Text segmentation: A topic modeling perspective

Hemant Misra [a,b,*], François Yvon [c], Olivier Cappé [d], Joemon Jose [a]

[a] *Department of Computing Science, University of Glasgow, Glasgow, United Kingdom*
[b] *Xerox Research Centre Europe, Meylan, France*
[c] *Univ Paris-Sud 11 and LIMSI-CNRS, Orsay, France*
[d] *TELECOM ParisTech and LTCI/CNRS, Paris, France*

In this paper, the task of text segmentation is approached from a topic modeling perspective. We investigate the use of two unsupervised topic models, latent Dirichlet allocation (LDA) and multinomial mixture (MM), to segment a text into semantically coherent parts. The proposed topic model based approaches consistently outperform a standard baseline method on several datasets. A major benefit of the proposed LDA based approach is that along with the segment boundaries, it outputs the topic distribution associated with each segment. This information is of potential use in applications such as segment retrieval and discourse analysis. However, the proposed approaches, especially the LDA based method, have high computational requirements. Based on an analysis of the dynamic programming (DP) algorithm typically used for segmentation, we suggest a modification to DP that dramatically speeds up the process with no loss in performance. The proposed modification to the DP algorithm is not specific to the topic models only; it is applicable to all the algorithms that use DP for the task of text segmentation.

1. Introduction

On several occasions, information needs of a user can be reasonably satisfied by presenting only the relevant part(s) of a document, and presenting the whole document may result in information overload. In information retrieval (IR), passage retrieval (Wilkinson, 1994) is a step in this direction, where instead of retrieving a set of documents in response to a query, a set of passages are retrieved. In this context, estimating passage boundaries reliably in an unsupervised manner is a key step in performing IR at the segment level.

Linear text segmentation is the task of dividing a given text data into topically coherent segments (Allan, Carbonell, Doddington, Yamron, & Yang, 1998; Beeferman, Berger, & Lafferty, 1999; Choi, 2000; Hearst, 1997; Kozima, 1993; Malioutov & Barzilay, 2006; Ponte & Croft, 1997; Reynar, 1998; Utiyama & Isahara, 2001). Text segmentation is a fundamental requirement for many IR applications, e.g., dividing a news broadcast transcription into stories (if possible, with a topic tag) could be very useful for browsing/retrieval. In a normal setting when no text segmentation is performed, if a user needs to access a particular story in a news broadcast, he may have to search the entire broadcast to get the story. In contrast, if the news is segmented (either manually or automatically) into stories and labeled, the relevant story can be retrieved directly. Text segmentation can also improve a user's retrieval experience by segmenting a document into topics and subtopics, and presenting only the relevant parts of the document during a search operation. Text segmentation assumes importance in text summarization and discourse analysis (the detection of topic changes) as well (Hearst, 1997).

* Corresponding author at: Department of Computing Science, University of Glasgow, Glasgow, United Kingdom.
E-mail addresses: hemant@dcs.gla.ac.uk (H. Misra), yvon@limsi.fr (F. Yvon), cappe@telecom-paristech.fr (O. Cappé), jj@dcs.gla.ac.uk (J. Jose).

Several approaches have been proposed in the past to perform this task. Most of the unsupervised approaches exploit *lexical chain* information, the fact that related or similar words tend to be repeated in topically coherent segments and segment boundaries are often linked to a change in the vocabulary (Choi, 2000; Hearst, 1997; Kozima, 1993; Malioutov & Barzilay, 2006; Stokes, Carthy, & Smeaton, 2004; Utiyama & Isahara, 2001). Different kinds of lexical cohesion are discussed in Stokes et al. (2004), repetition being one among them. These approaches typically do not require a training phase (and data), and can be directly applied to any text from any domain, subject to the only constraint that word boundaries can be identified. These approaches may fail to produce reliable segment boundaries if the word repetition is avoided by design.[1] A notable variation of these approaches (to overcome the issue of non-repetition because of words being replaced by synonyms or hypernyms) is when data is projected onto some latent space before performing the similarity match (Bestgen, 2001; Brants, Chen, & Tsochantaridis, 2002; Choi, Wiemer-Hastings, & Moore, 2001). Nevertheless, a potential drawback of most of these approaches is that even when the segment boundaries are estimated correctly, the segments are not associated (labeled) with any topic information.

