. The task of topic segmentation is to identify the topic changes in unannotated text, and split the text into a sequence of semantically coherent segments, e.g., sections. Recently, various extensions of topic models, such as Bayesseg [26], PLDA [27] and STSM [28], have been used for topic segmentation. We will compare our ordering-based models with these models in Section V.

## III. Topic Models with Top-$t$ Orderings

In this section we present a new probabilistic generative process, named a Topic Model of Top-$t$ Orderings (TMTO), that uses top-$t$ orderings to guide the topic assignments and topic segmentations. We start by discussing the GMM over top-$t$ orderings, and then describe our models in detail.

Let $\mathbf{\Pi}$ be the set of all $K!$ possible permutations over $K$ topics. The GMM defines a probability distribution over $\mathbf{\Pi}$. It is parameterised by dispersion parameters $\boldsymbol{\theta}$ (an $n-1$ dimensional vector of real values) and a canonical ordering $\boldsymbol{\sigma}$. The probability of any ordering $\boldsymbol{\pi}$ decreases exponentially with increasing distance from the canonical ordering, which is formally defined as $GMM(\boldsymbol{\pi};\boldsymbol{\sigma},\boldsymbol{\theta}) = \prod_{i=1}^{K-1} \frac{e^{-\theta_i s_i(\boldsymbol{\pi}|\boldsymbol{\sigma})}}{\psi_i(\theta_i)}$, where $\psi_i(\theta_i)$ is a partition function given by $\psi_i(\theta_i) = \sum_{s_i(\boldsymbol{\pi}|\boldsymbol{\sigma})=0}^{n-i} e^{-\theta_i s_i(\boldsymbol{\pi}|\boldsymbol{\sigma})} = \frac{1-e^{-(K-i+1)\theta_i}}{1-e^{-\theta_i}}$, and $s_i(\boldsymbol{\pi}|\boldsymbol{\sigma}) = \sum_{l>i} 1_{\boldsymbol{\sigma}(\boldsymbol{\pi}^{-1}(i))>\boldsymbol{\sigma}(\boldsymbol{\pi}^{-1}(l))}$. The dispersion parameters control how concentrated the probability mass is around $\boldsymbol{\sigma}$. One elegant computational property is that the probability mass function factors into a product of independent univariate exponential models, which makes it a special case of a multistage ranking model [9].

In a ranking process, a judge might report only his/her top $t < K$ preferences, which will result in a partial ordering with $t$ selected items (i.e., topics in our models), also known as a top-$t$ ordering. Similarly, a given document might discuss a subset of topics covered by its domain. A top-$t$ ordering, $\bar{\boldsymbol{\pi}} = (\boldsymbol{\pi}^{-1}(1), \boldsymbol{\pi}^{-1}(2), \ldots, \boldsymbol{\pi}^{-1}(t))$, indeed represents a set $\mathbf{\Omega}$ of full orderings that start with $\bar{\boldsymbol{\pi}}$. Moreover, in a top-$t$ ordering, the first $t$ values in $\boldsymbol{s}$, i.e., $s_1, \ldots, s_t$, are fixed, and the remaining $s_{t+1}, \ldots, s_{K-1}$ are undetermined. Fligner and Verducci [8] have shown that the probability mass function still applies to top-$t$ orderings. We will use the GMM$^t$ to denote the GMM over top-$t$ orderings in the rest of the paper.

The basic idea of our models can be summarised in four points: 1) there is a canonical topic ordering $\boldsymbol{\sigma}$ shared amongst documents from the same domain, 2) each document-specific top-$t$ ordering $\bar{\boldsymbol{\pi}}_d$ is generated from the GMM$^t$, and is allowed to vary from $\boldsymbol{\sigma}$, 3) each topic is discussed at most once by one part (e.g., section) of a document, and 4) the most common and important topics are assumed to be selected first in the ranking process. The full probabilistic generative process is defines as follows. For each document $d$,

1) **Generate the Number of Sections**: The number of sections $t_d$ is equal to the number of topics discussed by document $d$, and will also be the length of the top-$t$ ordering. We assume $t_d$ is distributed according to a Geometric distribution over integers, where the parameter $\epsilon$ is generated from a symmetric Beta$(\eta/2)$ prior. Let $\epsilon$ be the probability of generating the end of document. The probability of $t_d$ is $(1-\epsilon)^{t_d-1}\epsilon$, where $\epsilon \sim \text{Beta}(\frac{\eta}{2})$. One can also use a Poisson distribution with a Gamma prior, i.e., $p(t_d|\lambda) \propto \frac{\lambda^{t_d}}{t_d!}$.
2) **Generate a Top-$t$ Topic Ordering**: To generate a top-$t$ topic ordering for a document, we adapt the idea of multistage ranking models [9], [12] from the learning-to-rank literature. In the multistage ranking models, the ranking process is split into $K$-1 stages, where $K$ is the number of items to be ranked. At a given stage $k$, one item is selected from the set of remaining items, which excludes all the items selected by the previous $k-1$ stages. In our

case, we can view the process of generating topics for a sequence of $t_d$ ($t_d \leq K$) semantically coherent sections as a multistage ranking process, where the ordered sections correspond to a sequence of selection stages in ranking. In this process, each document selects the first topic to be discussed by the first section in the first stage. This topic is the topic that is usually discussed first according to the canonical ordering of the document's domain. For example, the *History* topic might be chosen in the first stage for Wikipedia articles about cities. The second topic is then selected from the remaining topics. For example, *Geography* might be chosen. This multistage selection process iterates until $t_d$ topics are selected. Using the factorized form of the GMM, one can compute the probability of choosing a topic in stage $j$ ($j \leq t_d$), $p\big(s_j(\bar{\boldsymbol{\pi}}_d|\boldsymbol{\sigma})\big) = \frac{e^{-\theta_j s_j(\bar{\boldsymbol{\pi}}_d|\boldsymbol{\sigma})}}{\psi_j(\theta_j)}$. The ideas of the GMM$^t$ and the multistage ranking process capture the four intuitions about topic ordering described at the beginning of this section. As in the GM, we assume that $\boldsymbol{\sigma}$ is always the identity ordering, $(1, 2, \ldots, K)$, because topics themselves are latent variables to be learnt from data.

