Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence

Topic Segmentation with An Ordering-Based Topic Model

Lan Du, John K Pate and Mark Johnson
Department of Computing, Macquarie University
Sydney, NSW 2109, Australia
{Lan.Du, John.Pate, Mark.Johnson}@mq.edu.au

Abstract

Documents from the same domain usually discuss similar topics in a similar order. However, the number of topics and the exact topics discussed in each individual document can vary. In this paper we present a simple topic model that uses generalised Mallows models and incomplete topic orderings to incorporate this ordering regularity into the probabilistic generative process of the new model. We show how to reparameterise the new model so that a point-wise sampling algorithm from the Bayesian word segmentation literature can be used for inference. This algorithm jointly samples not only the topic orders and the topic assignments but also topic segmentations of documents. Experimental results show that our model performs significantly better than the other ordering-based topic models on nearly all the corpora that we used, and competitively with other state-of-the-art topic segmentation models on corpora that have a strong ordering regularity.

Introduction

A great amount of effort in Machine Learning and Natural Language Processing has focused on developing more comprehensive topic models that can model discourse structures found in natural language texts. They seek to model various aspects of language by taking classical topic models, like Latent Dirichlet Allocation (LDA) (Blei, Ng, and Jordan 2003), as a building block, and manipulate the graphical structure to incorporate various kinds of information into the generative process, either as specific model variables or priors. One shared intuition among these models is to use not only the word co-occurrence patterns but also discourse, syntactic, or semantic constraints. They include, for example, syntactic topic models (Griffiths et al. 2004; Boyd-Graber and Blei 2009), topic collocation models (Wang, McCallum, and Wei 2007; Griffiths, Steyvers, and Tenenbaum 2007; Johnson 2010)), and Regularised Topic Models (Newman, Bonilla, and Buntine 2011), all of which have been shown to improve topic modelling accuracy.

Instead of considering discourse features at a word level, as in the aforementioned models, we are interested in structural features at the level of sentences and paragraphs. Natural language text usually exhibits a topic structure, where a topically coherent segment consists of a set of thematically related text units, e.g., sentences or paragraphs (Halliday and Hasan 1976; Hearst 1997). Topics discussed in documents do not appear in an arbitrary order, but are ordered to facilitate the reader's comprehension. Modelling these structures with an order-aware probabilistic generative model should improve modelling accuracy.

In this paper we are interested in modelling documents from the same domain, because these documents are likely to share a common structure that similar topics tend to be discussed in a similar order. For example, consider the three English Wikipedia articles about Sydney, Melbourne and Brisbane, the three biggest cities in Australia. These articles describe the cities in terms of *History*, *Geography*, *Culture*, *Economy*, *Education*, *Infrastructure*, etc. These topics are generally discussed in the same order. For instance, *History* and *Geography* appear first, and *Geography* is discussed after *History*; *Economy* is always discussed before *Education*, which in turn precedes *Infrastructure*. Each topic is discussed in only one section, and the paragraphs in a section share the same topic. There is variation in the topic ordering, e.g., the *Culture* and *Economy* sections appear in opposite order in the Sydney and Melbourne articles.

Therefore, we want a model that can capture the regularity discussed above so that there is a canonical topic ordering shared amongst documents from the same domain, and each document-specific topic ordering varies according to the canonical ordering. We propose a simple ordering-based topic model that constrains latent topic assignments by topic orderings, as in the global model proposed by Chen et al. (2009). In our model, we represent the topic structure of a document as an ordering of topics, and generate from each topic a sequence of adjacent paragraphs. Our representation guarantees that each topic appears at most once in each document. We then posit a generalised Mallows model (GMM) (Fligner and Verducci 1986; 1988) over the entire set of orderings.

Documents from the same domain do not mention all of the topics of the domain, which means the number of topics and the exact topics discussed in each individual document vary from document to document. To capture the variation, we use a finite feature model (FFM) (Griffiths and Ghahramani 2011), which defines a distribution over binary ma-

trices, to choose which topics are discussed in a particular document. If topics are absent from a document, the order of topics in that document will no longer be a complete order over all the topics in the domain. The incompleteness introduces a challenge for learning a topic model that allows incomplete topic orderings for each document. We show how this problem can be overcome by approximating incomplete orderings with their most probable linear extensions (Cheng, Hühn, and Hüllermeier 2009).

The rest of the paper is organised as follows. We first start with a brief review of related work. Then we discuss our model and the corresponding inference algorithms, which is followed by empirical analysis. Finally, we conclude the paper with future work.

Related Work

One advantage of generative models is modularity (Griffiths et al. 2004), which allows different models to be easily combined. In this paper we are interested in those models that can incorporate information about topic structure, in particular the topic orderings, into the generative process.

There is a body of work looking at topic structure. Much of this work has assumed an HMM-structure for topic sequences, including the aspect model (Blei and Moreno 2001), the content models (Barzilay and Lee 2004; Sauper, Haghighi, and Barzilay 2011), the hidden topic Markov model (Gruber, Weiss, and Rosen-Zvi 2007), sequential LDA (Du, Buntine, and Jin 2010), the structured topic model (Wang, Zhang, and Zhai 2011), etc. These models model topic dependencies by integrating an HMM in their modelling framework, and assume that words in a sentence/paragraph share the same topic or the same topic distribution. However, those models may not be appropriate for modelling topic orderings that we are interested in, where any given topic appears at most once in a topic ordering.