In order to overcome these limitations of the lexical chain approaches, this paper proposes a new methodology for text segmentation that builds upon unsupervised topic modeling techniques. The two topic models investigated in this paper are multinomial mixture (MM) (Nigam, McCallum, Thrun, & Mitchell, 2000; Rigouste, Cappé, & Yvon, 2007) and latent Dirichlet allocation (LDA) (Blei, Ng, & Jordan, 2002); both the models are generative. In our proposed system, first the models are trained and then used to segment running texts. During training, the models learn the semantics information from the dataset and hence do not rely on mere word repetitions to segment the text. This is a departure from the lexical chain approaches that are typically knowledge-free. Our proposed approach for text segmentation using topic models also differs from the lexical chain approaches in that the proposed approaches "jointly" perform segmentation and topic labeling (outputs the topic distribution associated with each segment). An expected benefit of these approaches is their ability to identify the topic of each segment, thus allowing one to track topics within a long document or within a collection. However, a fallout of this topic estimation process is that it increases the computational cost of the entire process by several orders of magnitude.

The main objective of this study is to investigate whether text segmentation can be achieved from a topic modeling perspective. The empirical contributions of this paper for text segmentation task are manyfold:

- We demonstrate that the topic model based approaches proposed in this paper yield a better performance than the standard baseline technique proposed in Utiyama and Isahara (2001), provided that in-domain data is available for training the topic models. Our study also provides new insights into the strengths and weaknesses of the methods.
- We also introduce a modification to the dynamic programming (DP) algorithm that results in a substantial reduction in computational cost associated with topic model based text segmentation approaches. This makes topic modeling a viable alternative even when the document needs to be segmented online. The proposed modification is not specific to the topic model based methods only; it can be straightaway applied to all other methods that use DP for text segmentation task.
- The proposed methods are first evaluated on a standard dataset. Later, a bigger and more realistic dataset is designed to do an in-depth investigation of the issues associated with the text segmentation task. Finally, we validate and reconfirm the findings of this paper on a third dataset which was a part of TRECVid 2003 evaluations.

The rest of the paper is organized as follows: In Section 2, we briefly review extant approaches to linear text segmentation and relate our methodology with state-of-the-art techniques. In Section 3.1, we recall the main aspects of two topic models, first LDA and then MM, along with their training and testing procedures. In Section 4, we recall the principles of dynamic programming (DP) for text segmentation, first reviewing the method proposed by Utiyama and Isahara (2001), one of the most successful approaches to date, and then explaining how to adapt these principles when fragments are scored under the two topic models. In Section 5.3, we compare the performance of the three methods on Choi's benchmarks (Choi, 2000). We then investigate these methods on a larger dataset derived from the Reuters News Corpus Volume 1 (RCV1) database (Lewis, Yang, Rose, & Li, 2004). Section 5.4 presents the results obtained on this dataset, and discusses the merits and weaknesses of all the approaches. We revisit the (DP) algorithm and analyze it in Section 6, proposing a solution which brings a significant reduction in the computational cost when using DP for text segmentation task. In Section 7, results are presented on the dataset proposed for TRECVid 2003 news story segmentation evaluation to revalidate the main findings of this paper. Conclusions of this study are drawn and the future prospects are discussed in Section 8. The proposed LDA based approach provides segment boundaries as well as the topic distribution associated with these segments. To show these two complementary outputs of the proposed approach, in Appendix A we show the top topic of the segments for a sample text.

2. Background

The approaches proposed for linear text segmentation can be primarily divided into two groups: (a) the ones that resort to linguistic information such as cue phrases, syntax or lexical features; (b) the ones that use the information regarding

[1] See, e.g., http://www.thinkage.ca/jim/prose/repetitionvselegantvariation.htm; link active as on May 2010.

character, word or phrase repetition. Though a few studies did use the linguistic cues for the task, for example (Beeferman et al., 1999; Reynar, 1998), most of the approaches proposed in the literature (Brants et al., 2002; Choi et al., 2001; Eisenstein, Barzilay, & Davis, 2008; Fragkou, Petridis, & Kehagias, 2004; Hearst, 1997; Kehagias, Pavlina, & Petridis, 2003; Kozima, 1993; Malioutov & Barzilay, 2006; Morris & Hirst, 1991; Ponte & Croft, 1997; Reynar, 1998; Stokes et al., 2004; Utiyama & Isahara, 2001; Youmans, 1991), were motivated by the observation that coherent text usually contains repeating or similar vocabulary, an observation often attributed to Halliday and Hasan (1976).

Linguistic information is typically a good indicator of a segment boundary or non boundary. For example, a pronoun in a sentence usually indicates that the sentence is related to the previous sentence(s). In a news broadcast, a greeting like "good morning" is an indicator of the beginning of a segment and greetings such as "next to follow" provide a reliable indication of a segment end (Reynar, 1998). However, these cues may be specific to a particular type of data or media, news broadcast in the above discussion, and cannot be used across the board in all application domains. For each new application, one needs to analyze the data and define the appropriate cues.