3) **Generate Topic Spans** The topic span is defined as a sequence of paragraphs that have the same topic. Given a top-$t$ ordering $\bar{\boldsymbol{\pi}}_d$ for document $d$, for each topic $k$ in $\bar{\boldsymbol{\pi}}_d$ we generate a topic span $l_{d,k}$ from a topic-specific Poisson distribution, Poisson($\lambda_k$). Here, we assume that different topics have different distributions over topic spans, and all the spans are Poisson distributed, i.e., $p(l_{d,k}|\lambda_k) = \frac{\lambda_k^{l_{d,k}}}{l_{d,k}!}e^{-\lambda_k}$. For example, if $\bar{\boldsymbol{\pi}}_d = (3, 1, 5)$, $l_{d,3} = 3$, $l_{d,1} = 3$ and $l_{d,5} = 3$, the generated document will have 9 paragraphs. The span of topic 3 will be from paragraph 1 to paragraph 3, the span of topic 1 from 4 to 6, and that of topic 4 from 7 to 9. It is worth pointing out that this generating process also gives a topic segmentation of $d$, which can be presented as a vector of paragraph topic assignments, $(3, 3, 3, 1, 1, 1, 5, 5, 5)$. We also used a Geometric distribution in our experiments.
4) **Generate Words** The final step in our generative process is to generate words from topics. We assume that words in paragraphs within a topic span are generated from the same topic with a Dirichlet-Multinomial model (DMM), which is the standard model for topic-word relations in topic modelling. The DMM assigns the following probability to a set of words $\boldsymbol{w}$ assigned to topic $k$: $p(\boldsymbol{w}|\boldsymbol{\beta}) = \frac{\Gamma(\sum_{v=1}^{V}\beta_v)}{\Gamma(\sum_{v=1}^{V}n_{k,v}+\beta_v)}\prod_{v=1}^{V}\frac{\Gamma(\beta_v)}{\Gamma(n_{k,v}+\beta_v)}$, where $\boldsymbol{\beta}$ is symmetric Dirichlet Prior, $\boldsymbol{\phi}_k$ the word distribution of topic $k$, $n_{k,v}$ the count of words $v$ assigned to topic $k$.

Compared with the GM, the difference resides in how the topics discussed by each document are generated. Chen et al. [7] use a bag of topics model, which is a Dirichlet-Multinomial distribution over topics. All the documents share the same Dirichlet-Multinomial. The topic ordering for a document is generated from a full topic ordering by simply ignoring unaddressed topics. In contrast, our TMTO model generates topics from a multistage ranking model over top-$t$ orderings. The number of sections, generated from either a Geometric or Poisson distribution, is viewed as the number of stages in the ranking process. This multistage ranking process models each document's selectional preference of topics. Section V shows that capturing this preference improves text analysis tasks.

## IV. Posterior Inference

Inspired by the boundary sampling algorithm used in Bayesian segmentation [28], [29], we developed a point-wise sampling algorithm that jointly samples top-$t$ orderings, topic assignments and topic segmentations from the posterior. However, the generative process of the model depicted in Section III does not explicitly show how to use a point-wise sampler. Here, we reformulate the model so that point-wise sampling becomes straightforward.

Similar to the word boundary indicator variable used in word segmentation, we introduce a topic-and-boundary indicator variable $\rho_{d,i}$ after each paragraph $i$ in document $d$. The value of $\rho_{d,i}$ indicates whether there is a topic (i.e., section) boundary between the $i$-th and $i+1$-th paragraphs, and, if there is a boundary, the identity of the topic on the left of the boundary. Concretely, if there is no boundary after the i-th paragraph, then $\rho_{d,i} = 0$. Otherwise, there is a section to the left of the boundary, which consists of a sequence of paragraphs from $j+1, \ldots, i$ where $j = \max\{p|1 \leq p \leq i-1 \wedge \rho_{d,p} \neq 0\}$, and the topic of the section is $\rho_{d,i}$, which takes values in $\{1, \ldots, K\}$. Now we illustrate the point-wise sampling algorithm using toy examples with $\mathbb{K} = \{1, 2, \ldots, 6\}$. Let the sampler start with the following state: $\bar{\boldsymbol{\pi}}_d = (3, 1, 5)$ and $\boldsymbol{\rho}_d = (0, 0, 3, 0, 0, 1, 0, 0, 5)$, from which one can induce that $t_d = 3$, $l_{d,3} = 3$, $l_{d,1} = 3$ and $l_{d,5} = 3$. There are three sections that are denoted respectively by $\mathbb{S}_1$, $\mathbb{S}_2$ and $\mathbb{S}_3$. If we consider resampling $\rho_{d,3}$ whose current value is 3, we have to consider two hypotheses—putting or not putting a section boundary after the third paragraph, which correspond to $\rho_{d,3} = 0$ and $\rho_{d,3} > 0$ respectively. The two hypotheses will have different settings for $\bar{\boldsymbol{\pi}}_d$, $\boldsymbol{\rho}_d$, $t_d$ and $\boldsymbol{l}_d$, and are specified as follows.