Recently, the generalised Mallows model (GMM) has gained increasing popularity in NLP, due to its elegant computational properties (Fligner and Verducci 1990; Meilă et al. 2007; Meilă and Chen 2010). Chen et al. (2009) integrated the GMM into a simple generative model, called the global model (GM), and found a benefit in text analysis tasks. Frermann et al. (2014) also showed that using the GMM to encode ordering information about event types can improve the accuracy of the event clustering induction. Our model may be regarded as a variant of the global model (Chen et al. 2009). The main differences reside in the inference algorithm and the handling of incomplete topic orderings, which will be discussed in the inference section.

Our work is also related to topic segmentation, since the by-product of the inference algorithm is topic segmentations of documents. Recently, the classic topic models have been extended in various ways to improve the topic segmentation performance,[1] for example, Bayesseg (Eisenstein and Barzilay 2008), PLDA (Purver et al. 2006) and STSM (Du, Buntine, and Johnson 2013). Bayesseg ignores topic structure, and assumes that the words in each segment are generated from a segment-specific Dirichlet-Multinomial language model. PLDA assumes a simple Markov structure on topic distributions of segments. In contrast, the STSM can learn a simple hierarchical topic structure by segmenting documents into sections. However, none of them consider topic ordering regularities.

[1] See (Purver 2011; Misra et al. 2011; Riedl and Biemann 2012) for more details about text segmentation with topic models.

Modelling with Incomplete Topic Orderings

In this section we present a topic model of incomplete ordering (TMIO). The basic idea of our model can be summarised in four points: 1) each paragraph is assigned to only one topic; 2) each topic is discussed at most once in a document in a continuous fragment of text, i.e., a sequence of adjacent sentences or paragraphs; 3) a canonical topic ordering is shared amongst a set of documents; 4) the incomplete topic ordering of each document is generated by intersecting a complete ordering generated from the GMM and a subset of topics chosen by the FMM. The details of the probabilistic generative process are as follows.

Generating Complete Topic Orderings: Given a set of K topics, denoted by $\mathbb{K} = \{1, 2, 3, \ldots, K\}$, the number of possible permutations of the topics is $K!$. Let $\mathbf{\Pi}$ be the set of all $K!$ permutations. A single permutation $\boldsymbol{\pi} \in \mathbf{\Pi}$ can be represented as either a ranking or an ordering. For the ranking representation, $\boldsymbol{\pi} = [\boldsymbol{\pi}(1), \boldsymbol{\pi}(2), \ldots, \boldsymbol{\pi}(K)]$, where $\boldsymbol{\pi}(i)$ is the rank of topic i; for the ordering representation, $\boldsymbol{\pi} = [\boldsymbol{\pi}^{-1}(1), \boldsymbol{\pi}^{-1}(2), \ldots, \boldsymbol{\pi}^{-1}(K)]$, where $\boldsymbol{\pi}^{-1}(j)$ is the topic at rank j. We will use the *ordering* representation unless otherwise stated. Given a canonical ordering $\boldsymbol{\sigma}$, any ordering $\boldsymbol{\pi}$ can be uniquely identified by $K-1$ non-negative integers, $s_1(\boldsymbol{\pi}|\boldsymbol{\sigma}), s_2(\boldsymbol{\pi}|\boldsymbol{\sigma}), \ldots, s_{K-1}(\boldsymbol{\pi}|\boldsymbol{\sigma})$, defined by $s_j(\boldsymbol{\pi}|\boldsymbol{\sigma}) = \sum_{l>j} 1_{\boldsymbol{\sigma}(\boldsymbol{\pi}^{-1}(j))>\boldsymbol{\sigma}(\boldsymbol{\pi}^{-1}(l))}$. In words, s_j counts the number of topics ranked after topic $\boldsymbol{\pi}^{-1}(j)$ in the proposed ordering $\boldsymbol{\pi}$ but before $\boldsymbol{\pi}^{-1}(j)$ in the canonical ordering $\boldsymbol{\sigma}$. The GMM (Fligner and Verducci 1986; 1988) 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 an ordering $\boldsymbol{\pi} \in \mathbf{\Pi}$ is defined as

$$GMM(\boldsymbol{\pi}; \boldsymbol{\sigma}, \boldsymbol{\theta}) = \prod_{j=1}^{K-1} \frac{e^{-\theta_j s_j(\boldsymbol{\pi}|\boldsymbol{\sigma})}}{\psi_j(\theta_j)}, \quad (1)$$

where $\psi_j(\theta_j)$ is a partition function given by $\psi_j(\theta_j) = \frac{1-e^{-(K-j+1)\theta_j}}{1-e^{-\theta_j}}$. Eq (1) assigns probability to an ordering, which decreases exponentially with the distance from the canonical ordering. In this paper, we are interested in the case where $\theta_i > 0$ for $i = 1, \ldots, K-1$, which means that the probability distribution has a unique maximum for $\boldsymbol{\pi} = \boldsymbol{\sigma}$. The canonical ordering and the dispersion parameters elegantly capture the topic ordering regularity observed in documents from the same domain.

Generating Incomplete Topic Orderings: Given an actual set of documents from a domain, such as a set of English Wikipedia articles about major cities, it is unlikely that each document will discuss every topic that is relevant to the domain. Instead, some documents will not discuss some topics, and the number of topics, and the topics themselves, will vary from document to document. For example, in the three Wikipedia articles about Sydney, Melbourne, and Bris-

bane, *Annual events* and *Environment* are discussed only in the articles about Brisbane and Melbourne respectively, but neither is mentioned in Sydney article.