Vocabulary repetition is usually linked with continuation of a segment, however all the words may not be equally important. For example, function words such as *the*, *of*, *for*, are common to all the segments in a text and must be discounted for while counting the word repetition. Statistics regarding the relationship between changes in vocabulary size (number of unique words) with changes in word tokens (number of word occurrence) was used in Youmans (1991) to estimate the segment boundaries; this study was based on the assumption that the vocabulary size increases more rapidly when a new segment starts as compared to when inside a segment. Attempts have also been made to use a broader definition of repetition, for instance taking into account repeated occurrences of synonyms. Along these lines, the effort of Morris and Hirst (1991) was to (manually) integrate a thesaurus, thus avoiding the problem of synonyms when identifying the *lexical chains* and subsequently finding the segment boundaries. In Kozima (1993), "semantic similarity between words" (called as *Lexical Cohesion Profile (LCP)*) was measured by manually designing a semantic network from a small English dictionary and spreading activation on this network. LCP was used for finding the segment boundaries. Though these earlier works were promising, they were applied only to small text databases and were not backed by any measurable results.

In the *TextTiling* approach of Hearst (1997), blocks of words were represented as word count vectors and similarity between adjacent blocks was measured using dot-product in the vector space. It was asserted that a drop in similarity was typically associated with a segment boundary. The proposed method was also compared with the method described in Youmans (1991) and the results were presented in terms of *Precision* and *Recall*. To illustrate the influence of this work, it is worth noting that in the TRECVid 2003 story segmentation task, *TextTiling* remained the most common method for segmenting the stories based on the text stream (Chua, Chang, Chaisorn, & Hsu, 2004). Since then, the same intuition has given rise to several important variants and extensions most of which incorporate two main ingredients: (i) a measure of similarity for contiguous chunks of text, or a measure of the internal cohesiveness of text chunk; (ii) a strategy for computing the best segment boundaries based on these similarity/cohesiveness patterns.

Various ways to compute these similarities have been put forward, for instance, Choi (2000) replaced the numerical similarity values with rank values and used a divisive clustering algorithm (Reynar, 1998) to find the segment boundaries; this paper also introduced a synthetic database which was later used as a benchmark in several publications. Utiyama and Isahara (2001) used a probabilistic measure of cohesiveness; this method is used as a baseline in this study and is further discussed in Section 4.2. Another line of improvement was to compute the similarity in so-called latent semantic spaces so as to capture not only exact repetitions, but also repetitions of related words (Bestgen, 2001; Brants et al., 2002; Choi et al., 2001; Stokes et al., 2004; Sun, Li, Luo, & Wu, 2008). Finally, recent unsupervised text segmentation techniques (Fragkou et al., 2004; Kehagia et al., 2003; Malioutov & Barzilay, 2006; Ponte & Croft, 1997; Sun et al., 2008; Utiyama & Isahara, 2001) have replaced the heuristic computation of the optimal segment boundary location with an exact computation based on dynamic programming (DP). In DP, similarity between each possible segment pair is computed, whereas in *TextTiling*, similarity is computed only between adjacent blocks. We will discuss DP and its computational complexity in Section 4; in Section 6 we investigate why computing similarity between all the segment pairs is not a necessity, and the information about *similarity between the number of neighbouring segments that must be computed* can be extracted in an unsupervised manner from the text data to be segmented.

Our work draws inspiration from these proposals, and relies on DP to locate segment boundaries. Where we depart from most of the previous work is in our assessment of the cohesiveness of a segment using probabilistic topic models. In our proposal, the topic model is first trained on a large amount of text data, and is then used to segment *running texts* it has not seen earlier (text not used for training).

Different probabilistic models have been used in the past for automatically segmenting texts: for instance, probabilistic models in Utiyama and Isahara (2001), hidden Markov model (HMM) in a fully supervised setting (Blei & Moreno, 2001; Mulbregt, Carp, Gillick, Lowe, & Yamron, 1998), and more recently LDA in Sun et al. (2008). The latter approach, however, is very different from the topic model based approaches proposed in this paper. In Sun et al. (2008), *the data to be segmented is used to train the LDA model* which makes the approach unfit for segmenting running texts. After training LDA model on the data to be segmented, based on the Fisher kernel the likelihood score of each possible segment is used to estimate similarity score between contiguous chunks. The authors compared the performance of their LDA based method with the approaches suggested in Fragkou et al. (2004); Hearst (1997) on a synthetic dataset but failed to demonstrate an improvement over the much simpler *TextTiling* method of Hearst (1997). The reason for this may be as follows: the data to be segmented is usually limited therefore the LDA parameters may not be estimated reliably in this kind of setup.