**Not putting a boundary**: This hypothesis corresponds to changing the boundary indicator vector $\boldsymbol{\rho}_d$ to $(0, 0, 0, 0, 0, 1, 0, 0, 5)$, which merges two sections into one, i.e., $\mathbb{S}_0 = \mathbb{S}_1 \cup \mathbb{S}_2$. Instead of merging $\mathbb{S}_1$ into $\mathbb{S}_2$ and sharing the topic of $\mathbb{S}_2$, we sample a new topic for the merged section $\mathbb{S}_0$ by resampling $\rho_{d,6}$. Note that the value of $\rho_{d,6}$ must be in $\mathbb{T} = \{1, 2, 3, 4, 6\}$, because topic 5 already appears elsewhere in the document. The new state with $\rho_{d,6} = k$ is $\boldsymbol{\rho}_d^k = (0, 0, 0, 0, 0, k, 0, 0, 5)$, $\bar{\boldsymbol{\pi}}_d^k = (k, 5)$, $t_d^k = 2$ and $\boldsymbol{l}_d^k = (6, 3)$. Let $\boldsymbol{\mu}$ indicate all parameters and statistics not affected by the boundary that is currently resampled. The probability of a new state $p(\boldsymbol{\rho}_d^k, \bar{\boldsymbol{\pi}}_d^k, t_d^k, \boldsymbol{l}_d^k \mid \boldsymbol{\mu})$ is proportional to

$$\frac{n' + \frac{\eta}{2}}{N_s^- + \eta}\prod_{i=1}^{t_d^k}\frac{e^{-\theta_i s_i(\bar{\boldsymbol{\pi}}_d^k|\boldsymbol{\sigma})}}{\psi_i(\boldsymbol{\theta}_i)}\frac{\lambda_k^{l_{d,k}}}{l_{d,k}!}\frac{\prod_{v\in\mathbb{S}_0}\left(n_{k,v}^- + \beta_v \,|\, 1\right)_{n_v^{\mathbb{S}_0}}}{\left(\sum_v(n_{k,v}^- + \beta_v)\,|\,1\right)_{N^{\mathbb{S}_0}}}. \quad (1)$$

where $N_s^-$ is the total number of sections, $n' = n_{\$}^-$ (the number of document-final sections) if $\mathbb{S}_2$ is the final section of $d$ and $N_s^- - n_{\$}^-$ otherwise, $n_{k,v}^-$ is the number of words assigned to topic $k$, $n_v^{\mathbb{S}_0}$ is the number of words $v$ in $\mathbb{S}_0$, $N^{\mathbb{S}_0} = \sum_v n_v^{\mathbb{S}_0}$ and $[x|y]_N = x(x+y)...(x+(N-1)y)$. All counts with $^-$ exclude counts from the putative segments $\mathbb{S}_1$ and $\mathbb{S}_2$. Now

the probability of not putting a boundary at $\rho_{d,3}$ is $p(\rho_{d,3} = 0) = \sum_{k\in\mathbb{T}} p(\boldsymbol{\rho}_d^k, \bar{\boldsymbol{\pi}}_d^k, t_d^k, \boldsymbol{l}_d^k \mid \boldsymbol{\mu})$.

**Putting a boundary**: If there is boundary after the third paragraph, the segmentation of document $d$ will not change. However, we will resample the topics for both $\mathbb{S}_1$ and $\mathbb{S}_2$, which corresponds to change the values of $\rho_{d,3}$ and $\rho_{d,6}$. Let $\rho_{d,3} = k_1$ and $\rho_{d,6} = k_2$, where $k_1 \neq k_2$ and both are in $\mathbb{T}$. The new state is then $\boldsymbol{\rho}_d^{(k_1,k_2)} = (0, 0, k_1, 0, 0, k_2, 0, 0, 5)$, $\bar{\boldsymbol{\pi}}_d^{(k_1,k_2)} = (k_1, k_2, 5)$, $t_d^{(k_1,k_2)} = 3$ and $\boldsymbol{l}_d^{(k_1,k_2)} = (3, 3, 3)$. Its probability $p(\boldsymbol{\rho}_d^{(k_1,k_2)}, \bar{\boldsymbol{\pi}}_d^{(k_1,k_2)}, t_d^{(k_1,k_2)}, \boldsymbol{l}_d^{(k_1,k_2)} \mid \boldsymbol{\mu})$ is proportional to

$$\frac{N_s^- - n_\$^- + \frac{\eta}{2}}{N_s^- + \eta} \frac{n^* + \frac{\eta}{2}}{N_s^- + 1 + \eta} \prod_{j=1}^{t_d^{(k_1,k_2)}} \frac{e^{-\theta_j s_j(\bar{\boldsymbol{\pi}}_d^{(k_1,k_2)}|\boldsymbol{\sigma})}}{\psi_j(\boldsymbol{\theta}_j)}$$

$$\prod_{i=1,2} \frac{\prod_{v\in\mathbb{S}_i} \left(n_{k_i,v}^- + \beta_v \mid 1\right)_{n_v^{\mathbb{S}_i}}}{\left(\sum_v (n_{k_i,v}^- + \beta_v) \mid 1\right)_{N^{\mathbb{S}_i}}} \frac{\lambda_{k_i}^{l_{d,k_i}}}{l_{d,k_i}!}.$$

where $n^* = n_\$^-$ if $\mathbb{S}_2$ is the final section of $d$ and $N_s^- - n_\$^- + 1$ otherwise. The probability of putting a boundary at $\rho_{d,3}$ is $p(\rho_{d,3} > 0) = \sum_{k_1,k_2} p(\boldsymbol{\rho}_d^{(k_1,k_2)}, \bar{\boldsymbol{\pi}}_d^{(k_1,k_2)}, t_d^{(k_1,k_2)}, \boldsymbol{l}_d^{(k_1,k_2)} \mid \boldsymbol{\mu})$.

