Multi-document Topic Segmentation

Minwoo Jeong
Saarland University
Saarbrücken, Germany
m.jeong@mmci.uni-saarland.de

Ivan Titov
Saarland University
Saarbrücken, Germany
titov@mmci.uni-saarland.de

ABSTRACT

Multiple documents describing the same or closely related sets of events are common and often easy to obtain: for example, consider document clusters on a news aggregator site or multiple reviews of the same product or service. Even though each such document discusses a similar set of topics, they provide alternative views or complimentary information on each of these topics. We argue that revealing hidden relations by jointly segmenting the documents, or, equivalently, predicting links between topically related segments in different documents would help to visualize documents of interest and construct friendlier user interfaces. In this paper, we refer to this problem as multi-document topic segmentation. We propose an unsupervised Bayesian model for the considered problem that models both shared and document-specific topics, and utilizes Dirichlet process priors to determine the effective number of topics. We show that topic segmentation can be inferred efficiently using a simple split-merge sampling algorithm. The resulting method outperforms baseline models on four datasets for multi-document topic segmentation.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing; I.2.7 [**Artificial Intelligence**]: Natural Language Processing

General Terms

Algorithms, Experimentation

Keywords

Topic segmentation, non-parametric Bayesian model

1. INTRODUCTION

Multiple documents conveying the same or closely related information are common on the Web and often easy to obtain. For example, a query to a search engine often returns documents describing the same facts, a document cluster on a news aggregator site covers the same events, and multiple online customer reviews express evaluations of the same product or service. Even though each document from such a set discusses a similar set of topics, they provide alternative views or complimentary information on each of these topics. Additionally, many documents in the set are likely to include fragments specific only to this very document, such as subjective evaluations or information not available to other media sources. That is, each document is likely to consist of a set of fragments related to the entire set and document-specific fragments introducing new facts or opinions. Discovery of these inherent relations between content in such groups of documents could offer a great convenience to users: for instance, an individual following an event through multiple media could find related segments and use them to detect complimentary information or reveal inherent biases of each media source. In this paper, we argue that this hidden relation can be revealed by jointly segmenting documents, or, equivalently, predicting links between topically related segments in different documents. We refer to such a problem as *multi-document topic segmentation*.

Topic segmentation of the multiple related documents is a novel and challenging problem, as previous research has mostly focused on linear segmentation of isolated texts (e.g., [21]). The most straightforward approach would be to use a pipeline strategy, where an existing segmentation algorithm finds topic boundaries of each document in isolation, and then the segments are aligned. Or, conversely, a sentence-alignment stage can be followed by a segmentation stage. However, as we will see in our experiments, these strategies may result in poor segmentation and alignment quality. Therefore, joint modeling of segmentation and alignment is required in multi-document topic segmentation. While unsupervised topic modeling [7, 18, 19] can be used to address this problem, little has been done to directly address joint modeling in the context of topic segmentation.

In this paper we explore generative probabilistic modeling for multi-document topic segmentation. We present a non-parametric Bayesian model for unsupervised joint segmentation and alignment of multiple documents. In contrast to the discussed pipeline approaches, our method leverages the inter-document *lexical cohesion* [20] in modeling multiple related documents. We hypothesize that topically related segments display a compact and consistent lexical distribution, and this insight is encoded in our model. In comparison with related work [9, 33], our method has two important advantages: (1) it induces two types of topics, one type

CIKM'10, October 26–30, 2010, Toronto, Ontario, Canada.

for shared topics and one for document-specific information, and (2) ensures that the effective number of segments can grow adaptively. In addition, we propose a simple split-merge sampling algorithm which is fast to converge in our experiments.

We evaluate the proposed model on texts coming from 4 different domains: news, biographies, biological reports and lectures. In our experiments, we demonstrate that joint modeling is beneficial for the considered problem and the improvement is consistent across the domains. For example, on segmentation and alignment of news and biography documents our model achieves F1-scores of 73.9% and 65.9%, and this compares favorably with 53.7% and 47.0% shown by the pipeline approach.

The remainder of this paper is structured as follows. In Section 2 we give the background of the considered problem and a formal definition of the multi-document topic segmentation task. Section 3 presents our unsupervised generative model which utilizes Dirichlet process priors to determine the effective number of shared and document-specific topics. In Section 4 we describe a simple split-merge Metropolis-Hastings algorithm for our model. Section 5 provides an empirical evaluation of the proposed method. In Section 6 we conclude with examination of additional related work.

2. PROBLEM FORMULATION

In this section we will formulate the task of multi-document topic segmentation. Here, we will assume that we are provided with groups of documents presenting the same or similar information. Our task is, roughly, to induce a segmentation of each document in a group, and to align segments between documents within each group.

2.1 Motivating Example

As an illustration of why uncovering inherent parallel structure may be beneficial, consider Figure 1, where we present three examples taken from three different domains (the chosen datasets are discussed in Section 5). In the figure, the documents are segmented, and, for each group of documents, segments indexed with the same numeral are assumed aligned. For example, in Figure 1a each news article consists of two or more segments describing the same or closely related facts (*motivation*, *results* and *implications* of a study of social media addiction). The second example is drawn from a dataset of biographies. Biographies of the same persons tend to focus on common facts, and we can observe here parts describing *early life*, *education*, and *entrance in politics* of Abraham Lincoln (Figure 1b). Similarly, consider the third example from the lecture domain (Figure 1c), where the pair of documents are an original text and the corresponding commentary given by a language teacher. In this example, the alignment describes the following relation between fragments: a segment of the original text is discussed in the corresponding fragment of the commentary.

We conclude that joint segmentation provides two types of information to the user. First, relations between *shared* segments simplify analysis of multiple information sources: the user can easily compare descriptions of a specific fact provided in different documents. This comparison would help to obtain maximal information about the fact (e.g., only one of the texts in Figure 1a mentioned that the study was held at the University of Maryland), or to reveal biases of the considered media sources (e.g., conservative vs. liberal media). Second, *document-specific* topics are likely to correspond to the information not mentioned in any other document; hence they could be potentially more interesting to the user. Summarizing the above observations, we can conclude that both of these types of information are important, and therefore, the goal in multi-document topic segmentation is to discover both shared and document-specific topics.

2.2 Formal Definition

In this section we formally define our problem. Let us assume that we are given a set of K related documents $\boldsymbol{x} = \{\boldsymbol{x}^k\}_{k=1:K}$.[1] A document $\boldsymbol{x}^k$ is a sequence of bag-of-word vectors $\{\boldsymbol{x}^k_1, \boldsymbol{x}^k_2, \ldots, \boldsymbol{x}^k_{N_k}\}$, where N_k is the length of document k. Note that the bag-of-word vector, $\boldsymbol{x}^k_n$, could be defined at any level of granularity, e.g. word, sentence, or paragraph. While we use both sentences and paragraphs in our experiment, we refer to $\boldsymbol{x}^k_n$ as a sentence for convenience. Our task is to break a sequence of sentences into topically coherent contiguous segments. We represent the segmentation as a sequence of hidden topic variables, $\boldsymbol{z}$. More formally, a topic sequence for document k is defined as $\boldsymbol{z}^k = \{z^k_1, z^k_2, \ldots, z^k_{N_k}\}$, where each topic variable $z \in \boldsymbol{z}$ is described by a topic type $z.t \in \{\textsc{Shared}, \textsc{DocSpecific}\}$ and a topic label $z.l \in \mathbb{N}$. Then, we say that a segment (or sentence) is *global* when $z.t = \textsc{Shared}$ and otherwise, *local*.