In contrast to the previous attempts using probabilistic models, our approach does not require any annotated data or an analysis of the test corpus in advance. Our approach requires building a model prior to segmentation on some similar data source; these models are our primary source of information when computing the cohesiveness of text segments.

3. Unsupervised topic models

In this section, we briefly explain two unsupervised topic models, LDA and MM, that are investigated in this paper for the task of text segmentation.

3.1. Latent dirichlet allocation model

LDA is a generative unsupervised topic model (Blei et al., 2002; Griffiths & Steyvers, 2004). In Blei et al. (2002), the authors showed that the model can capture semantic information from a collection of documents, and demonstrated its superiority vis-à-vis several other models including multinomial mixture (MM) model (Nigam et al., 2000; Rigouste et al., 2007) and probabilistic latent semantic analysis (Hofmann, 2001). Moreover, they investigated the use of LDA for the tasks of text modeling, text classification and collaborative filtering. In Griffiths and Steyvers (2004), LDA model was used to identify "hot topics" by analyzing the temporal dynamics of topics. Lately, the use of LDA model has been investigated in several applications where topic detection plays a key role, including, but not limited to, unsupervised language model adaptation for automatic speech recognition (ASR) (Heidel, an Chang, & shan Lee, 2007; Tam & Schultz, 2006), fraud detection in telecommunications (Xing & Girolami, 2007), and detecting semantically incoherent documents (Misra, Cappé, & Yvon, 2008).

This paper explores the use of topic modeling properties of LDA for yet another important application, the task of text segmentation. Our approach is based on the premise that using a topic model may allow better detection of segment boundaries because a segment change should be associated with a significant change in the topic distribution.

3.1.1. LDA: basics

LDA adopts the traditional view that texts are represented as word count vectors, and relies upon a two step generation process for these vectors. A key assumption is that *each document is represented by a specific topic distribution* and *each topic has an underlying word distribution*.

The probabilistic generative story of a document is as follows: assuming a fixed and known number of topics, T, for each topic t, a distribution ϕ_t is drawn from a Dirichlet distribution of order V, where V is the vocabulary size. The first step for generating a document is to draw a *topic distribution*, $\Theta = \{\theta_t, t = 1 \ldots T\}$ from a Dirichlet distribution over the T-dimensional simplex. Next, assuming that the document length is fixed, for each word occurrence in the document a topic, z, is chosen from Θ and a word is drawn from the word distribution associated with the topic z. Given the topic distribution, each word is thus drawn independently from every other word using a *document specific* mixture model.

Given Θ, the probability of w_i, the ith word token in a document, is thus

$$P(w_i|\Theta, \Phi) = \sum_{t=1}^{T} P(z_i = t|\Theta) P(w_i|z_i, \Phi) \tag{1}$$

$$= \sum_{t=1}^{T} \theta_t \phi_{tw_i} \tag{2}$$

where $P(z_i = t|\Theta)$ is the probability that the tth topic was chosen for the ith word token and $P(w_i|z_i = t, \Phi)$ is the probability of w_i given topic t. The likelihood of a document, represented as a count vector C, is a mere product of terms such as (2)

$$P(C|\Theta, \Phi) = \prod_{v=1}^{V} \left[\sum_{t=1}^{T} (\theta_t \phi_{tv}) \right]^{C_v} \tag{3}$$

where C_v is the count of word v in the document.

3.1.2. LDA: training

During training, the following two sets of parameters are estimated from a set of documents: the topic distribution in each document d (θ_{dt}, $t = 1 \ldots T$, $d = 1 \ldots D$) and the word distribution in each topic (ϕ_{tv}, $t = 1 \ldots T$, $v = 1 \ldots V$). Both Θ and Φ are interpreted as the parameters of a multinomial distribution and indicate which topics are important for a particular document and which words are important for a particular topic respectively.

The task of estimating parameters can be accomplished using statistical techniques such as variational Bayes (Blei et al., 2002), expectation propagation (Minka & Lafferty, 2002) and Gibbs sampling (Griffiths & Steyvers, 2004); the latter technique has been used in this study for training. In Gibbs sampling, two hyper-parameters α and β define the non-informative Dirichlet priors on Θ and Φ respectively.

The estimation procedure for LDA model using Gibbs sampling is detailed in Griffiths and Steyvers (2004). In short, for each word token in the training data, the probability of assigning the current token to each topic is conditioned on the topic

assigned to all other tokens. A topic is then sampled from this conditional distribution and assigned to the current token. The assignments computed for all the word tokens in the training data constitutes a Gibbs sample. For a particular Gibbs sample, the estimates for Θ and Φ are derived from the counts of hypothesized topic assignments as

$$\phi_{tv} = \frac{J_{tv} + \beta}{\sum_{k=1}^{V} J_{tk} + V\beta} \tag{4}$$

$$\theta_{dt} = \frac{K_{dt} + \alpha}{\sum_{k=1}^{T} K_{dk} + T\alpha} \tag{5}$$

where J_{tv} is the number of times word v is assigned to topic t and K_{dt} is the number of times topic t is assigned to a word token in document d.