## V. Experimental Results

In the experiments we compare our TMTO with the Global Model [7], Bayesseg [26], LDASeg [27] and STSM [28], on either the topic segmentation task or the cross-document alignment task. The empirical evaluation demonstrates that TMTO performs much better than the earlier ordering-based models (i.e., the Global Model) in both tasks, and also achieves state-of-the-art results in topic segmentation as well.

We evaluate these models on two collections of corpora shown in Table I. The first collection contains four corpora whose documents are assumed to exhibit the ordering regularities. Articles in each corpus usually have a similar section ordering, and we consider section boundaries to be gold topic boundaries. See [7] for the details on the generation of these datasets. The second collection contains four of Choi's datasets [30] and *ICSI* meeting transcripts [31], [32], which have been often used in the topic segmentation task. Our modelling assumptions about topic orderings do not apply for the last two kinds of datasets. We used them to study how sensitive our new model is to the absence of ordering structure.

Parameter settings for the models[1] were: 1) **Bayesseg**: we used the configuration included in the source code package, named *dp.config*; 2) **The Global Model (GM)**: We used the settings reported in [7]. Results reported are the average of 10 samples drawn from 10 randomly initialised Markov chain; 3) **LDASeg** and **STSM**: Settings reported in [28] were used, we ran 10 randomly initialised Markov chains with 30,000 iterations. After 25,000 iterations, we drew a sample every 25 iterations from each chain. The marginal boundary probabilities from the total of 2,000 samples were thresholded to give document segmentations. 4) **TMTO**: We tried both a Geometric (**G**) distribution and a Poisson (**P**) distribution on document lengths and topic spans. Different combinations of these two distributions give us the following four models: **G-G**: **G-P**, **P-G** and **P-P**, where the model on document lengths is to the left of the dash and the model on topic spans is to the right. The GMM parameters were exactly the same as in the GM. We used a symmetric Dirichlet prior in the DMM, i.e., $\boldsymbol{\beta}$ = 0.1. If Poisson distribution was used, we initialised its parameter $\lambda_k$ to 1, and then sampled with a Gamma prior. If Geometric geometric distribution was used, the parameter of the Beta prior $\eta$ was set to 1.0. We ran 10 Markov chains for 20,000 iterations. A sample was drawn at the last iteration from each chain.

[1]The source code for Bayesseg and the GM were downloaded from http://groups.csail.mit.edu/rbg/code/. The source codes for the STSM and TMTO are available at http://web.science.mq.edu.au/∼ldu/.

TABLE I: Statistics of the nine corpora.

| Corpus | #Docs | #Segs | #Pass | Voc |
|---|---|---|---|---|
| *Wikielement* | 118 | 910 | 2,810 | 18,008 |
| *Cellphones* | 100 | 656 | 2,401 | 13,522 |
| *WikicitiesEnglish* | 100 | 1,319 | 6,670 | 41,978 |
| *WikicitiesFrench* | 100 | 1,040 | 4,074 | 30,999 |
| *Choi-3-11* | 400 | 4,000 | 27,876 | 6,390 |
| *Choi-3-5* | 100 | 1,000 | 3,928 | 3,878 |
| *Choi-6-8* | 100 | 1,000 | 6,962 | 5,266 |
| *Choi-9-11* | 100 | 1,000 | 9,855 | 6,376 |
| *ICSI* | 25 | 132 | 24,864 | 6,140 |

### A. Topic Segmentation Performance

For our models, the section boundaries can be induced from the samples drawn from the inference procedure, as in Section IV. Segmentation quality is evaluated using several metrics: PK [33], Window Diff (WD) [34] and an extension of WD (WDE) [35]. For all metrics, lower scores are better.

**Performance on Domain-Specific Corpora**: We evaluated all models on those Wikipedia articles and cellphone reviews. Bayesseg, LDASeg and STSM do not assume any topic orderings, and are given the gold standard number of segments. Table II shows the segmentation results for $K$=20 and 40. It shows that all the ordering-based models outperform the Bayesseg model on all four datasets over the three measures. The GM performs better than LDASeg, but worse than STSM, which yields the best results in the non-ordering-based models. The four variants of our TMTO outperform LDASeg and compare favourably with STSM on the three sets of Wikipedia articles. Specifically, both **G-G** and **P-P** outperform STSM on the *Wikielements* datasets at $K$=40; all the four variants perform much better than STSM on the *WikicitiesEnglish* dataset at both $K$=20 and $K$=40; on the *WikicitiesFrench* dataset, both **P-G** and **P-P** outperforms STSM at $k$=20, but only **P-G** is better than STSM at $k$=40. It is interesting that all the ordering-based models perform worse than the STSM on the *Cellphones* dataset. The poor performance of the ordering-based model is not surprising, given the formulaic nature of the *Cellphones* dataset [7].