The task we will consider is to find segmentation $\boldsymbol{z}$ which results in a compact distribution of words assigned to each topic label $z.l$. The previous work on segmentation [3, 6, 10, 15, 17, 21, 24], has been mostly focused on modeling individual documents, and therefore the produced segmentation did not define topical links between segments in different documents. One approach would be to use a straightforward pipeline strategy, where topic boundaries are first independently found for each document and only then their alignment is predicted. An alternative technique would be to use unsupervised topic segmentation methods [19, 30] which induce collection-level topics, and therefore can produce the desired parallel structure on each set of related documents. However, as we will see in our experiments, both of these strategies fail to achieve good results on the multi-document segmentation task. To tackle their limitations, we propose a new model for our problem.

3. MODEL

We propose a hierarchical Bayesian model for joint segmentation of multiple documents. Similarly to previous work on segmentation of isolated documents [15, 21], our model leverages the lexical cohesion phenomenon [20] in the generative framework. Unlike this previous work, we hypothesize that not only a segment in each isolated document but also each group of topically related segments in different documents displays a compact and consistent lexical distribution, and our generative model leverages this inter-document cohesion assumption.

More formally, we rewrite the document $\boldsymbol{x}^k$ as a set of segments, i.e. $\boldsymbol{x}^k = \{\boldsymbol{s}^k_z\}$, in which $\boldsymbol{s}^k_z = \{\boldsymbol{x}^k_n : z^k_n.l = z.l, z^k_n.t = z.t\}_{n=1:N_k}$. Note that the effective number of segments is unknown. We treat each text in a segment as

[1] In practice, we are normally given multiples sets of such related documents. However, we do not exploit relations between different sets in our model, as our model factorizes over these sets.

[Here was the challenge given to **200** University of Maryland **students** from a variety of majors: Abstain from **social media** for **24 hours** ... no iPhone or text messaging ... no Facebook ...]$_1$ [... the blogs these **students** wrote after the traumatic experience ... "I clearly am **addicted** and ...]$_2$ ... [Being **cut** off from the wired world also meant being **cut** off from news and **information** ...]$_3$

[**According** to a **new study**, **college students** are addicted to cell phones, **social media** and the **Internet** ... titled "**24 Hours**: Unplugged" is based on the experiences of **200 students** ...]$_1$ [**Students** who participated in the study said that abstaining from social **media** made them feel anxious, jittery, antsy and ...]$_2$ [... they **care** about being **cut** off from that instantaneous flow of **information** that comes ...]$_3$ ...

... [American **college students** are hooked on cellphones, **social media** and the **Internet** and showing symptoms similar to drug and alcohol addictions, **according** to a **new study** ...]$_1$ ... ["I clearly am **addicted** and ... said one **student** ... people have become unable to shed their **media** skin." ...]$_2$ ... [... In one extreme example in South Korea reported by the media ...] ...

(a) A news cluster of '24 Hours: Unplugged' in News domain

[**Abraham Lincoln** (February 12, **1809** – April 15, **1865**) served as the **16th President** of the **United States** from ...]$_1$ [**Abraham Lincoln** was **born** on February 12, 1809 ... In 1816, the Lincoln **family** left **Kentucky** ...]$_2$ ... [... the Lincoln Memorial in Washington, D.C. ... sculpture on Mount Rushmore ... To commemorate his 200th birthday in February 2009 ...] ...

[**Abraham Lincoln** (February 12, **1809** – April 15, **1865**) was the **16th President** of the **United States** ... the "Great Emancipator" ...]$_1$ [**Abraham Lincoln** was **born** on ... **Kentucky** ... had one brother and ... his **family** moved to Indiana, and later to Illinois ...]$_2$ ... [... joined the Whig Party ... The **Republican** Party **nominated** him for the **Presidential election** of 1860.]$_3$ ...

[**Lincoln** summarized his early life as "the short and simple annals of the poor." He was **born** in a **Kentucky** log cabin ... grew out of his hard lot as a youth.]$_2$ [The repeal of the Missouri Compromise in 1854 ... and gained national attention in 1858 from his debates with Stephen A. Douglas ... won the **Republican presidential nomination** and was **elected**.]$_3$ ...

(b) A document set for biography of 'Abraham Lincoln' in Biography domain

[I have been **suffering** from back **pain** for **years**.]$_1$ [I've **tried** several **treatments prescribed** by my **doctor**, but nothing has had a **lasting effect**.]$_2$ [I **finally decided** to **try alternative medicine**.]$_3$ [My friend, Amelia, swore by acupuncture.]$_4$ ...

[This episode is called "Alternative Medicine" ... called non-traditional medicine.] [The story begins by me saying that I have **suffered** "from back **pain** for **years**".]$_1$ ["I've **tried** several **treatments** ... To **prescribe**, "**prescribe**," is to give a patient - a **doctor** giving a patient some medicine to take ... **Lasting**, "**lasting**," means permanent or continuing for a long time ...]$_2$ [The story continues by saying that "I **finally decided** to **try alternative medicine**."]$_3$ ...

(c) An episode of 'Alternative Medicine' in Lecture domain

Figure 1: Examples of multi-document topic segmentation. Sentences in brackets refer to a segment, and topically related segments are indexed with the same numeral. Co-occurred words are represented in boldfaces.

draws from a multinomial language model. As we use sparse Dirichlet priors, the model will assign higher probability to segments with more sparse (i.e. compact) distributions of word counts, whereas fragments with a very flat distribution of counts will have lower probability under the model. This property is exactly what we need to model lexical cohesion: from the Figure 1, we may conclude that there is a tendency of word repetition between each shared segment, illustrating our hypothesis of compactness of their joint distribution.

As it was previously considered in the context of contextual text mining [27], our model defines two distinct types of topics, *global* and *local*. We assume that each sentence of a document k can be assigned to a single topic, e.g., a global topic i or a local topic j^k. To model this, we define global language models $\{\boldsymbol{\phi}_i\}_{i=1,2,\ldots}$ and local language models $\{\boldsymbol{\psi}_j^k\}_{j=1,2,\ldots}$ for each document k. Note that we do not assume that we know the effective number of both global topics and local topics. Each language model $\boldsymbol{\phi}_i$ corresponds to a shared topic, and each local language model $\boldsymbol{\psi}_j^k$ corresponds to a single topic of the document k. Thus, each global language model creates a topical link or alignment between sentences assigned to the corresponding topics in different texts of the document set. The local language models, in turn, define a linear segmentation of the remaining unaligned text. In this description, we do not encode the constraint that each topic corresponds to a contiguous segment. For simplicity, we assume that this property can be enforced as a constraint on admissible states of the model. However, as we will discuss later, it can be directly encoded in the generative story.