3.1.3. LDA: testing

Training LDA on a text collection reveals the thematic structure of the collection, and has been the primary application of LDA in, e.g., Blei et al. (2002); Griffiths and Steyvers (2004). Being a generative model, LDA can also be used to make predictions regarding novel documents (*assuming that they use the same vocabulary as the training corpus – vocabulary mismatch issue in LDA model is explained in* Table 2). As the topic distribution of a test document gives its representation along the latent semantic dimensions, computing this distribution is important in many applications, including the present task of text segmentation.

This computation can be performed using the iterative procedure suggested in Heidel et al. (2007); Misra et al. (2008), which relies on the following update rule

$$\theta_{dt} \quad \frac{1}{l_d} \sum_{v=1}^{V} \frac{C_{dv}\theta_{dt}\phi_{tv}}{\sum_{t'=1}^{T} \theta_{dt'}\phi_{t'v}} \tag{6}$$

where l_d is the document length computed as the number of running words. As discussed in Misra et al. (2008), this update rule converges monotonically towards a local optimum of the likelihood, and convergence is typically reached in less than a dozen iterations. Once the Θ has been obtained for a document, the likelihood of the document can be computed by (3). This recently proposed method for computing Θ for unseen documents is key to computing the likelihood of a document. In this paper, we extend this idea to compute likelihood of a segment and use the estimated likelihood of segments as scores for performing the text segmentation task.

3.2. Multinomial mixture model

In an MM model, it is assumed that every word in a document belongs to the same topic and each document is thus represented by only one topic. In other words, in the topic space, the document specific topic distribution lies on one vertex of the $[0,1]^T$ simplex. Assuming that θ_t denotes the position independent probability of topic t in the entire collection, the probability of a document is:

$$P(C_d|\theta_t, \Phi) = \sum_{t=1}^{T} \theta_t \prod_{v=1}^{V} \Phi_{tv}^{C_{dv}} \tag{7}$$

This model can be trained through expectation maximization (EM) using the following re-estimation formulas:

$$P(t|C_d, \Theta, \Phi) = \frac{\theta_t \prod_{v=1}^{V} (\Phi'_{tv})^{C_{dv}}}{\sum_{t=1}^{T} \theta_t \prod_{v=1}^{V} (\Phi_{tv})^{C_{dv}}} \tag{8}$$

$$\theta'_t \propto \alpha + \sum_{d=1}^{D} P(t|C_d, \Theta, \Phi) \tag{9}$$

$$\Phi'_{tv} \propto \beta + \sum_{d=1}^{D} C_{dv} P(t|C_d, \Theta, \Phi) \tag{10}$$

where (8) defines the E-step, and (9) and (10) define the M-step.

As suggested in Rigouste et al. (2007), we initialize the EM algorithm by drawing initial topic distributions from a prior Dirichlet distribution with hyper-parameters $\alpha = 1$ and $\beta = 0.1$ in all the experiments.

During testing, the parameters of the MM models are used to estimate the posterior topic distribution in an unseen document using (8). The likelihood of the unseen document is given by (7). As in the case of LDA model, we extend this idea to compute likelihood of a segment and use the estimated likelihood of segments as scores for performing the text segmentation task.

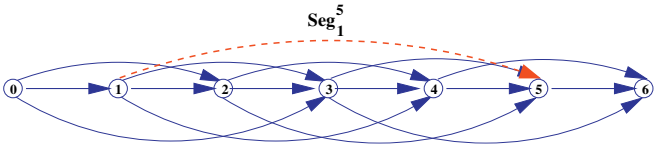

Fig. 1. Nodes and segments in dynamic programming.

```
0-1
1-2 0-2
2-3 1-3 0-3
3-4 2-4 1-4 0-4
4-5 3-5 2-5 1-5 0-5
5-6 4-6 3-6 2-6 1-6 0-6
..........
(N-3)-(N-2) (N-4)-(N-2) .... 0-(N-2)
(N-2)-(N-1) (N-3)-(N-1) .... 0-(N-1)
(N-1)-N (N-2)-N (N-3)-N ............ 0-N
```

Fig. 2. Node pairs (segments) in standard DP. Segment "B–E" represents segment staring from "B" node and ending on "E" node. For example, Line 2: node 2 can be reached from node 1 (1–2) or node 0 (0–2).