To decide which topics are covered by a document, we use the FMM, where topics are viewed as features and documents as objects. Assuming there are D documents, one can construct a $D \times K$ binary matrix $\mathbf{B}$ that indicates which topics are discussed by each individual document, with $b_{d,k} = 1$ if document d discusses topic k and 0 otherwise. In the FFM, the topics are generated independently according to their own Binomial distributions with a shared Beta prior, Beta$(\frac{\alpha}{K}, 1)$. The conditional probability of d discussing k is $p(b_{d,k} = 1|\mathbf{b}_k^{-d}) = \frac{m_k^{-d} + \frac{\alpha}{K}}{D + \frac{\alpha}{K}}$, where m_k^{-d} indicates the number of documents discussing topic k without document d. Given a full topic ordering $\boldsymbol{\pi}_d$ generated from the GMM, and a binary vector $\boldsymbol{b}_d$, an incomplete ordering $\boldsymbol{\pi}'_d$ can be expressed as the intersection of $\boldsymbol{\pi}_d$ and a subset of topics induced from $\boldsymbol{b}_d$, i.e., $\boldsymbol{\pi}'_d = \boldsymbol{\pi}_d \cap (\boldsymbol{\sigma} \circ \boldsymbol{b}_d)$, e.g., if $\boldsymbol{\pi}_d = (3, 1, 5, 6, 2, 4)$, $\boldsymbol{b}_d = (1, 0, 1, 0, 1, 0)$ and $\boldsymbol{\sigma} = (1, 2, 3, 4, 5, 6)$, then $\boldsymbol{\pi}'_d = (3, 1, 5)$.

There is a set of complete orderings that are compatible with $\boldsymbol{\pi}'_d$, and the members of this set are called the linear extensions of the incomplete ordering $\boldsymbol{\pi}'_d$. For example, both $(3, 1, 5, 6, 2, 4)$ and $(3, 2, 1, 4, 5, 6)$ are compatible with $\boldsymbol{\pi}'_d$, because 3 precedes 1 and 1 precedes 5 in both orderings. We therefore compute the probability of $p(\boldsymbol{\pi}'_d)$ by summing over the probabilities of all its linear extensions, i.e., $p(\boldsymbol{\pi}'_d) = \sum_{\boldsymbol{\pi}^* \in \boldsymbol{\Pi}(\boldsymbol{\pi}'_d)} GMM(\boldsymbol{\pi}^*; \boldsymbol{\sigma}, \boldsymbol{\theta})$, where $\boldsymbol{\Pi}(\boldsymbol{\pi}'_d)$ denotes the set of linear extensions of $\boldsymbol{\pi}'_d$. The extension of the FFM to incomplete ordering allows us to model each document's partial preference on topics. In other words, each document selects the subset of topics that it will discuss, and then discusses them in a specific order.

Generating Topic Span and Words: The next step in our generative process is to generate topic assignments of paragraphs in a document, and then generate the words in them. We have generated the incomplete order $\boldsymbol{\pi}'_d$ for document d from the joint model of the GMM and the FFM. $\boldsymbol{\pi}'_d$ also specifies the number of sections that need to be generated in d. For each topic k in $\boldsymbol{\pi}'_d$, we first generate the number of paragraphs discussing this topic, called the topic span length $l_{d,k}$ from a Poisson distribution Poiss(λ). Varying λ allows one to segment a document in a more or less fine-grained way. Here we assume all the topics share the same Poisson distribution over spans. Words in paragraphs within the topic span are then generated from topic k with a Dirichlet-Multinomial model. We have assumed that each paragraph can discuss only one topic, so words in the same paragraph are generated from the same topic.

Given a set of topics $\mathbb{K}$ and a set of documents, $\mathbb{D} = \{1, 2, 3, \ldots, D\}$, the full generative process can be read off from the above components as the following:

1. For each topic $k \in \{1, \ldots, K\}$,
   (a) Draw word distribution $\boldsymbol{\phi}_k \sim \text{Dirichlet}_V(\boldsymbol{\beta})$.
   (b) Draw Bernoulli parameter in the FMM, $\mu_k \sim \text{Beta}(\frac{\alpha}{K}, 1)$.
2. For each document $d \in \{1, \ldots, D\}$,
   (a) Draw a full ordering $\boldsymbol{\pi}_d$ from the GMM with Eq (1).
   (b) For each topic k, draw $b_{d,k} \sim \text{Bernoulli}(\mu_k)$.
   (c) Generate incomplete ordering $\boldsymbol{\pi}'_d = \boldsymbol{\pi}_d \cap (\boldsymbol{\sigma} \circ \boldsymbol{b}_d)$.
   (d) For each section $k \in \boldsymbol{\pi}'_d$,
      i. Draw topic span $l_{d,k} \sim \text{Poiss}(\lambda)$.
      ii. For each paragraph j in the span, let $z_{d,j} = k$ and generate a set of words $\boldsymbol{w}_{d,p}$ from Discrete(ϕ_k).

where we assume that $\boldsymbol{\sigma}$ is always the identity ordering, i.e., $\boldsymbol{\sigma} = (1, 2, \ldots, K)$, as in the global model (Chen et al. 2009), because topics themselves are latent variables to be learnt.

Posterior Inference

Inspired by the boundary sampling algorithm used in Bayesian segmentation (Goldwater, Griffiths, and Johnson 2009; Du, Buntine, and Johnson 2013), we reformulate TMIO with a set of topic boundary indicator variables so that a tractable point-wise sampling algorithm can be used.