When analyzing non-structured sources, such as news, a model needs to decide on the number of effective segments, or, in our case, on the number of effective segments of both types. In contrast with most previous systems which assume that the number of segments is given, we do not make this assumption, and use the Dirichlet process priors [16] to determine the effective number of segments. We incorporate them in our model in a way that is similar to how it has previously been done for the latent Dirichlet allocation (LDA) model [34]. Unlike the standard LDA, the topic proportions are chosen not from a Dirichlet prior but from the marginal distribution $GEM(\alpha_0)$ and $GEM(\beta_0^k)$ defined by the *stick breaking construction* [32], where α_0 and β_0^k are the concentration parameter of the underlying Dirichlet process for global and local topics, respectively. $GEM(\cdot)$ defines a distribution of partitions of the unit interval into a countable number of parts.

In Figure 2 we present a representation of the graphical model. The underlying generative process is as follows:

- Draw global topic proportion $\boldsymbol{\alpha} \sim GEM(\alpha_0)$
- For each topic $i \in \{1, 2, \ldots\}$:
  - Draw global language model $\boldsymbol{\phi}_i \sim Dirichlet(\boldsymbol{\phi}_0)$
- For each document $k \in \{1, \ldots, K\}$:
  - Draw local topic proportion $\boldsymbol{\beta}^k \sim GEM(\beta_0^k)$
  - For each topic $j \in \{1, 2, \ldots\}$:
    - Draw local language model $\boldsymbol{\psi}_j^k \sim Dirichlet(\boldsymbol{\psi}_0^k)$
  - Draw $\boldsymbol{\eta}^k \sim Beta(\boldsymbol{\eta}_0^k)$.
  - For each sentence $n \in \{1, \ldots, N_k\}$:
    - Choose topic type $z_n^k.t \sim Bernoulli(\boldsymbol{\eta}^k)$.
    - If $z_n^k.t =$ SHARED then choose topic $z_n^k.l \sim \boldsymbol{\alpha}$; generate words $\boldsymbol{x}_n^k \sim Multinomial(\boldsymbol{\phi}_{z_n^k.l})$.
    - Otherwise, choose topic $z_n^k \sim \boldsymbol{\beta}^k$; generate words $\boldsymbol{x}_n^k \sim Multinomial(\boldsymbol{\psi}_{z_n^k.l}^k)$.

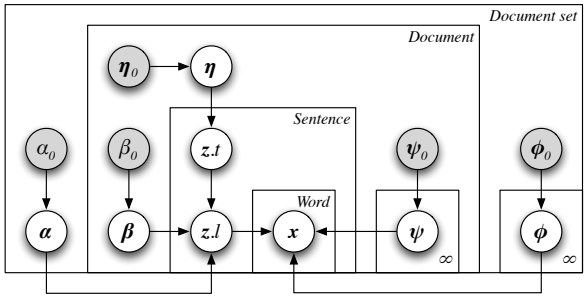

Figure 2: A graphical model representation of our model. Shaded circles denote hyper-parameters.

Our model is similar in spirit to standard topic models [7, 18]. But rather than generating a collection of documents using a mixture of topic-specific language models, we constrain the induced topic labels to correspond to contiguous fragments of the document.

In fact, we can obtain an essentially equivalent model by modifying the generative story. In this generative story instead of choosing a topic for a sentence and then immediately generating words in the sentence, we would first generate a bag of topics for each topic type (local or global), and then generate a random contiguous segmentation using only topic labels form the previously generated bag. This segmentation would be chosen uniformly over all the legal segmentations. A similar technique has been previously considered in [9]. We chose our way of presenting this model, as it results in a much simpler generative story.

4. INFERENCE BY SAMPLING

In this section we describe the inference algorithm we use for our model. Our goal is to estimate the model and obtain a likely segmentation for each document according to this model. As exact inference is intractable, we have to resort to approximate inference methods. We chose to use Markov Chain Monte Carlo (MCMC) methods, as they are easy to implement, do not require additional assumptions about the properties of the distributions, and have been successfully used in previous text segmentation methods [15]. MCMC methods are a class of algorithms which construct a Markov chain with the stationary distribution coinciding with the joint distribution of the considered model. A state of such a chain corresponds to an assignment to all the latent variables of the model with the observed variables held fixed. Therefore, a sample from the joint distribution can be obtained by running this Markov chain for long enough. We use a collapsed MCMC sampler [18] which analytically integrates over all the distributional parameters and samples only assignment of sentences to topics (and topic types). Before starting with our sampling algorithm, we will clarify the computation of the joint distribution under our model, as it is needed to construct our sampler.

4.1 Joint Probability

The joint distribution $P(\boldsymbol{z}, \boldsymbol{x})$ can be decomposed into the product of the likelihood $P(\boldsymbol{x}|\boldsymbol{z})$ and the prior $P(\boldsymbol{z})$. The likelihood $P(\boldsymbol{x}|\boldsymbol{z})$ factorizes over segments and can be written as $P(\boldsymbol{x}|\boldsymbol{z}) =$

$$\prod_{i=1}^{I} P_{seg}(\{\boldsymbol{x}_n^k : z_n^k.l = i, z_n^k.t = \textsc{Shared} \text{ for all } k\}|\boldsymbol{\phi}_0) \cdot \prod_{k=1}^{K} \left(\prod_{j=1}^{J_k} P_{seg}(\{\boldsymbol{x}_n^k : z_n^k.l = j, z_n^k.t = \textsc{DocSpecific}\}|\boldsymbol{\psi}_0^k) \right), \quad (1)$$

where the effective number of topics, I and J_k, are selected using the Dirichlet process priors. Now, we consider, for both shared and document-specific segments, the probability of each segmentation, $P_{seg}(\boldsymbol{y}|\boldsymbol{\lambda})$, in which $\boldsymbol{y}$ indicates a set of sentences assigned to one type of topics, and $\boldsymbol{\lambda}$ is a Dirichlet prior (either $\boldsymbol{\phi}_0$ or $\boldsymbol{\psi}_0^k$). Given the text $\boldsymbol{y}$, let $c(w, \boldsymbol{y})$ be the number of appearances of word w in $\boldsymbol{y}$, where w ranges from 1 to the vocabulary size W. For $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_W)$, the probability of a segment is equal to:

$$\begin{aligned} P_{seg}(\boldsymbol{y}|\boldsymbol{\lambda}) &= \int_{\boldsymbol{\theta}} P(\boldsymbol{y}|\boldsymbol{\theta}) P(\boldsymbol{\theta}|\boldsymbol{\lambda}) d\boldsymbol{\theta} \\ &= \frac{\Gamma(m)}{\Gamma(l+m)} \prod_{w:c(w,\boldsymbol{y})\geq 1} \frac{\Gamma(c(w,\boldsymbol{y}) + \lambda_w)}{\Gamma(\lambda_w)} \end{aligned} \quad (2)$$

where $\boldsymbol{\theta}$ is a hidden language model, $l = \sum_w^W c(w, \boldsymbol{y})$ is the document length, $m = \sum_w^W \lambda_w$ is the sum of the Dirichlet priors parameters ($W\lambda$ if the prior is symmetric), and Γ is the gamma function. This distribution (2) is known as the Dirichlet compound multinomial distribution or the multivariate Polya distribution, the result of integrating out multinomial parameters $\boldsymbol{\theta}$ with a Dirichlet prior $\boldsymbol{\lambda}$ [4]. Note that, we can evaluate this probability as a function of non-zero counts $c(w, \boldsymbol{y})$ only, since $\Gamma(c(w, \boldsymbol{y}) + \lambda_w)/\Gamma(\lambda_w) = 1$ if $c(w, \boldsymbol{y}) = 0$.