Compared only with the GM, the four variants of our TMTO show significant improvement regardless of the value of $K$ on all datasets except for *Cellphones*. In the four variants, **P-P** has the best segmentation performance on the *Wikipedia* articles. It outperforms all the state-of-the-art topic segmentation models and the GM. On the *Cellphones* dataset, the best performance of the ordering-based models is given by the GM, but the unequal variance t-test shows there is no statistically significant difference between the GM and our models. The comparison demonstrates that on domain-specific corpora, constraining the topic assignments with topic

TABLE II: Topic segmentation results on the Wikipedia articles and cellphone reviews (scores in %). Scores in bold are the best results among the models. '*' indicates that scores derived by our models are statistically better than those by the GM with the p-value less than 0.01 from a two-tail t-test with unequal variance.

| $K$ | Model | Wikielements | | | | Cellphones | | | | WikicitiesEnglish | | | | WikicitiesFrench | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | PK | WD | WDE | Segs | PK | WD | WDE | Segs | PK | WD | WDE | Segs | PK | WD | WDE | Segs |
| Bayesseg | | 29.7 | 31.8 | 29.6 | *7.7* | 35.4 | 38.0 | 35.1 | *9.6* | 29.7 | 34.1 | 33.7 | *13.2* | 27.0 | 31.9 | 31.1 | *10.4* |
| 20 | LDASeg | 19.5 | 22.9 | 22.2 | *7.7* | 34.5 | 38.8 | 36.0 | *9.6* | 24.0 | 29.8 | 29.5 | *13.2* | 22.1 | 27.1 | 26.4 | *10.4* |
| | STSM | 18.0 | 21.0 | 20.3 | *7.7* | **30.0** | **33.8** | **31.1** | *9.6* | 22.4 | 27.9 | 27.8 | *13.2* | 21.3 | 26.2 | 25.7 | *10.4* |
| | GM | 20.2 | 24.8 | 23.3 | 8.7 | 31.8 | 37.3 | 34.3 | 10.1 | 22.7 | 28.7 | 28.0 | 14.3 | 23.2 | 29.0 | 27.4 | 10.6 |
| | G-G | 19.5 | 22.8* | 21.6 | 7.9 | 32.3 | 37.1 | 34.1 | 9.1 | 19.3* | 24.1* | 23.7* | 13.0 | 21.9 | 26.4* | 25.1* | 8.4 |
| | G-P | 18.4 | 21.6* | 20.4* | 7.9 | 34.1 | 38.3 | 35.3 | 8.4 | 19.1* | 23.7* | 23.3* | 12.9 | 21.6* | 25.9* | 24.7* | 8.0 |
| | P-G | 18.2 | 21.5* | 20.3* | 8.0 | 33.0 | 37.5 | 34.6 | 9.0 | 19.2* | 24.1* | 23.7* | 13.2 | 21.3* | 26.0* | 24.6* | 8.5 |
| | P-P | **17.8*** | **21.0*** | **19.7*** | 8.0 | 33.9 | 0.381 | 35.0 | 8.6 | **19.0*** | **23.3*** | **22.8*** | 12.6 | **21.2*** | **25.6*** | **24.4*** | 8.3 |
| 40 | LDASeg | 19.2 | 22.3 | 21.5 | *7.7* | 34.9 | 39.1 | 36.4 | *9.6* | 23.1 | 28.8 | 28.5 | *13.2* | 20.6 | 25.4 | 24.8 | *10.4* |
| | STSM | 17.9 | 21.3 | 20.6 | *7.7* | **29.5** | **33.2** | **30.6** | *9.6* | 22.9 | 28.4 | 28.4 | *13.2* | 20.7 | 25.8 | 25.2 | *10.4* |
| | GM | 20.9 | 25.7 | 24.2 | 8.9 | 30.7 | 36.8 | 33.7 | 11.2 | 21.6 | 29.0 | 28.3 | 16.3 | 23.4 | 29.1 | 27.4 | 10.6 |
| | G-G | 17.7* | 21.2* | 19.8* | 8.1 | 328 | 37.1 | 34.1 | 9.0 | 19.0* | 24.8* | 24.4* | 14.7 | 20.7* | 25.2* | 24.0* | 8.7 |
| | G-P | 18.2 | 21.5* | 20.1* | 7.9 | 33.3 | 37.3 | 34.3 | 8.7 | 18.9* | 24.5* | 24.1* | 14.6 | 21.0* | 25.3* | 24.2* | 8.5 |
| | P-G | 18.5 | 22.0* | 20.6* | 8.0 | 32.8 | 37.3 | 34.3 | 9.1 | 19.1* | 24.9* | 24.4* | 14.8 | **20.6*** | **25.1*** | **23.9*** | 8.7 |
| | P-P | **17.6*** | **20.8*** | **19.5*** | 7.9 | 33.1 | 37.0 | 34.1 | 8.7 | **18.8*** | **24.3*** | **23.8*** | 14.4 | 21.2* | 25.5* | 24.3* | 8.3 |

TABLE III: Segmentation results on Choi's datasets (scores in %). '*' indicates that scores derived by our models are statistically better than those derived by the GM with the p-value less than 0.001 from a two-tail t-test with unequal variance.