In Bayesian word segmentation (Goldwater, Griffiths, and Johnson 2009), the sampler samples word boundaries between phonemes. Our sampler instead samples topic boundaries between paragraphs, which segment a sequence of paragraphs in a document into thematically coherent sections. Here we show how one can reparameterise the model presented in the previous section 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 $j = \max\{p \mid 1 \leq p \leq i-1 \wedge \rho_{d,p} \neq 0\}$, and the topic of the section is $\rho_{d,i}$, which takes values in $\{1, \ldots, K\}$. For example, given the canonical ordering $\boldsymbol{\sigma} = (1, 2, 3, 4, 5, 6)$ and a vector of boundary indicators is $(0, 0, 3, 0, 0, 1, 0, 0, 5)$, we can induce the paragraph topic assignments of a document, $(3, 3, 3, 1, 1, 1, 5, 5, 5)$, topic span lengths $\boldsymbol{l}_d = (3, 3, 3)$, the partial ordering $\boldsymbol{\pi}'_d = (3, 1, 5)$, and $\boldsymbol{b}_d = (1, 0, 1, 0, 1, 0)$.

Now we illustrate the point-wise sampling algorithm using the above example. Currently, 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—not putting (**H1**) or putting (**H2**) 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 $\boldsymbol{\pi}'_d$, $\boldsymbol{\rho}_d$, $\boldsymbol{b}_d$ and $\boldsymbol{l}_d$, and are specified as follows.

H1: There is not a boundary. This hypothesis corresponds to setting $\rho_{d,3} = 0$ and 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 current topic of $\mathbb{S}_2$, we sample a new topic for the merged section $\mathbb{S}_0$ by resampling $\rho_{d,6}$.[2] Note

[2] We have also tried the algorithm that uses the current topic assignment of $\mathbb{S}_2$ as the topic assignment of the merged section $\mathbb{S}_0 = \mathbb{S}_1 \cup \mathbb{S}_2$ for a merge move, and only resamples the topic assignment of $\mathbb{S}_1$ without changing that of $\mathbb{S}_2$ for a split move. We found that the Markov chain mixed very slowly, and gave bad segmentation results.

that the value of $\rho_{d,6}$ must be in $\mathbb{T} = \{1, 2, 3, 4, 6\}$, because topic 5 has already been assigned to the last section of the document. Thus, the new state with $\rho_{d,3} = 0$ and $\rho_{d,6} = k$ is $\boldsymbol{\rho}_d = (0, 0, 0, 0, 0, k, 0, 0, 5)$, $\boldsymbol{\pi}'_d = (k, 5)$, $\boldsymbol{l}_d = (6, 3)$ and $\boldsymbol{b}_d$ with k^{th} and 5^{th} elements equal to 1 and the others equal to 0. Let $\boldsymbol{\mu}$ indicate all model parameters and statistics not affected by the boundary that is currently resampled. The probability of the new state is

$$p(\rho_{d,3} = 0, \rho_{d,6} = k \mid \boldsymbol{\mu}) \propto \left(\sum_{\boldsymbol{\pi}^* \in \boldsymbol{\Pi}(\boldsymbol{\pi}'_d)} \prod_{j=1}^{K-1} \frac{e^{-\theta_j s_j(\boldsymbol{\pi}^* | \sigma)}}{\psi_j(\theta_j)} \right)$$

$$\frac{m_k^{-d} + \frac{\alpha}{K}}{D + \frac{\alpha}{K}} \frac{\lambda^{l_{d,k}}}{l_{d,k}!} \frac{\prod_{v \in \mathbb{S}_0} \left[n_{k,v}^{-\mathbb{S}_0} + \beta_v \mid 1 \right]_{n_v^{\mathbb{S}_0}}}{\left[\sum_v (n_{k,v}^{-\mathbb{S}_0} + \beta_v) \mid 1 \right]_{N^{\mathbb{S}_0}}}, \quad (2)$$

where $n_{k,v}^{-\mathbb{S}_0}$ is the number of words assigned to topic k but not in $\mathbb{S}_0$, $n_v^{\mathbb{S}_0}$ is the number of words v in the merged $\mathbb{S}_0$, $N^{\mathbb{S}_0} = \sum_v n_v^{\mathbb{S}_0}$ and $[x|y]_N = x(x + y)...(x + (N - 1)y)$. Now the probability of not putting a boundary at $\rho_{d,3}$ is $\sum_{k \in \mathbb{T}} p(\rho_{d,3} = 0, \rho_{d,6} = k | \boldsymbol{\mu})$.

H2: There is a boundary. If there is boundary after the third paragraph, the segmentation of document d will not change. However, we will sample 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 $\boldsymbol{\rho}_d = (0, 0, k_1, 0, 0, k_2, 0, 0, 5)$, $\boldsymbol{\pi}'_d = (k_1, k_2, 5)$, $\boldsymbol{l}_d = (3, 3, 3)$ and with k_1^{th}, k_2^{th}, and 5^{th} elements equal to 1 and the others equal to 0. The probability of the new state is

$$p(\rho_{d,3} = k_1, \rho_{d,6} = k_2 \mid \boldsymbol{\mu}) \propto \left(\sum_{\boldsymbol{\pi}^* \in \boldsymbol{\Pi}(\boldsymbol{\pi}'_d)} \prod_{j=1}^{K-1} \frac{e^{-\theta_j s_j(\boldsymbol{\pi}^* | \sigma)}}{\psi_j(\theta_j)} \right)$$