Similarly, to the standard Dirichlet process mixture model, the latent segmentation $\boldsymbol{z}$ can be sampled using the generalized Polya urn scheme (for more details, we refer the readers to [5] and [16]), which yields the following prior distribution: $P(\boldsymbol{z}) =$

$$(\alpha_0)^I \frac{\prod_{i=1}^{I}(c_{\boldsymbol{z},i} - 1)!}{\prod_{n=1}^{N^s}(\alpha_0 + n - 1)} \prod_{k=1}^{K} (\beta_0^k)^{J_k} \frac{\prod_{j=1}^{J_k}(c_{\boldsymbol{z},j} - 1)!}{\prod_{n=1}^{N_k^d}(\beta_0^k + n - 1)} \quad (3)$$

where $c_{\boldsymbol{z},i}$ is the number of sentences assigned to topic i in $\boldsymbol{z}$, and the total number of global and local topics is defined as $N^s = \sum_{i=1}^{I} c_{\boldsymbol{z},i}$ and $N_k^d = \sum_{j=1}^{J_k} c_{\boldsymbol{z},j}$. As we will see in the following section, our MCMC sampler will only modify adjacent segments at each sampling move. If we denote by $\boldsymbol{z}_{ij}$ the subset of $\boldsymbol{z}$ which consists of two segments i and j, the probability of this change in segmentation simplifies considerably as $P(\boldsymbol{z}_{ij}) \propto (c_{\boldsymbol{z},i} - 1)!(c_{\boldsymbol{z},j} - 1)!$. We will use this fact to sample a new state of the Markov chain efficiently.

4.2 Split-Merge Sampler

We use MCMC sampling techniques to infer our model. Although the Gibbs sampling method is straightforward and easy to implement, it can be slow to mix in our case. In its basic form it assigns only one hidden variable at a time, making it difficult to introduce large changes in the topic segmentation, as local minima in the joint probability are often separated by regions of low probability. We instead use a split-merge algorithm [11, 22], which relies on Metropolis-

Hastings (MH) proposals to merge two segments into one or split an existing segment into two.

At each iteration of the MH algorithm, a new potential segmentation $\boldsymbol{z}'$ is drawn from a proposal distribution $\pi(\boldsymbol{z}'|\boldsymbol{z})$, where $\boldsymbol{z}$ is the current segmentation. The proposal state $\boldsymbol{z}'$ is accepted with the probability

$$\min\left(1, \frac{P(\boldsymbol{z}'|\boldsymbol{x})}{P(\boldsymbol{z}|\boldsymbol{x})}\frac{\pi(\boldsymbol{z}|\boldsymbol{z}')}{\pi(\boldsymbol{z}'|\boldsymbol{z})}\right). \quad (4)$$

In order to implement the MH algorithm for our model, we need to define the set of potential *moves* (i.e. admissible changes from $\boldsymbol{z}$ to $\boldsymbol{z}'$), and the proposal distribution over these moves. For the standard Dirichlet process mixture model, two standard proposals, *split* and *merge* of the clusters, are sufficient [22]. In our case, however, a more complex set of moves is required.

The set of moves we consider in this work is divided into two groups: *split-merge* and *increase-decrease*. The former one is the set of proposal moves that generates a new topic or removes one of the topics, and the latter one reassigns the topic variables in a single considered document. More specifically, we define four split-merge moves (two for each topic type):

(a) *GlobalSplit*: A move that splits an existing global topic i into two. First, all the segment labeled with i are collected from the set of documents. Then each such segment is split at one of the spanned sentences. Topics i and $I+1$ are then assigned to the new segments. (Figure 4a)

(b) *GlobalMerge*: A move that removes an existing global topic, and merges its sentences into a neighboring segment. Given a global topic i, this move selects all segments corresponding the topic i and merges each segment into one of the adjacent segments (either local or global ones). (Figure 4b)

(c) *LocalSplit*: A move that splits an existing segment in a single document into two, one of which will be a new local topic $J_k + 1$. (Figure 4c)

(d) *LocalMerge*: A move that removes an existing local topic, and merges it into a neighboring segment. (Figure 4d)

Although the split-merge moves allow for relatively complex joint segmentation, they constrain unnecessarily the segmentations: namely, all the documents are assumed to have the same sequence of global topics. When there are only two relevant documents in the considered document set (i.e. $K = 2$), the split-merge moves are sufficient, but in more general cases $K > 2$ (e.g. Figure 1a and 1b), we need a way to assign a different sequence of global topics to each document. In addition, to decrease the mixing time, a move which shifts a segment border is introduced:

(e) *Increase*: A move that introduces a global topic into the set of topics discussed in a given document. This move splits an existing segment $\boldsymbol{s}_z^k$ into two, and assigns a global topic i chosen from $\{i : z_n^k.l \neq i \text{ for all } n, 1 \leq n \leq N_k\}_{i=1:I}$ to a new segment. (Figure 4e)

(f) *Decrease*: A move that removes a global topic from a set of topics discussed in a given document. Given a segment $\boldsymbol{s}_{z_1}^k$ and an adjacent segment $\boldsymbol{s}_{z_2}^k$, where z_1 is the considered global segment, this move merges them, and assigns z_2 to merged sentences. (Figure 4f)

Algorithm: MultiSeg
Input: $\boldsymbol{x}$ [document set]
Initialize $\boldsymbol{z}$ with one global topic, and $iter := 1$
repeat
  Randomly select two sentences, $z_i^{k_1}$ and $z_j^{k_2}$
  if $z_i^{k_1}.t =$ SHARED **then**
    $\boldsymbol{z} \leftarrow \begin{cases} \text{GlobalSplit}(\boldsymbol{x}, z_i^{k_1}) & z_i^{k_1} = z_j^{k_2} \\ \text{GlobalMerge}(\boldsymbol{x}, z_i^{k_1}) & \text{otherwise} \end{cases}$
  end if
  for all document $k \in \{1, \ldots, K\}$ **do**
    Randomly select two segments z_i and z_j
    if z_i and z_j are the same segment **then**
      $r \sim Uniform(0, 1)$
      $\boldsymbol{z}^k \leftarrow \begin{cases} \text{LocalSplit}(\boldsymbol{x}^k, z_i) & r = 0 \\ \text{Increase}(\boldsymbol{x}^k, z_i) & r = 1 \end{cases}$
    else if z_i and z_j are adjacent **then**
      $\boldsymbol{z}^k \leftarrow \begin{cases} \text{LocalMerge}(\boldsymbol{x}^k, z_i, z_j) & z_i.t = \text{SHARED} \\ \text{Decrease}(\boldsymbol{x}^k, z_i, z_j) & z_i.t = \text{DOCSPECIFIC} \end{cases}$
    end if
    Randomly select a segment z; $\boldsymbol{z}^k \leftarrow \text{Shift}(\boldsymbol{x}^k, z)$
  end for
  $iter \leftarrow iter + 1$
until converge or $iter \leq iter^{max}$
Output: $\boldsymbol{z}$ [segmentation]

Figure 3: Inference algorithm (MultiSeg)

(g) *Shift*: A move that shifts the segment boundary. Given a segment $\boldsymbol{s}_z^k$, this move shifts its border. (Figure 4g)

Note that this set of moves changes only the number of topics discussed in a single document, whereas moves *GlobalSplit* and *GlobalMerge* affect the entire set of documents.