4. Algorithmics of segmentation

4.1. Dynamic Programming (DP) with probabilistic scores

As discussed in Utiyama and Isahara (2001); Kehagia et al. (2003), text segmentation can be efficiently implemented with DP techniques. Assuming the text is represented as a linear graph, a segment is defined by two nodes, the begin (B) and the end (E) nodes. For instance, segment $\mathbf{Seg_1^5}$ (dotted line) in Fig. 1 is from begin node $B = 1$ (excluding $B = 1$) to end node $E = 5$ (including $E = 5$). Node 0 is treated as null node for convenience.

In the standard DP approach, scores for all the possible node pairs are computed. Therefore, if the graph contains N nodes, one has to consider $N*(N+1)/2$ node pairs as shown in Fig. 2.

Text segmentation thus proceeds as follows (Utiyama & Isahara, 2001): We denote $d = w_1 w_2 \cdots w_{l_d}$ a document of length l_d; and $S = S_1 S_2 \cdots S_m$ a particular segmentation S made up of m segments. The likelihood of S is thus

$$P(S|d) = \frac{P(d|S)P(S)}{P(d)} \tag{11}$$

In (11), $P(d|S)$ is the probability of d under segmentation S and $P(S)$ is a prior over segmentations that corresponds to a penalty factor. Assuming that S_i contains n_i word tokens, and that w_i^j denotes the jth word token in S_i, we denote $W_i = w_i^1 \cdots w_i^{n_i}$; therefore, $d = W_1 \cdots W_m$ with $l_d = \sum_{i=1}^{m} n_i$. Under these assumptions, W_i and S_i are in a one to one correspondence. Further, assuming that segments are independent of each other, (11) can be rewritten as[2]:

$$\begin{aligned} P(S|d) &\propto \left[\prod_{i=1}^{m} P(W_i|S)\right] P(S) \\ &\propto \left[\prod_{i=1}^{m} P(W_i|S_i)\right] P(S) \\ &\propto \left[\prod_{i=1}^{m} \prod_{j=1}^{n_i} P(w_i^j|S_i)\right] P(S) \end{aligned} \tag{12}$$

The most likely segmentation, $\widehat{S}$, is defined as $\widehat{S} = \underset{S}{\operatorname{argmax}}\, P(S|d)$, and can be recovered using DP in a manner similar to the resolution of shortest path problems. During the forward-pass, for each pair of nodes (B,E), the score of Seg_B^E is computed. The path that maximizes the cumulative score from the first to the last node is searched for, and for each E node the value of the best start node B is stored. The information about the best start node is used during trace back to find the path that maximizes the score, and in turn, the segment boundaries.

[2] For a given document d, $P(d)$ is constant for all the segmentations and can be dropped from the equation.

4.2. Scoring segments by baseline

The method proposed in Utiyama and Isahara (2001) consists of modeling each segment using the conventional multinomial model, assuming that segment specific parameters are estimated using the usual maximum-likelihood estimates with Laplace smoothing. This approach has often been used as a standard baseline in literature and shown to deliver competitive results on several databases (Fragkou et al., 2004; Malioutov & Barzilay, 2006; Utiyama & Isahara, 2001). The factors in (12) are thus computed as

$$P(w_i^j|S_i) = (C_i^j + 1)/(n_i + V_d) \tag{13}$$

where C_i^j is the frequency of word w_i^j in S_i and V_d is the number of unique words in d. Therefore, the log of $P(W_i|S_i)$ can be written as

$$\begin{aligned}\log P(W_i|S_i) &= \log\left(\prod_{j=1}^{n_i}\left[\frac{C_i^j+1}{n_i+V_d}\right]\right) = \sum_{j=1}^{n_i}\left[\log(C_i^j+1)\right] - n_i\log(n_i+V_d)\\ &= \sum_{v=1}^{V_{W_i}}\left[C_{W_iv}\log(C_{W_iv}+1)\right] - n_i\log(n_i+V_d)\end{aligned} \tag{14}$$

In (14), C_{W_iv} is the count of word v in W_i. The second term in (12) is the penalty factor; it defines the penalty that a segmentation algorithm incurs for every proposed segment change. In Utiyama and Isahara (2001), it was optimized to $\log P(S) = -m\log(l_d)$ to yield the best performance.