| $K$ | Model | Choi-3-11 | | | | Choi-3-5 | | | | Choi-6-8 | | | | Choi-9-11 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | PK | WD | WDE | Segs | PK | WD | WDE | Segs | PK | WD | WDE | Segs | PK | WD | WDE | Segs |
| Bayesseg | | 9.5 | 10.5 | 9.7 | *10.0* | 9.1 | 9.7 | 8.9 | *10.0* | 6.2 | 6.7 | 6.0 | *10.0* | 5.2 | 5.7 | 5.3 | *10.0* |
| 10 | LDASeg | 1.6 | 2.3 | 2.1 | *10.0* | 4.0 | 5.2 | 4.8 | *10.0* | 2.4 | 3.4 | 3.2 | *10.0* | 2.2 | 3.3 | 3.1 | *10.0* |
| | STSM | **0.8** | **1.1** | **1.0** | *10.0* | **2.0** | **2.7** | **2.5** | *10.0* | **2.1** | **2.8** | **2.7** | *10.0* | **1.5** | **2.3** | **2.2** | *10.0* |
| | GM | 15.9 | 18.3 | 17.1 | 8.8 | 16.9 | 18.7 | 17.4 | 8.6 | 15.7 | 18.1 | 16.9 | 8.9 | 15.1 | 18.0 | 16.8 | 9.0 |
| | G-G | 13.1* | 14.1* | 13.0* | 8.0 | 14.6 | 15.4* | 14.2* | 7.8 | 14.8 | 15.6* | 14.3* | 7.9 | 15.0 | 16.0* | 14.7* | 7.9 |
| | G-P | 13.0* | 13.7* | 12.6* | 7.9 | 14.1* | 14.6* | 13.4* | 7.8 | 14.1* | 14.4* | 13.2* | 7.8 | 14.1 | 14.6* | 13.3* | 7.9 |
| | P-G | 13.3* | 14.3* | 13.2* | 8.0 | 14.5* | 15.3* | 14.1* | 7.9 | 15.0 | 15.9* | 14.6* | 7.9 | 14.9 | 15.8* | 14.5* | 8.0 |
| | P-P | 12.9* | 13.6* | 12.5* | 7.9 | 14.3* | 14.7* | 13.5* | 7.8 | 14.4* | 14.8* | 13.5* | 7.8 | 14.5 | 14.8* | 13.5* | 7.8 |
| 20 | LDASeg | 0.9 | 1.4 | 1.3 | *10.0* | 1.8 | 2.3 | 2.1 | *10.0* | 1.8 | 2.4 | 2.3 | *10.0* | 1.4 | 2.1 | 1.9 | *10.0* |
| | STSM | **0.6** | **0.9** | **0.9** | *10.0* | **1.1** | **1.4** | **1.3** | *10.0* | **1.7** | **2.3** | **2.2** | *10.0* | **1.2** | **1.9** | **1.7** | *10.0* |
| | GM | 16.5 | 24.6 | 23.3 | 13.8 | 13.9 | 20.1 | 19.1 | 12.5 | 14.7 | 24.1 | 23.1 | 13.8 | 15.5 | 26.2 | 25.0 | 14.5 |
| | G-G | 6.3* | 8.1* | 7.6* | 10.1 | 6.4* | 7.7* | 7.2* | 9.8 | 7.4* | 9.4* | 8.8* | 10.1 | 7.9* | 10.6* | 10.0* | 10.5 |
| | G-P | 6.3* | 7.6* | 7.1* | 10.0 | 6.4* | 7.3* | 6.7* | 9.6 | 7.1* | 8.3* | 7.7* | 9.9 | 7.5* | 8.9* | 8.2* | 10.1 |
| | P-G | 6.3* | 8.3* | 7.8* | 10.2 | 6.7* | 8.1* | 7.5* | 9.9 | 8.2* | 10.2* | 9.6* | 10.3 | 7.4* | 10.2* | 9.5* | 10.5 |
| | P-P | 6.4* | 7.7* | 7.1* | 10.0 | 6.2* | 7.0* | 6.5* | 9.6 | 6.7* | 7.7* | 7.1* | 9.8 | 7.7* | 9.0* | 8.3* | 10.1 |

TABLE IV: Text alignment results for different number of topics. Higher scores are better. Scores in bold are the best results. '*' indicates that scores derived by our models are statistically better than those by the GM with the p-value less than 0.01.

| #Topics | Model | Wikielements | | | Cellphones | | | WikicitiesEnglish | | | WikicitiesFrench | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | R | P | F | R | P | F | R | P | F | R | P | F |
| 20 | GM | 0.594 | 0.526 | 0.557 | 0.666 | **0.538** | **0.595** | 0.647 | 0.485 | 0.554 | 0.631 | 0.430 | 0.511 |
| | G-G | 0.630* | 0.546 | 0.585* | **0.686** | 0.519 | 0.590 | 0.739* | 0.529* | 0.617* | 0.723* | 0.446 | 0.551* |
| | G-P | 0.647* | 0.572* | 0.607* | 0.681 | 0.487 | 0.568 | 0.746* | **0.536*** | **0.623*** | 0.732* | 0.456 | 0.562* |
| | P-G | **0.656*** | 0.567* | 0.608* | 0.678 | 0.510 | 0.582 | 0.740* | 0.534* | 0.620* | 0.728* | 0.448 | 0.555* |
| | P-P | 0.655* | **0.575*** | **0.612*** | 0.685 | 0.495 | 0.574 | **0.754*** | 0.530* | 0.622* | **0.744*** | **0.469*** | **0.575*** |
| 40 | GM | 0.566 | 0.520 | 0.541 | 0.648 | **0.561** | **0.601** | 0.634 | 0.520 | 0.571 | 0.626 | 0.435 | 0.513 |
| | G-G | 0.632* | 0.573* | 0.601* | 0.660 | 0.517 | 0.579 | 0.727* | 0.558* | 0.632* | 0.717* | 0.470* | 0.568* |
| | G-P | 0.634* | 0.573* | 0.602* | 0.665 | 0.504 | 0.573 | **0.737*** | **0.560*** | **0.636*** | 0.725* | 0.481* | 0.578* |
| | P-G | 0.635* | 0.568* | 0.600* | 0.648 | 0.513 | 0.572 | 0.729* | 0.560* | 0.633* | 0.727* | 0.467* | 0.568* |
| | P-P | **0.646*** | **0.576*** | **0.609*** | **0.667** | 0.512 | 0.579 | 0.732* | 0.557* | 0.632* | **0.735*** | **0.482*** | **0.581*** |

orderings benefits the topic-ordering based models, and the use of topic distribution gives LDASeg and the STSM a degree of freedom in learning.