An important property of the proposal moves defined here is that they only depend on a pair of adjacent segments in each document. Therefore, the acceptance ratio can be computed efficiently. Formally, the ratio of the posterior distributions $P(\boldsymbol{z}'|\boldsymbol{x})/P(\boldsymbol{z}|\boldsymbol{x})$ can be rewritten as the ratio of the likelihoods and the priors, i.e. $[P(\boldsymbol{x}|\boldsymbol{z}')/P(\boldsymbol{x}|\boldsymbol{z})] \cdot [P(\boldsymbol{z}')/P(\boldsymbol{z})]$, and we can calculate each ratio over the partition $\boldsymbol{z}_{ij}$ or the set of partitions $\{\boldsymbol{z}_{ij}\}$.

Similarly, we use the following ratio of proposal distributions:

$$\frac{\pi(\boldsymbol{z}|\boldsymbol{z}')}{\pi(\boldsymbol{z}'|\boldsymbol{z})} = \begin{cases} 1/\left[\left(\frac{1}{2}\right)^{c_{\boldsymbol{z}',i}+c_{\boldsymbol{z}',j}-2}\right] & \text{for split} \\ \left(\frac{1}{2}\right)^{c_{\boldsymbol{z},i}+c_{\boldsymbol{z},j}-2} & \text{for merge} \\ 1 & \text{otherwise.} \end{cases} \quad (5)$$

In our experiments, we start with one global topic, and repeatedly sample the potential state $\boldsymbol{z}'$ and evaluate it using Eq. (1)-(5). If the proposal is accepted, $\boldsymbol{z}'$ is selected as the next state of the Markov chain, and otherwise the original segmentation $\boldsymbol{z}$ is preserved. The algorithm is presented in Figures 3-4.

5. EXPERIMENTS

In this section we present quantitative and qualitative experiments. For quantitative experiments we show that our model outperforms baselines on four different datasets. For qualitative analysis we inspect types of errors made by our model and the considered baselines, and discuss properties of topics discovered by our method.

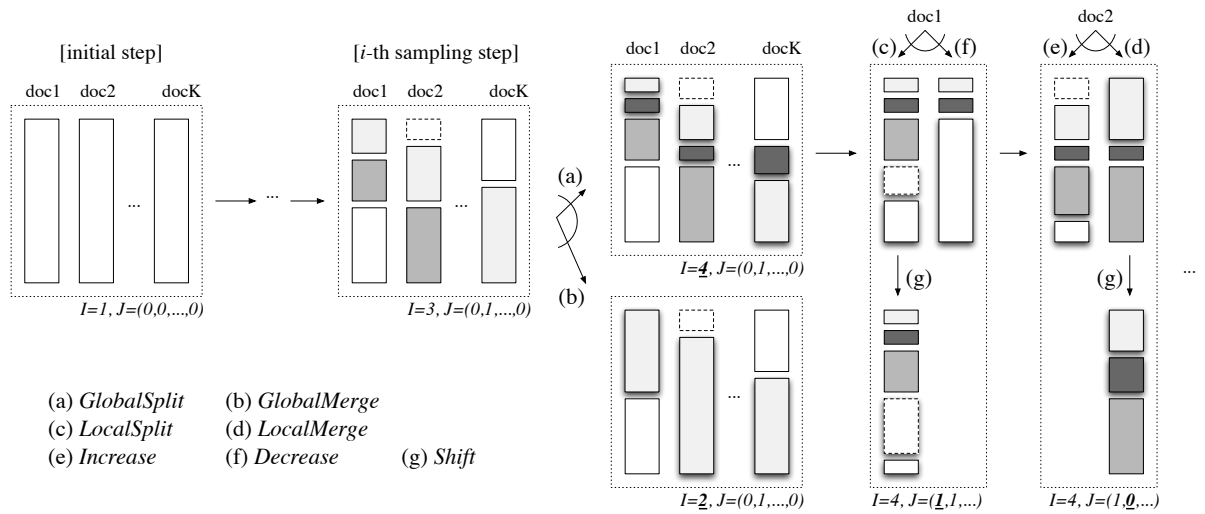

Figure 4: An illustration of the MultiSeg algorithm. Dashed line boxes indicate the document-specific parts, and solid line boxes denote the shared parts. The considered potential moves are highlighted. The values on the bottom of groups represent the effective number of topics at each sampling move: I indicates the number of global topic and each value for J is the number of local topics (i.e. segments) for each document.

5.1 Datasets

We evaluate our method on four datasets: News, Biography, Report and Lecture (Table 1). All the datasets were collected and annotated by human annotators, and automatically tokenized. The number of annotated segments vary across the datasets and across groups of related documents within each dataset. For the Report and Lecture dataset, a document is a sequence of automatically-split sentences, but for the News and Biography domains we treat each document as a sequence of paragraphs separated by the HTML tag `<p>`. The paragraph break information was not available for the Report and Lecture datasets and, therefore, we could not apply our approach at this granularity level. For the News domain we remove only punctuation symbols, whereas for the Biography, Report and Lecture domains we also use a standard list of stop-words and a stemmer.

In the News and Biography domains, each group of documents refers to the same event or person (see examples in Figure 1a and 1b). To create the News dataset, we collected document clusters of 50 news events over the period of April 28 – May 6, 2010 from *science and technology* section in `news.google.com`. The size of each set of related documents in the News dataset varies between 2 and 6. For the Biography dataset, we collected biographies of 30 persons from four web sites; `en.wikipedia.org`, `simple.wikipedia.org`, `biography.com`, and `notablebiographies.com`. Documents in these datasets contain both shared and document-specific segments, and the proportion of paragraphs assigned to global topics vary across documents in each dataset. Our purpose is collecting this set of texts to uncover hidden shared topics as well as document-specific topics by jointly segmenting multiple news stories and biography articles.

The Report dataset consists of reports describing a plant growth lab, an assignment for a biology class [33]. There are only two sets of related documents: for the first set of documents (100 examples) there are two annotated global segments (*an introduction of plant hormones* and *description of the experiment*), for the second set (60 documents) there are four global segments. Note that this dataset is annotated only with global segments.

Table 1: Datasets: the number of sets of related documents (#Set), the total number of documents (#Doc), the average number of sequences per document (paragraphs for News and Biography, sentences for Report and Lecture) per domain (AvgSeq), the average number of annotated segments per document (AvgSeg) and the average size of vocabulary per set (AvgVoc).

Domain	#Set	#Doc	AvgSeq	AvgSeg	AvgVoc
News	50	184	11.1	3.0	563.8
Biography	30	120	34.5	8.1	2,222.6
Report	2	160	13.8	2.4	1,440.0
Lecture	200	400	51.4	18.2	209.0

For the Lecture domain, we use a dataset of 'English as a second language' (ESL) podcast [28] containing 200 episodes (podcast no. 174 $\sim$ 373 posted over June 19, 2006 – May 16, 2008) from `www.eslpod.com`. Each episode consists of two parts: a *story* (an example monologue or dialogue) and an explanatory *lecture* discussing meaning and usage of English expressions appearing in the story (see example in Figure 1c). The objective here is to divide the lecture transcript into discourse units (i.e. focused paragraphs of the lecture) and to align each unit to the related segment of the story. Predicting the segmentation and alignment for the ESL podcast could be the first step in development of an e-learning system and a podcast search engine for ESL learners.