4.3. Scoring segments by LDA

The LDA based method proposed in this paper is based on the following premise: if a segment is made up of only one story, it will have only a few active topics, whereas if a segment is made up of more than one story, it will have a comparatively higher number of active topics. In Misra et al. (2008), the authors showed that documents with a few active topics have higher log-likelihood than documents that have several active topics. Extending this reasoning to segments, if a segment is coherent (the topic distribution for a segment has only a few active topics), the log-likelihood for that segment is typically high as compared to the log-likelihood in the case when a segment is not coherent. This observation is of critical importance in the success of the proposed LDA based approach for text segmentation task, and has been left unexplored except for its original use in detecting coherence of a document (Misra et al., 2008).

It is thus tempting to use the log-likelihood of each possible segment as a score in the DP algorithm and to recover the segmentation from the path that yields the highest log-likelihood. The information about topic distribution (Θ) for a text segment can be estimated by (6),[3] assuming that the parameter Φ (distribution of words in each topic) of the LDA model has been learned on a training corpus. This in turns allows to compute the likelihood of the segment using (3).

The proposed LDA based approach for text segmentation task works like this:

1. For each possible segment, S_i,
   (a) Compute its Θ by performing 20 iterations of (6):

$$\theta_{S_it} \quad \frac{1}{n_i}\sum_{v=1}^{V}\frac{C_{W_iv}\theta_{S_it}\phi_{tv}}{\sum_{t'=1}^{T}\theta_{S_it'}\phi_{t'v}}$$

   (b) Compute its log-likelihood using (3):

$$P(W_i|\Theta,\Phi) = \prod_{v=1}^{V}\left[\sum_{t=1}^{T}(\theta_{S_it}\phi_{tv})\right]^{C_{W_iv}}$$

   (c) The likelihood of the segment is treated as its score:

$$\log P(W_i|S_i) = \log P(W_i|\Theta,\Phi)$$

2. Substitute the scores of the segments in (12), and use DP to find the segmentation which maximizes the score.

The penalty factor we have used is defined as $\log P(S) = -p\cdot m\cdot\log(l_d)$, where $p = 3$ was empirically found to yield the best performance on a held out condition (on Choi's dataset with words as nodes) and has been used throughout.[4]

[3] This approach requires to run this iterative procedure on every possible segment, which proves extremely resource demanding. As we will discuss in Section 6, a modification is proposed to make this computation feasible.

[4] Though a small change in performance was observed with a change in p, it did not affect the performance significantly.

4.4. Scoring segments with MM

The MM topic model can be used to segment a text if in the algorithm described above the likelihood is estimated by an MM model. The segmentation algorithm in this case proceeds like this:

1. For each possible segment, S_i,
   (a) Compute its log-likelihood using (7):

$$P(W_i|\theta_t, \Phi) = \sum_{t=1}^{T} \theta_t \prod_{v=1}^{V} \Phi_{tv}^{C_{W_i v}}$$

   (b) The likelihood of the segment is treated as its score:

$$\log P(W_i|S_i) = \log P(W_i|\Theta, \Phi)$$

2. Substitute the scores of the segments in (12), and use DP to find the segmentation which maximizes the score.

Notice that in MM, unlike LDA, the topic distribution need not be estimated in order to estimate the likelihood. This makes the whole approach much faster than the LDA based approach (where an iterative procedure was needed to estimate the topic distribution before computing the log-likelihood); however, a drawback of this approach is that segments are not associated with a topic distribution.

5. Experimental results

5.1. Databases

The first dataset used in this study is the Choi's dataset,[5] which has often been used in benchmarking text segmentation algorithms. Choi's dataset is derived from Brown corpus. This corpus consists of running text of edited English prose printed in US during the calendar year 1961 and has a small sample size of 500 proses of approximately 2000 words each (http://129.177.24.52/icame/manuals/brown/; link active as on June 2010). Brown corpus contains "informative prose" (374 documents) as well as "imaginative prose" (126 documents), including 44 documents from the "press reportage" genre.

Choi's dataset is divided into four subsets ("3–5", "6–8", "9–11" and "3–11") depending upon the number of sentences in a segment/story. For example, in subset "*X*–*Y*", a segment is derived by (randomly) choosing a story from Brown corpus, followed by selecting first *N* (a random number between *X* and *Y*) sentences from that story. Exactly 10 such segments are concatenated to make a document. Further, in each subset, there are 100 documents to be segmented. By design, the segments are not complete stories. As expected, because of the small coverage of Brown corpus, the coverage of Choi's dataset is also quite limited.