**Performance on Randomly Generated Documents**: The objective of this set of experiments is to study how tolerant our models are to documents that do not have an ordering structure, even though they are designed to be applied to documents from one domain. We ran all experiments with $k$ = 10 and 20 on Choi's datasets. The results are reported in Table III. The four variants of TMTO and the GM perform worse than the other three segmentation models that do not consider topic orderings. The poor performance of the ordering-based models is expected, since Choi's four datasets consist of documents that are randomly created without any canonical topic ordering structure. Constraining the topic assignments by topic orderings cannot improve topic segmentation on such a dataset. However, according to the t-test, the four TMTO variants perform significantly better than the GM, which gives the worst results in all the models. The gap between our models and the other three topic segmentation models, i.e., Bayesseg, LDASeg, and STSM, decreases at $K$=20.

**Performance on Meeting Transcripts**: We further studied our models by using a more realistic ICSI dataset of meeting transcripts. The topic transitions are not as sharp as in the previous eight corpora, which makes the segmentation of these transcripts more challenging. Segmentation results are

TABLE V: Segmentation results on meeting transcript

| $K$ | Model | ICSI | | | |
|---|---|---|---|---|---|
| | | PK | WD | WDE | Segs |
| Bayesseg | | 25.7 | 31.8 | 33.4 | *5.3* |
| 10 | LDASeg | 28.9 | 32.2 | **31.7** | *5.3* |
| | STSM | 23.5 | 31.5 | 32.3 | *5.3* |
| | GM | 59.8 | 67.6 | 60.9 | 10.0 |
| | G-G | **22.5** | **29.9** | 33.4 | 6.0 |
| | G-P | 31.8 | 35.4 | 34.2 | 4.6 |
| | P-G | 23.2 | 30.3 | 34.0 | 5.9 |
| | P-P | 30.7 | 34.5 | 33.8 | 4.6 |
| 20 | LDASeg | 31.0 | 35.1 | 35.3 | *5.3* |
| | STSM | **22.5** | **30.8** | **31.6** | *5.3* |
| | GM | 66.7 | 87.4 | 79.1 | 20.0 |
| | G-G | 23.0 | 33.6 | 37.5 | 9.6 |
| | G-P | 33.9 | 39.3 | 38.3 | 5.9 |
| | P-G | 24.4 | 35.6 | 39.7 | 9.9 |
| | P-P | 32.6 | 38.3 | 37.3 | 5.7 |

in Table V. The GM model suffers significantly from an over-segmentation problem. In contrast, both **G-G** and **P-G** have lower PK and WD scores than the STSM while $k = 10$. Our models start over-segmenting the transcripts as the number of topics increases, which worsens the scores.

### *B. Text Alignment Performance*

Although our models are designed for the topic segmentation task, they can also be used for the cross-document alignment task, where text passages that address similar topics are clustered together. The goal of this task is to study how text passages from different documents semantically correlate. We compare the performance of the GM and TMTO variants on this task. Two text passages are aligned in the reference annotation if they have the same section heading, and they are aligned in the proposal if they have the same topic assignment. The alignment outputs are quantified with recall, precision and f-scores. Table IV shows the alignment results. On all three Wikipedia datasets, the best performance is achieved by the four variants of our TMTO by a sizeable margin. The t-test shows that the improvement gained by our TMTO is statistically significant. The only case when the GM outperforms our models is the *Cellphones* dataset, which is not surprising given TMTO's segmentation performance. However, the difference on the *Cellphones* dataset is not statistically significant.

## VI. Conclusion

In this paper we have presented new topic models that can capture the ordering regularities of how topics are discussed in documents from the same domain. Using the generalised Mallows model of top-$t$ orderings, the new models take the generation of a sequence of topics in each document as a multistage ranking process. In order to develop a tractable point-wise sampler we reformulated the model with a set of n-ary boundary indicator variables. The experimental results on the domain-specific sets of documents show that our model outperforms the other ordering-based models and state-of-the-art segmentation models. Those on the non-domain-specific sets of documents demonstrate that our model is tolerant to the absence of ordering structure. The most promising extension is to consider how to generalise TMTO to model infinite orderings with the infinite GMM [11]. The point-wise sampling algorithm proposed in Section IV can still be applied to the infinite model with simple modification.


## Acknowledgment

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).


## References


[1] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, 2003.

[2] M. Steyvers, P. Smyth, M. Rosen-Zvi, and T. Griffiths, "Probabilistic author-topic models for information discovery," in *KDD*, 2004, pp. 306–315.

[3] R. M. Nallapati, A. Ahmed, E. P. Xing, and W. W. Cohen, "Joint latent topic models for text and citations," in *KDD*, 2008, pp. 542–550.