5.2 Evaluation

To measure the quality of segmentation predicted by our model, we compute the precision and recall scores of a pre-

dicted segmentation against a reference segmentation. Precision is the fraction of correctly identified and aligned segments among all the predicted segments; recall is the fraction of correct segments that are identified by the algorithm. We also present the F1-score in our results, which is the harmonic mean of recall and precision. Although precision and recall are standard measures for many information retrieval tasks, some researchers [29] argue that these measures are not always appropriate, as they are not sensitive to near misses, i.e. small shifts of segmentation boundaries. Consequently, we also use two standard metrics, Pk [3] and WindowDiff [29], but both of these metrics disregard topic labels and alignment, and score only segmentation. In addition, we present the ratio of the predicted number of segments to the reference number of segments (NR). This ratio will be close to 1 if the method is able to detect the correct number of segments.

Our MultiSeg algorithm presented in Figures 3 and 4 has two important properties: (1) it can model both document-specific and shared topics and (2) each document can include a different subset of shared topics, as result of application of the increase-decrease moves. To quantify the benefits of these two properties of our method, we compare our model not only with other baselines but also with three more restricted versions of the model itself. The resulting 4 versions of our model are:

- MultiSeg-1: This model induces no document-specific topics (we set $\boldsymbol{\eta}_0^k = (1, 0)$) and the sampler does not perform increase-decrease moves. Thus, the predicted segmentation will include the same sequence of global topics for every document.
- MultiSeg-2: The sampler for this model supports increase-decrease moves, but still no document-specific topics are included.
- MultiSeg-2': This model induces the document-specific topics, but no increase-decrease moves are permitted. We set $\boldsymbol{\eta}_0^k$ to $(5, 1)$ which reflects the intuition that most sentences should be assigned to shared topics.
- MultiSeg-3: Our full model, $\boldsymbol{\eta}_0^k = (5, 1)$.

The baseline models are:

- K-means: This method uses 1000 iterations of the *k-means* clustering algorithm with tf-idf cosine similarity to induce clusters of similar sentences (paragraphs for Report and News). The number of clusters is chosen according to an oracle. As k-means is sensitive to initialization, we use k-means++ algorithm [2], which uses a heuristic strategy to find good initialization.
- Pipeline: This method is a cascaded approach. First, the BayesSeg method [15], a state-of-the-art method for single-document segmentation, is used to segment isolated documents, then the predicted segments are clustered using the k-means algorithm. The numbers of segments and clusters are chosen according to an oracle.
- HTMM: This baseline is given by the hidden topic Markov model [19]. Their method can jointly model segmentation and alignment as it models topic distributions on the collection level. However, it uses a Markov chain on topics and, therefore, it cannot constrain topics to correspond to contiguous fragments of documents. We use the publicly available implementation[2], with hyperpriors equal to the ones used in the original work [19].

Table 2: Result on News dataset. Columns are precision (Prec), recall (Rec), and F1-score (F1); Pk and WindowDiff (WD); the ratio of the predicted number of segments to the reference number of segments (NR).

	Prec	Rec	F1	Pk	WD	NR
Coarse	.309	1.00	.472	.335	.335	0.41
Fine	1.00	.129	.228	.502	.755	4.15
K-means	.402	.718	.516	.436	.502	1.76
Pipeline	.393	.849	.537	.318	.326	0.70
HTMM	.378	.735	.499	.415	.467	1.47
MultiSeg-1	.554	**.860**	.674	.287	.307	0.90
MultiSeg-2	.624	.831	.713	.262	.288	0.89
MultiSeg-2'	.572	.789	.663	.275	.309	**1.09**
MultiSeg-3	**.712**	.769	**.739**	**.236**	**.275**	1.14

Table 3: Result on Biography dataset

	Prec	Rec	F1	Pk	WD	NR
Coarse	.086	1.00	.159	.233	.233	0.19
Fine	1.00	.136	.239	.351	.501	3.99
K-means	.254	.500	.337	.312	.440	2.89
Pipeline	.338	.768	.470	.186	.208	0.81
HTMM	.208	.443	.283	.312	.431	2.51
MultiSeg-1	.335	**.879**	.485	.202	.213	0.77
MultiSeg-2	.548	.853	**.667**	.174	**.191**	0.89
MultiSeg-2'	.395	.684	.501	.175	.193	0.85
MultiSeg-3	**.577**	.769	.659	**.168**	.192	**0.97**

In order to infer segmentation with our model, we run the inference algorithm from five randomly chosen initialization states, and take the 100,000th iteration of each chain as a sample. Results are then averaged over these five samples. Similarly, results of the three baselines are also initialized with five random states and averaged. While the global topic Dirichlet hyperprior $\boldsymbol{\phi}_0$ is set to 0.2, all document-specific topic Dirichlet priors $\boldsymbol{\psi}_0^k$ are set to 0.1, encouraging sparser distributions. All the concentration parameters, α_0 and β_0^k, are set to 0.1.

Our method is implemented in Java and the experiments were executed on a dual-core 3.06 GHz machine. The running time varies across domains from 2 seconds per document on the Lecture domain to 19 seconds per document on the Biography domain.

5.3 Results

First, we show that our proposed model is appropriate for multi-document topic segmentation (Table 2 - 5), and then we provide some more detailed analysis of its behavior.

In the tables, the first two lines present results of the minimal baselines: one which places all sentences in the same segment (Coarse) and one which assigns each sentence to its own segment (Fine). We use these baselines only to give a better sense for the performance metrics. For all the datasets, the segmentation errors of the Pipeline approach, measured

[2]http://code.google.com/p/openhtmm/

Table 4: Result on Report dataset

	Prec	Rec	F1	Pk	WD	NR
Coarse	.375	1.00	.546	.227	.227	0.45
Fine	1.00	.003	.005	.755	.958	9.63
K-means	.609	.669	.638	.509	.594	3.35
Pipeline	.401	.772	.526	.254	.255	0.76
HTMM	.578	.677	.623	.450	.515	2.58
MultiSeg-1	.693	**.916**	.789	.089	.089	0.90
MultiSeg-2	**.796**	.858	**.826**	**.075**	**.077**	**1.05**

Table 5: Result on Lecture dataset

	Prec	Rec	F1	Pk	WD	NR
Coarse	.061	1.00	.115	.199	.199	0.06
Fine	1.00	.227	.370	.313	.463	2.93
K-means	.454	.641	.531	.260	.357	2.17
Pipeline	.422	.763	.543	.173	.190	0.91
HTMM	.414	.599	.489	.228	.270	1.14
MultiSeg-1	.708	**.828**	.764	**.138**	**.160**	**1.04**
MultiSeg-2'	**.772**	.777	**.775**	.141	.172	1.22

with Pk and WD metrics, are generally lower than those of K-means and HTMM (lower numbers for Pk and WD metrics correspond to the better segmentation quality). For the K-means baseline every sentence is considered in isolation, whereas HTMM captures only Markovian dependencies in topic transitions and disregards contiguity constraints. Both of these properties result in over-generation of segments (not topics), even though the true number of clusters and the true number of topics were given to these baselines by an oracle. This result suggests that encoding the discourse-level constraint, that topics are normally non-recurring in text [20], is crucial for the success of the method. For all the domains except for the Report domain, Pipeline performs better than two other baselines. The results of the HTMM model come close to those of K-means.