$$\prod_{i=1,2} \frac{m_{k_i}^{-d} + \frac{\alpha}{K}}{D + \frac{\alpha}{K}} \frac{\lambda^{l_{d,k_i}}}{l_{d,k_i}!} \frac{\prod_{v \in \mathbb{S}_i} \left[n_{k_i,v}^{-\mathbb{S}_0} + \beta_v \mid 1 \right]_{n_v^{\mathbb{S}_i}}}{\left[\sum_v (n_{k_i,v}^{-\mathbb{S}_0} + \beta_v) \mid 1 \right]_{N^{\mathbb{S}_i}}}. \quad (3)$$

where $n_v^{\mathbb{S}_i}$ is the number of words v in section $\mathbb{S}_i$, $N^{\mathbb{S}_i} = \sum_v n_v^{\mathbb{S}_i}$. The probability of placing a boundary at $\rho_{d,3}$ is $\sum_{k_1,k_2} p(\rho_{d,3} = k_1, \rho_{d,6} = k_2 \mid \boldsymbol{\mu})$.

We have noticed that it is computationally infeasible to enumerate all the possible linear extensions of a given incomplete ordering. To have a tractable sampler, we approximate $p(\boldsymbol{\pi}'_d)$ by using the probability of the most probable linear extension $\boldsymbol{\pi}^*_d$ of $\boldsymbol{\pi}'_d$ (Cheng, Hühn, and Hüllermeier 2009). $\boldsymbol{\pi}^*_d$ is the complete topic ordering compatible with $\boldsymbol{\pi}'_d$ that minimises the Kendall distance, i.e., $\sum_{j=1}^{K-1} s_j(\boldsymbol{\pi}^*_d | \boldsymbol{\sigma})$. Finding the most likely extension is much more efficient than summing because we know that, in the max, the unchosen topics should appear in linear order. With this approximation, Eqs (2) and (3) can be simplified by replacing the summation with a maximisation. Those linear extensions are also used in sampling the dispersion parameters $\boldsymbol{\theta}$ of the GMM. As a member of the exponential family, the GMM has a natural conjugate prior (Fligner and Verducci 1990; 1988; Meilă and Chen 2010), which is $p(\boldsymbol{\sigma}, \boldsymbol{\theta}; \boldsymbol{s}', s_0) \propto e^{-s_0 \sum_{j=1}^{K-1} (\theta_j s'_j + \ln \psi_j(\theta_j))}$, where $\boldsymbol{s}' > 0$. Intuitively, s_0 indicates the number of prior trials, and $s'_j s_0$ denotes the total number of prior inversions. The hyperparameter s'_j is set to $s'_j = \frac{1}{e^{\theta_0} - 1} - \frac{K-j+1}{e^{(n-j+1)\theta_0} - 1}$, which is derived by setting the ML estimate of θ_j to θ_0, a common prior on θ_j (Chen et al. 2009). Given D observed orderings, the posterior is proportional to $\prod_{j=1}^{K-1} \frac{e^{-\theta_j (\sum_{i=1}^{D} s_j(\boldsymbol{\pi}_i^* | \boldsymbol{\sigma}) + s_0 s'_j)}}{\psi_j(\theta_j)^{s_0 + D}}$, from which we sample the values of $\boldsymbol{\theta}$ with a slice sampler (Neal 2003).

The use of a point-wise sampler and the approximation by maximisation differentiates our model from the global model (Chen et al. 2009). The sampler of the global model works as follows. Let $\boldsymbol{t}_d = (1, 5, 3, 3, 1, 5, 5, 1, 3)$ and $\boldsymbol{\pi}_d = (3, 1, 5, 6, 2, 4)$ be the topics and topic ordering drawn for document d. The global model reorders $\boldsymbol{t}_d$ according to $\boldsymbol{\pi}_d$ to derive the topic assignments of paragraphs, i.e., (3, 3, 3, 1, 1, 1, 5, 5, 5), and all the topics in $\boldsymbol{\pi}_d$ but not in $\boldsymbol{t}_d$ will simply be ignored. The first step in its sampler is to resample the i^{th} topic draw in $\boldsymbol{t}_d$ and reorder the new $\boldsymbol{t}_d$ according to the fixed $\boldsymbol{\pi}_d$ to derive the new topic assignments of paragraphs. The second step is to sample a new $\boldsymbol{\pi}_d$ and reorder the fixed $\boldsymbol{t}_d$ according to the new $\boldsymbol{\pi}_d$. Instead of using a two-stage Gibbs sampler, our sampler reformulates the sampling problem by associating a boundary indicator variable with each paragraph. Changing one indicator's value simultaneously changes the topic assignment of a set of consecutive paragraphs and the topic ordering. The incomplete topic orderings caused by unused topics are approximated by their most probable extensions in our sampler. These extensions are computed according to their distances to the canonical ordering. The experimental results show that our model with the new sampler significantly outperforms the global model in both the segmentation and alignment tasks.

Experiments

In this section we compare our TMIO model to four state-of-the-art topic segmentation models, one ordering-based and four with no sense of ordering, in two text analysis tasks (topic segmentation and cross-document alignment) with two different kinds of documents. The first kind of document matches the modelling assumptions: they are from the same domain and have shared ordering regularities. We show that our model significantly improves performance on both tasks with these documents. The second kind of documents is benchmark documents that do not match the modelling assumptions in that they lack shared ordering regularities. We show that our model degrades in performance more gracefully than the other ordering-based topic model, although the non-ordering based models perform best here.