[4] Y. Liu, A. Niculescu-Mizil, and W. Gryc, "Topic-link lda: Joint models of topic and author community," in *ICML*, 2009, pp. 665–672.

[5] X. Wang and A. McCallum, "Topics over time: A non-markov continuous-time model of topical trends," in *KDD*, 2006, pp. 424–433.

[6] D. Ramage, C. D. Manning, and S. Dumais, "Partially labeled topic models for interpretable text mining," in *KDD*, 2011, pp. 457–465.

[7] H. Chen, S. R. K. Branavan, R. Barzilay, and D. R. Karger, "Content modeling using latent permutations," *J. Artif. Int. Res.*, vol. 36, no. 1, pp. 129–163, 2009.

[8] M. A. Fligner and J. S. Verducci, "Distance based ranking models," *J. R. Stat. Soc. Series B (Methodological)*, pp. 359–369, 1986.

[9] ——, "Multistage ranking models," *JASA*, vol. 83, no. 403, pp. 892–901, 1988.

[10] M. Meilă and H. Chen, "Dirichlet process mixtures of generalized Mallows models," in *UAI*, 2010, pp. 358–367.

[11] M. Meilă and L. Bao, "An exponential model for infinite rankings," *J. Mach. Learn. Res.*, vol. 11, pp. 3481–3518, 2010.

[12] X. Liqun, "A multistage ranking model," *Psychometrika*, vol. 65, no. 2, pp. 217–231, 2000.

[13] D. Blei and P. Moreno, "Topic segmentation with an aspect hidden Markov model," in *SIGIR*, 2001, pp. 343–348.

[14] R. Barzilay and L. Lee, "Catching the drift: Probabilistic content models, with applications to generation and summarization," in *NAACL*, 2004, pp. 113–120.

[15] C. Sauper, A. Haghighi, and R. Barzilay, "Content models with attitude," in *ACL*, 2011, pp. 350–358.

[16] A. Gruber, Y. Weiss, and M. Rosen-Zvi, "Hidden topic Markov models," *AISTATS*, vol. 2, pp. 163–170, 2007.

[17] L. Du, W. Buntine, and H. Jin, "Sequential latent Dirichlet allocation: Discover underlying topic structures within a document," in *ICDM*, 2010, pp. 148–157.

[18] H. Wang, D. Zhang, and C. Zhai, "Structural topic model for latent topical structure analysis," in *ACL*, 2011, pp. 1526–1535.

[19] L. Du, W. Buntine, and H. Jin, "A segmented topic model based on the two-parameter poisson-dirichlet process," *Mach. Learn.*, vol. 81, no. 1, pp. 5–19, 2010.

[20] ——, "Modelling sequential text with an adaptive topic model," in *EMNLP*, 2012, pp. 535–545.

[21] D. E. Critchlow, M. A. Fligner, and J. S. Verducci, "Probability models on rankings," *J. Math. Psychol.*, vol. 35, no. 3, pp. 294 – 318, 1991.

[22] L. Nie, B. D. Davison, and X. Qi, "Topical link analysis for web search," in *SIGIR*, 2006, pp. 91–98.

[23] D. Duan, Y. Li, R. Li, R. Zhang, and A. Wen, "Ranktopic: Ranking based topic modeling," in *ICDM*, 2012, pp. 211–220.

[24] Y. Zhu, Y. Lan, J. Guo, P. Du, X. Cheng, Y. Zhu, J. Guo, Y. Lan, X.-Q. Cheng, Y., "A novel relational learning-to-rank approach for topic-focused multi-document summarization," in *ICDM*, 2013, pp. 927–936.

[25] L. Frermann, I. Titov, and M. Pinkal, "A hierarchical Bayesian model for unsupervised induction of script knowledge," in *EACL*, 2014, pp. 49–57.

[26] J. Eisenstein and R. Barzilay, "Bayesian unsupervised topic segmentation," in *EMNLP*, 2008, pp. 334–343.

[27] M. Purver, T. L. Griffiths, K. P. Körding, and J. B. Tenenbaum, "Unsupervised topic modelling for multi-party spoken discourse," in *ACL*, 2006, pp. 17–24.

[28] L. Du, W. Buntine, and M. Johnson, "Topic segmentation with a structured topic model," in *NAACL*, 2013, pp. 190–200.

[29] S. Goldwater, T. L. Griffiths, and M. Johnson, "A Bayesian framework for word segmentation: Exploring the effects of context," *Cognition*, vol. 112, no. 1, pp. 21–53, 2009.

[30] F. Y. Y. Choi, "Advances in domain independent linear text segmentation," in *NAACL*, 2000, pp. 26–33.

[31] A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, and C. Wooters, "The ICSI Meeting Corpus," in *ICASSP*, 2003, pp. 364–367.

[32] M. Galley, K. R. McKeown, E. Fosler-Lussier, and H. Jing, "Discourse segmentation of multi-party conversation," in *ACL*, 2003, pp. 562–569.

[33] D. Beeferman, A. Berger, and J. Lafferty, "Statistical models for text segmentation," *Mach. Learn.*, vol. 34, no. 1-3, pp. 177–210, 1999.

[34] L. Pevzner and M. A. Hearst, "A critique and improvement of an evaluation metric for text segmentation," *Comput. Linguist.*, vol. 28, no. 1, pp. 19–36, 2002.

[35] S. Lamprier, T. Amghar, B. Levrat, and F. Saubion, "On evaluation methodologies for text segmentation algorithms," in *ICTAI*, 2007, pp. 19–26.