Our model MultiSeg-1 substantially outperforms the baselines on all the datasets. The difference is statistically significant at $p < .001$ level measured with the permutation test [14]. The significant improvement over the Pipeline results demonstrates benefits of joint modeling for the considered problem. Moreover, additional improvement is obtained by using different sequences of global topics for different documents (MultiSeg-2), and from using document-specific topics (MultiSeg-2'). Note that for two domains more restricted models are sufficient: no document-specific topics are needed for the Report dataset and no increase-decrease move are required for the Lecture dataset. Table 2 - 3 show that performance is improving as the model is becoming more and more flexible. For the News domain, our full model (MultiSeg-3) performs better than other constrained versions. Although the F1-score of MultiSeg-2 is slightly higher than that of the full model on Biography domain, we can observe that this does not translate into improvement in recall, as the segmentation produced by MultiSeg-2 is too fine-grained (see the NR column).

Next, we compare our method to previous work [33], which also considered the Report dataset. However, their dataset, experimental set-up and evaluation metrics are somewhat different, and, therefore, we had to replicate their set-up and ran an additional experiment. They used a subset of the Report dataset that consists of 102 documents and two topics. Our model, MultiSeg-2, achieves the error rate of 3.4% whereas they reported the error rate of 2.8%. This difference is certainly not significant, and also it is important to note that they assume that the number of segments is given, whereas we detect the number of segments automatically. One reason for the very high accuracy of both models on this dataset is that it is easy to induce language models for 2 segment types given a large set of related documents. These 2 language models present compact but, to significant degree, non-overlapping distribution over the vocabulary. For example, an introduction part in this dataset often discusses background knowledge on plant hormones using a specific set of biological terms (e.g., *cell*, *auxin* and *cytokinin*), and the experimental parts is characterized by terms mostly related to plant growth or stages of the experiment (e.g., *treatment*, *experiment*, and *data*).

	doc1	doc2
Reference	1234...	00000122222222222344444444444...
MultiSeg-2'	1234...	00000122222222333344444444000...
MultiSeg-1	1234...	11111122222222333344444444444...
HTMM	2434...	15511241155662227548987784411...
Pipeline	0004...	22222222222222222244444444222...

Figure 5: Comparison of segmentation on 'Alternative Medicine' episode in the Lecture domain

In order to better understand types of errors made by the different methods, we now turn to qualitative analysis. In Figure 5 we show predictions of our model against two baselines on an example from the Lecture domain. Each digit indicates a distinct topic identifier assigned to the sentence, the digit 0 denotes any document-specific segment.[3] The topic assignments predicted by HTMM are non-contiguous and therefore the number of segments is exceedingly large. Conversely, the Pipeline method results in too coarse-grained segmentation. MultiSeg-1 is only able to model shared topics, and the first sequence of errors it made (sentence 1-5) is due to this deficiency. Extending the sampler for our model with two new moves (MultiSeg-2') remedies this problem, but errors are still made due to over-generation of the document-specific topics. In fact, often there is more than a single granularity level for a 'good' segmentation, and prior knowledge encoded in the form of prior distributions (or a limited amount of labeled data) may be required to force the model to use the required granularity level.

Top words for 3 topics discovered by our model in News and Biography domains are presented in Table 6. To improve readability, we manually de-stemmed words. The topics are also manually labeled to reflect our interpretation of their meaning. The top words suggest that the discovered segments indeed correspond to semantically coherent topics. Importantly, many of these top words (and the corresponding semantic topics) are irrelevant to any other document in the collection (for example, "motivation of the study" or "Civil war") and therefore they are unlikely to be induced if topic modelling is done at the collection level [7].

[3]We rearranged the randomly ordered topic numbers to improve readability.

Table 6: Top words extracted from our model. All words are listed in order of descending values.

(a) '24 Hours: Unplugged' in News

Label	Top words
Motivation of the study	media addicted university 24 college hours center journalism professor director study social new going according symptoms unplugged
Personal impact	friends gives feeling comfort did luxuries quite secluded texting felt life instant school thousands fact able communicate unbearable
Implication	care families tied single device application outlet news large going friends information world researcher instantaneous flow comes sides worked

(b) 'Abraham Lincoln' in Biography

Label	Top words
Early life	illinois moved father salem mary springfield family law county lincolns todd new abraham died sangamon began did served kentucky 1837 later robert
Political career	douglas party republican slavery national won convention 1860 new power votes candidate election divided slave whig stephen debates senate decision
Civil war	slaves union proclamation gettysburg 1863 freed war grant nation address emancipation soldiers armies army battle people freeing draft cemetery

6. RELATED WORK

Topic segmentation has been an active area of research in the last decade. However, previous research has mostly focused on the linear segmentation of isolated texts [3, 6, 10, 15, 17, 21, 24]. Our work can be regarded as an extension of the Bayesian segmentation model [15] for the multi-document topic segmentation problem. A preliminary version of this work was presented in [23], where a more restricted inference algorithm was considered, and the technique was evaluated only on the simpler ESL dataset.

A similar problem has been studied in the context of topic detection and tracking (TDT) [1]. There, after segmenting a news stream into stories discussing a single topic (segmentation stage), similar stories are grouped into a cluster (detection stage). Joint segmentation and detection are closely related to the multi-document topic segmentation problem considered in this paper. However, most previous studies have focused on solving these tasks independently. Moreover, rather than modeling multiple related documents, TDT often assumes that only a single long stream is available.

Although interest in exploiting multiple related documents has been growing recently, the task of multi-document topic segmentation has not received much attention. We are aware of only two previous methods [33, 9] focusing on joint segmentation and alignment of multiple texts. In [33] it is shown that leveraging multiple documents improves the segmentation performance. Even though the problem definition is similar, there are differences in the set-up, and our approaches are also quite different. First, we assume that there exist both document-specific and shared topics, and not all the shared topics are necessary mentioned in each document, whereas they assume that the same set of topics is discussed in each document. We believe that our problem formulation is not only more general but also more realistic in multi-document topic segmentation. Second, we do not assume that the number of segments is provided to the model, whereas they used the actual number as an input. Though it is a common practice in segmentation community, it is not a realistic assumption in many cases (e.g., consider analysis of newswire). Our model uses Dirichlet process priors to determine the effective number of topics. Finally, while they use similarity functions and an iterative greedy algorithm [33], we use the generative framework to model lexical cohesion.

The closest model to ours is that of [9] which is focused on learning preferred ordering of topics in a given collection. As they were considering modeling of wikipedia text collections, their main goal was to detect a compatible set of topics for a large collection of documents. In our case we focused on detecting shared and document-specific topics for a set of related documents in less structured data-sources (e.g., news). Here, a significant part of each document is not related to any shared topic (and to any other document), and, as our results suggest, modeling both types of topics is beneficial. Also, conceptually these set-ups are quite different, as we can assume that each segment is weakly equivalent to an aligned segment in other documents, and instead of generating all aligned segments from a single language model, a translation model can be used to generate them jointly. This approach, which we regard as a potential future direction, is not appropriate for modeling general collections of documents but only appropriate for modeling groups of related documents. Again, the authors also assume that the number of topics is given, which may be a reasonable assumption for modeling collection level topics, but clearly results in sub-optimal performance with small sets of related documents.