We use the following sets of corpora. The first set contains four corpora (Chen et al. 2009) whose documents are assumed to exhibit the ordering regularity. Specifically, *WikicitiesEnglish* and *WikicitiesFrench* contain Wikipedia articles about the world's 100 largest cities by population in English and French respectively, *Wikielements* contains 118 English Wikipedia articles about chemical elements, and *Cellphones* contains 100 cellphone reviews. We consider section boundaries to be gold topic boundaries. These corpora are used in both topic segmentation and text alignment tasks. The second set consists of four of Choi's data sets (Choi 2000), which have been treated as benchmark data sets for evaluating topic segmentation models. Specifically, we used *Choi-3-11*, *Choi-3-5*, *Choi-6-8* and *Choi-9-11*. Each document in these data sets is a concatenation of 10 segments. n varies

Table 1: Topic segmentation results. PK, WD and WDE scores are in %. * indicates best scores yielded by all the five models, and the boldface indicates those yielded by the ordering-based models.

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*
10	LDASeg	19.2	22.4	21.6	*7.7*	35.4	39.5	36.4	*9.6*	26.7	32.2	31.8	*13.2*	22.9	28.3	27.6	*10.4*
	STSM	18.3*	21.3*	20.5*	*7.7*	31.3*	35.2*	32.4*	*9.6*	23.7*	29.4	29.0	*13.2*	22.2*	27.1*	26.4*	*10.4*
	GM	22.0	24.9	23.6	6.7	**33.2**	**37.5**	**34.5**	8.0	25.6	29.1	28.7	8.9	26.4	30.7	29.5	7.4
	TMIO	**21.0**	**23.5**	**22.2**	**6.4**	34.3	38.0	35.0	7.6	**23.7***	**26.8***	**26.5***	8.9	**24.3**	**28.4**	**27.0**	7.0
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
	TMIO	**17.2***	**20.8***	**19.5***	8.2	32.3	**36.9**	**33.9**	9.3	**18.8***	**23.8***	**23.5***	13.7	**20.7***	**25.8***	**24.5***	9.3

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
	TMIO	**13.0**	**13.6**	**12.5**	7.9	**14.2**	**14.6**	**13.4**	7.7	**14.3**	**14.6**	**13.3**	7.8	**14.1**	**14.4**	**13.1**	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
	TMIO	**6.4**	**7.7**	**7.2**	10.0	**6.0**	**6.8**	**6.3**	9.6	**6.8**	**7.7**	**7.1**	9.8	**7.9**	**9.4**	**8.7**	10.2

from 3 and 11. Since they are randomly concatenated documents, the topic ordering regularities do not apply. The final "*Clinical*" corpus (Eisenstein and Barzilay 2008) contains 227 documents, each of which is a chapter of a medical textbook. Section breaks are selected to be the true boundaries.

The models we compared were[3]: **Bayesseg**: we used the configuration included in the source code package, named *dp.config*. It was given the gold-standard number of segments in training. **The Global Model (GM)**: We used the settings reported by Chen et al. (2009). For each dataset, we run the GM 10 times with different initialisation. Results reported are the average of 10 runs. **LDASeg** & **STSM**: We ran 10 randomly initialised Markov chains. 200 samples were drawn from each chain after 25,000 iterations. The marginal boundary probabilities from the total of 2,000 samples were thresholded to give the gold-standard number of segments. We used the same parameter settings as Du, Buntine, and Johnson (2013). **TMIO**: the GMM parameters were exactly the same as in the GM. We used a symmetric Dirichlet prior in the Dirichlet-Multinomial model, i.e., $\beta = 0.1$. We set the parameter of the Poisson distribution to 1.0 (unless otherwise stated), and set α in the FFM to the number of topics. Results are the average of the samples drawn from 10 runs. Except for Bayesseg, the other four models take a parameter K, the number of topics. We report results in the two text analysis tasks using both K=10 and K=20.[4]

[3]The source code for Bayesseg and GM were downloaded from http://groups.csail.mit.edu/rbg/code/. The source code for STSM was downloaded from http://web.science.mq.edu.au/∼ldu/code.html.

[4]In general, the higher K is, the finer-grained topics are, which may result in a more finer-grained segmentation of a document. For those ordering-based models, K also controls the upper bound on the number of segments they can learn. In the preliminary experiments, we varied K from 10 to 40, and observed that the optimal K value is approximately twice the average true number of segments in a corpus. Due to the space limitation, we report K=10 and K=20.

Table 2: Topic segmentation results on clinical books.

K	Model	*Clinical*			
		PK	WD	WDE	#Segs
Bayesseg		34.5	35.3*	33.4*	*4.0*
10	LDASeg	34.9	38.3	36.4	*4.0*
	STSM	31.8*	35.9	33.9	*4.0*
	GM	43.5	54.3	51.7	8.1
	TMIO	**33.7**	**35.9**	**34.8**	3.0
20	LDASeg	34.1	37.4	35.5	*4.0*
	STSM	32.8*	36.5	34.5	*4.0*
	GM	50.4	68.2	64.5	14.1
	TMIO	**34.9**	**37.8**	**36.2**	3.5