Our work is also related to research on multi-document summarization [8, 13, 31, 35]. In multi-document summarization, the goal is to generate a summary consisting of sentences extracted from a set of documents. Our work is different in that we do not try to extract summary sentences but rather aim to find coherent fragments with maximally overlapping lexical distributions. Similarly, passage retrieval (e.g., [25]) and abstract/document alignment [12, 26] are also related but they focus on selection of most relevant passages and sentences given a query (or an abstract) rather than on jointly segmenting multiple related documents.

7. SUMMARY AND FUTURE WORK

We studied the problem of multi-document topic segmentation, where the goal is to jointly segment multiple documents detecting both aligned and non-aligned segments. Our model achieves favorable results on four datasets, demonstrating that the use of the Dirichlet process priors and structured topic models can lead to improved segmentation quality. Accurate prediction of these hidden relations between documents would open interesting possibilities for constructing friendlier user interfaces. One example being an application which, given a document cluster of a news event, produces a graph-based visualization of the shared topic segments.

In future research, we plan to investigate models which are specifically suited to jointly modeling small sets of documents, as word overlap between related segments in such documents may be too small and not sufficient to detect their relevance. This approach may borrow techniques from statistical machine translation and document summariza-

tion. Another interesting direction would be integration of user feedback to decide on the necessary granularity level.

8. ACKNOWLEDGMENTS

The authors acknowledge the support of the Excellence Cluster on Multimodal Computing and Interaction (MMCI), and also thank Andrew Gargett and the anonymous reviewers for their valuable comments, and Bingjun Sun for providing the Report dataset.

9. REFERENCES

[1] J. Allan, editor. *Topic detection and tracking: event-based information organization*. Kluwer Academic Publishers, Norwell, MA, USA, 2002.

[2] D. Arthur and S. Vassilvitskii. k-means++: the advantages of careful seeding. In *Proc. of the Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1027–1035, 2007.

[3] D. Beeferman, A. Berger, and J. Lafferty. Statistical models for text segmentation. *Machine Learning*, 34(1–3):177–210, 1999.

[4] J. Bernardo and A. Smith. *Bayesian Theory*. John Wiley, 1994.

[5] D. Blackwell and J. B. MacQueen. Ferguson distributions via Polya urn schemes. *Annals of Statistics*, 1:353–355, 1973.

[6] D. M. Blei and P. J. Moreno. Topic segmentation with an aspect hidden Markov model. In *Proc. of ACM SIGIR*, pages 343–348, 2001.

[7] D. M. Blei, A. Ng, and M. I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[8] G. Carenini, R. Ng, , and A. Pauls. Multi-document summarization of evaluative text. In *Proc. of EACL*, 2006.

[9] H. Chen, S. Branavan, R. Barzilay, and D. R. Karger. Global models of document structure using latent permutations. In *Proc. of NAACL*, pages 371–379, 2009.

[10] F. Y. Y. Choi, P. Wiemer-Hastings, and J. Moore. Latent semantic analysis for text segmentation. In *Proc. of EMNLP*, pages 109–117, 2001.

[11] D. B. Dahl. Sequentially-allocated merge-split sampler for conjugate and nonconjugate Dirichlet process mixture models, 2005.

[12] H. Daumé and D. Marcu. A phrase-based hmm approach to document/abstract alignment. In *Proc. of EMNLP*, pages 137–144, 2004.

[13] H. Daumé and D. Marcu. Bayesian query-focused summarization. In *Proc. of ACL*, pages 305–312, 2006.

[14] P. Diaconis and B. Efron. Computer-intensive methods in statistics. *Scientific American*, pages 116–130, 1983.

[15] J. Eisenstein and R. Barzilay. Bayesian unsupervised topic segmentation. In *Proc. of EMNLP*, pages 334–343, 2008.

[16] T. S. Ferguson. A Bayesian analysis of some non-parametric problems. *Annals of Statistics*, 1:209–230, 1973.

[17] M. Galley, K. R. McKeown, E. Fosler-Lussier, and H. Jing. Discourse segmentation of multi-party conversation. In *Proc. of ACL*, pages 562–569, 2003.

[18] T. L. Griffiths and M. Steyvers. Finding scientific topics. *Proc. of the National Academy of Sciences of the United States of America*, 101(Suppl 1):5228–5235, April 2004.

[19] A. Gruber, Y. Weiss, and M. Rosen-Zvi. Hidden topic Markov models. In *Proc. of AISTATS*, 2007.

[20] M. A. K. Halliday and R. Hasan. *Cohesion in English*. Longman, 1976.

[21] M. Hearst. Multi-paragraph segmentation of expository text. In *Proc. of ACL*, pages 9–16, 1994.

[22] S. Jain and R. Neal. A split-merge Markov chain Monte Carlo procedure for the Dirichlet process mixture model. *Journal of Computational and Graphical Statistics*, 13:158–182, 2004.

[23] M. Jeong and I. Titov. Unsupervised discourse segmentation of documents with inherently parallel structure. In *Proc. of the ACL 2010 Conference Short Papers*, pages 151–155, July 2010.

[24] X. Ji and H. Zha. Domain-independent text segmentation using anisotropic diffusion and dynamic programming. In *Proc. of ACM SIGIR*, pages 322–329, 2003.

[25] X. Liu and W. B. Croft. Passage retrieval based on language models. In *Proc. of CIKM*, pages 375–382, 2002.

[26] D. Marcu. The automatic construction of large-scale corpora for summarization research. In *Proc. of ACM SIGIR*, pages 137–144, 1999.

[27] Q. Mei and C. Zhai. A mixture model for contextual text mining. In *Proc. of the ACM SIGKDD*, pages 649–655, 2006.

[28] H. Noh, M. Jeong, S. Lee, J. Lee, and G. G. Lee. Script-description pair extraction from text documents of English as second language podcast. In *Proc. of Conf. on CSEDU*, 2010.

[29] L. Pevzner and M. Hearst. A critique and improvement of an evaluation metric for text segmentation. *Computational Linguistics*, 28(1):19–36, 2002.

[30] M. Purver, K. Kording, T. Griffiths, and J. Tenenbaum. Unsupervised topic modelling for multi-party spoken discourse. In *Proc. of EMNLP*, pages 17–24, 2006.

[31] D. R. Radev, H. Jing, M. Stys, and D. Tam. Centroid-based summarization of multiple documents. *Information Processing & Management*, 40(6):919–938, 2004.

[32] J. Sethuraman. A constructive definition of Dirichlet priors. *Statistica Sinica*, 4:639–650, 1994.

[33] B. Sun, P. Mitra, C. L. Giles, J. Yen, and H. Zha. Topic segmentation with shared topic detection and alignment of multiple documents. In *Proc. of ACM SIGIR*, pages 199–206, 2007.

[34] K. Yu, S. Yu, and V. Tresp. Dirichlet enhanced latent semantic analysis. In *Proc. of AISTATS*, 2005.

[35] L. Zhou, M. Ticrea, and E. Hovy. Multi-document biography summarization. In *Proc. of EMNLP*, 2004.