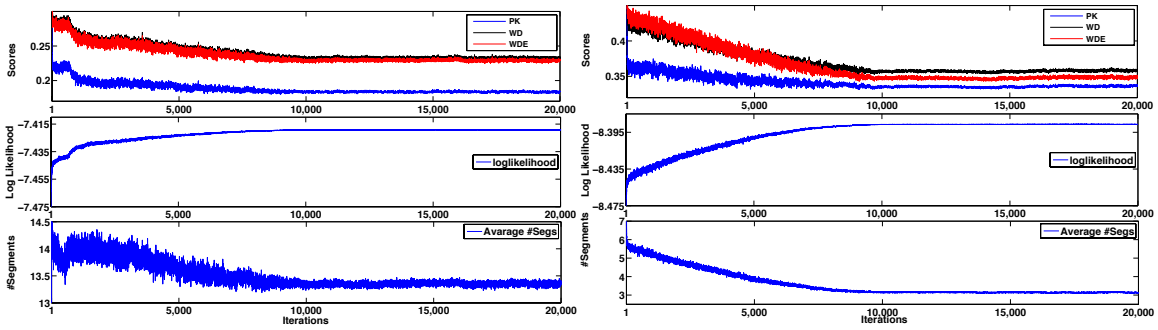
Topic Segmentation Performance

The goal of topic segmentation is to automatically divide a document into topically coherent segments (e.g., sections). Segmentation quality is evaluated using several metrics: PK (Beeferman, Berger, and Lafferty 1999), Window Diff (WD) (Pevzner and Hearst 2002) and an extension of WD (WDE) (Lamprier et al. 2007). Lower scores are better.

We first compare our TMIO with the other four models on a set of corpora that contain Wikipedia articles or cellphone reviews. Results are shown in the upper block of Table 1. The Bayesseg yields the highest (worst) scores among all five models. For K=10, the lower scores are usually obtained by the STSM and LDASeg, except for the *WikipediaEnglish* corpus. Both the STSM and LDASeg assume a topic distribution per section, which gives them a degree of freedom in learning. The ordering-based models undersegment the four corpora, since the number of segments they learn is bounded by K and is usually less than K. However, the performance of the ordering-based models improves for K=20. Our TMIO significantly outperforms all other models on the three datasets extracted from Wikipedia. In par-

Table 3: Text alignment results for different number of topics. Higher scores are better. Scores in bold are the best results.

K	Model	Wikielements			Cellphones			WikicitiesEnglish			WikicitiesFrench		
		R	P	F	R	P	F	R	P	F	R	P	F
10	GM	0.667	0.457	0.542	0.729	**0.477**	**0.576**	0.708	0.410	0.519	0.657	0.346	0.453
	TMIO	**0.698**	**0.487**	**0.573**	**0.740**	0.464	0.569	**0.779**	**0.447**	**0.568**	**0.727**	**0.374**	**0.494**
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
	TMIO	**0.654**	**0.580**	**0.614**	**0.677**	0.528	0.593	**0.759**	**0.542**	**0.633**	**0.736**	**0.479**	**0.580**



(a) WikicitiesEnglish with $k = 20$ (b) Clinical with $k = 10$

Figure 1: PK, WD, WDE, log-likilihood and average number of segments as a function of iterations for TMIO.

ticular, the TMIO achieves the lowest (best) scores on the *WikipediaEnglish* corpus for both K=10 and K=20. This might be because articles in *WikipediaEnglish* exhibit particularly strong topic ordering regularities. The poor performance of both ordering-based models on the *Cellphones* dataset may be due to its formulaic nature (Chen et al. 2009). It is worth highlighting the difference in decoding between the ordering-based models and the non-ordering based models; for both the TMIO and the GM, we simply used the last sample from each Markov chain, but for LDASeg and the STSM we used marginal boundary probabilities from 2,000 samples. If we instead use only the last sample from each Markov chain for LDASeg and STSM, segmentation performance decreases.

The lower block of Table 1 shows the performance of the five models 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 that do. Our TMIO model and the GM perform worse than the other three segmentation models, as expected (since the documents are randomly ordered). Constraining the topic assignments by topic orderings cannot improve topic segmentation on those datasets. At K=20 the gap between TMIO and Bayesseg, LDASeg, and STSM, decreases, and TMIO can recover the true number of segments most accurately for all Choi's four datasets. These results show that the structural properties of documents are important for the use of the ordering-based models.

We further compared the five models on a collection of clinical book chapters. The results with Poisson parameter λ=15 are reported in Table 2. The results show that for K=10 our TMIO compares favourably with Bayesseg and STSM, particularly on WD and WDE scores, and outperforms LDASeg. Our TMIO outperforms the GM on all the above segmentation evaluations. Finally, we observe that our posterior sampler mixes quickly. Figure 1 presents various diagnostics for the *WikicitiesEnglish* and *Clinical* datasets as a function of sampling iteration, and we can see that all diagnostics stabilise after around 10,000 iterations.

Cross-Document Alignment Performance

The cross-document alignment task is to cluster together text passages that address similar topics and to study how text passages from different documents topically correlate. We compare the performance of the GM and TMIO 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 results are quantified with Recall, Precision and F-scores and shown in Table 3. On all three Wikipedia datasets, the best performance is achieved by TMIO over the three measures by a notable margin. The only case when the GM outperforms TMIO is the *Cellphones* dataset, which is not unexpected given TMIO's segmentation performance.

In summary, comparing the two ordering-based model shows that our TMIO performs significantly better than the GM in both the segmentation task and the cross-document alignment task. Those results suggest that the use of pointwise sampling algorithm with the approximation of incomplete topic ordering leads to real improvements.

Conclusion

In this paper we have presented a simple probabilistic generative process that can model the topic ordering regularities in documents that from a similar domain. We reformulated the model and approximated the probability of an incomplete topic ordering by using the probability of its most probable

linear extension, based on which a tractable point-wise sampler has been developed. The experimental results show that our model with the new sampling algorithm significantly outperforms the other ordering-based models and competitively with other state-of-the-art segmentation models on those have strong topic-ordering regularity. We have also observed that the performance of the ordering-based topic models largely depends on whether documents in a corpus which they are applied to share a common topic structure or not. As future work, it is interesting to try modelling infinite orderings (Meilă and Bao 2010).

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).

References

Barzilay, R., and Lee, L. 2004. Catching the drift: Probabilistic content models, with applications to generation and summarization. In *Proceedings of NAACL*, 113–120.

Beeferman, D.; Berger, A.; and Lafferty, J. 1999. Statistical models for text segmentation. *Mach. Learn.* 34(1-3):177–210.

Blei, D., and Moreno, P. 2001. Topic segmentation with an aspect hidden Markov model. In *Proceedings of SIGIR*, 343–348.

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent Dirichlet allocation. *J. Mach. Learn. Res.* 3:993–1022.

Boyd-Graber, J. L., and Blei, D. 2009. Syntactic topic models. In *Proceedings of NIPS*. 185–192.

Chen, H.; Branavan, S. R. K.; Barzilay, R.; and Karger, D. R. 2009. Content modeling using latent permutations. *J. Artif. Int. Res.* 36(1):129–163.

Cheng, W.; Hühn, J.; and Hüllermeier, E. 2009. Decision tree and instance-based learning for label ranking. In *Proceedings of ICML*, 161–168.

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

Du, L.; Buntine, W.; and Jin, H. 2010. Sequential latent Dirichlet allocation: Discover underlying topic structures within a document. In *Proceedings of ICDM*, 148–157.

Du, L.; Buntine, W.; and Johnson, M. 2013. Topic segmentation with a structured topic model. In *Proceedings of NAACL*, 190–200.

Eisenstein, J., and Barzilay, R. 2008. Bayesian unsupervised topic segmentation. In *Proceedings of EMNLP*, 334–343.

Fligner, M. A., and Verducci, J. S. 1986. Distance based ranking models. *J. Roy. Statist. Soc. Ser. B* 359–369.

Fligner, M. A., and Verducci, J. S. 1988. Multistage ranking models. *J. Am. Statist. Assoc.* 83(403):892–901.

Fligner, M. A., and Verducci, J. S. 1990. Posterior probabilities for a consensus ordering. *Psychometrika* 55(1):53–63.

Frermann, L.; Titov, I.; and Pinkal, M. 2014. A hierarchical Bayesian model for unsupervised induction of script knowledge. In *Proceedings of EACL*, 49–57.

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

Griffiths, T. L., and Ghahramani, Z. 2011. The Indian buffet process: An introduction and review. *J. Mach. Learn. Res.* 12:1185–1224.

Griffiths, T. L.; Steyvers, M.; Blei, D. M.; and Tenenbaum, J. B. 2004. Integrating topics and syntax. In *Proceedings of NIPS*, 537–544.

Griffiths, T. L.; Steyvers, M.; and Tenenbaum, J. B. 2007. Topics in semantic representation. *Psychological Review* 114(2):211–244.

Gruber, A.; Weiss, Y.; and Rosen-Zvi, M. 2007. Hidden topic Markov models. *J. Mach. Learn. Res. - Proceedings Track* 2:163–170.

Halliday, M. A. K., and Hasan, R. 1976. *Cohesion in English*. Longman Pub Group.

Hearst, M. A. 1997. Texttiling: Segmenting text into multi-paragraph subtopic passages. *Comput. Linguist.* 23(1):33–64.

Johnson, M. 2010. PCFGs, topic models, adaptor grammars and learning topical collocations and the structure of proper names. In *Proceedings of ACL*, 1148–1157.

Lamprier, S.; Amghar, T.; Levrat, B.; and Saubion, F. 2007. On evaluation methodologies for text segmentation algorithms. In *Proceedings of ICTAI*, 19–26.

Meilă, M., and Bao, L. 2010. An exponential model for infinite rankings. *J. Mach. Learn. Res.* 11:3481–3518.

Meilă, M., and Chen, H. 2010. Dirichlet process mixtures of generalized Mallows models. In *Proceedings of UAI*, 358–367.

Meilă, M.; Phadnis, K.; Patterson, A.; and Bilmes, J. 2007. Consensus ranking under the exponential model. In *Proceedings of UAI*, 285–294.

Misra, H.; Yvon, F.; Cappé, O.; and Jose, J. 2011. Text segmentation: A topic modeling perspective. *Information Processing & Management* 47(4):528–544.

Neal, R. M. 2003. Slice sampling. *Annals of statistics* 705–741.

Newman, D.; Bonilla, E.; and Buntine, W. 2011. Improving topic coherence with regularized topic models. In *Proceedings of NIPS*. 496–504.

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

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

Purver, M. 2011. Topic segmentation. *Spoken Language Understanding: Systems for Extracting Semantic Information from Speech* 291–317.

Riedl, M., and Biemann, C. 2012. Text Segmentation with Topic Models. *JLCL* 27(47-69):13–24.

Sauper, C.; Haghighi, A.; and Barzilay, R. 2011. Content models with attitude. In *Proceedings of ACL*, 350–358.

Wang, X.; McCallum, A.; and Wei, X. 2007. Topical n-grams: Phrase and topic discovery, with an application to information retrieval. In *Proceedings of ICDM*, 697–702.

Wang, H.; Zhang, D.; and Zhai, C. 2011. Structural topic model for latent topical structure analysis. In *Proceedings of ACL*, 1526–1535.