To study the performances on a more realistic dataset, we created a dataset similar to Choi's dataset from the Reuters Corpus Volume 1 (RCV1) (Lewis et al., 2004). RCV1 is a collection of over 800,000 news items in English from August 1996 to August 1997. These news items belong to at least one of the following broad subject areas (topic codes): Corporate/Industrial (CCAT), Economics (ECAT), Government/Social (GCAT) and Markets (MCAT). We selected a set of 23,326 news items from RCV1 for generating the *Reuters dataset (RDS)* for text segmentation which is more comprehensive than Choi's dataset. As in Choi's dataset, four subsets are created in *RDS* ("3–5", "6–8", "9–11" and "3–11") depending upon the number of sentences in a segment. In addition, a fifth subset is created in *RDS* which has entire news items ("Fulltext"). Again, as in Choi's dataset, there are 100 documents to be segmented in each subset of *RDS*, and each document has 10 segments (**Set 1**) in it. However, with a realization that 10 segments in each document is too restrictive and may not be the only condition to occur in practice, we created two more sets varying the number of segments in a document. In *RDS*, **Set 2** and **Set 3** contain 50 and 100 segments respectively.

Though similar to Choi's dataset, *RDS* covers more situations. The number of segments in a document is not fixed to 10; it is 10, 50 or 100. Further, subset "Fulltext" in *RDS* has entire news items randomly chosen and then concatenated to form the documents.

From RCV1 collection, we selected another 27,672 news items for training the LDA model (*ReutersTrain*). The vocabulary size of this train set is approximately 93K and the number of word tokens, excluding stop words, are approximately 3.6M. In these experiments, number of topics (T) and Dirichlet priors (α and β) are set to the following values: $T = 50$, $\alpha = 1$ and $\beta = 0.01$. These values were chosen for two main reasons: (a) in our previous work on coherence detection (Misra et al., 2008), which was the initial motivation for using LDA for the task of text segmentation, these values yielded good performance, and (b) $T = 50$ is large enough to keep the model flexible but at the same time keep the computational complexity of the segmentation algorithm moderate; it may be recalled from (6) that theta estimation is an iterative procedure and its computational cost increases with an increase in T (though it was not a factor here, it may be noted that an increase

[5] http://www.freddychoi.co.uk/; link active as on September 2009.

Table 1
The performance of text segmentation algorithms for various setups on Choi's database in terms of P_k value (along with the average time taken to segment each text document). A lower P_k value indicates a higher accuracy in text segmentation.

Method	Stemmer	P_k, % (time, s)			
		3–5	6–8	9–11	3–11
Baseline	Porter	13	6	6	11
Baseline	None	14.5 (0.16)	7.1 (0.66)	6.4 (1.63)	11.3 (0.67)
LDA	None	22.9 (120.7)	15.4 (569.5)	13.8 (1503.4)	15.8 (596.0)
MM	None	38.4 (9.6)	33.9 (22.3)	33.9 (47.3)	33.1 (22.8)

in T also increases the computational cost associated with the training of LDA model). Similar values of these parameters were found to be a reasonable choice while training an LDA model on data of similar size in Griffiths and Steyvers (2004).

RCV1 was used for the following reasons: (a) It is a large database so we can have separate train (*ReutersTrain*) and test (*RDS*) sets. (b) RCV1 is built over a period of one year, using real news stories, so bias towards a particular domain or vocabulary are less likely. (c) *ReutersTrain* had enough train documents to cover minimum vocabulary size and do reliable LDA parameters estimation. (d) In *RDS*, the subset "Fulltext" simulated realistic conditions whereas other subsets were to emulate Choi's dataset. (e) In *RDS*, the number of segments in a document were varied (10, 50 and 100) in order to simulate a wider spectrum of the text segmentation problem.

5.2. Test conditions

In all the previous studies reported in the literature (Brants et al., 2002; Choi et al., 2001; Fragkou et al., 2004; Utiyama & Isahara, 2001), the information about the sentence end is used while performing segmentation. In such a case, each sentence start is a possible B node while each sentence end is a possible E node. The same setup is studied in this paper.

5.3. Choi's data: results and analysis

5.3.1. Baseline experiments

In this section, we compare the results of the following segmentation systems (see Table 1)[6]:

- The results reported in Utiyama and Isahara (2001) (using Choi's implementation of Porter stemmer)
- Our own implementation of Baseline method
- LDA based segmentation
- MM based segmentation

The results are presented in terms of P_k value, the probabilistic error metrics introduced in Beeferman et al. (1999). P_k is the probability that two randomly drawn sentences which are k sentences apart are classified incorrectly. As in Choi (2000), k is set to the average segment length in our experiments. A lower value of P_k indicates a higher accuracy in text segmentation.

The results reported in Table 1 suggest that the performance of the Baseline and LDA based methods consistently improves with an increase in segment size. This is an expected result: longer segments allow a better estimation of the multinomial parameters for the baseline method and of the topic distribution for LDA. This trend is observed in the performance of the MM based method as well, but it is much weaker.

Out of the two topic models, the LDA based method gives a better performance as compared to that of the MM based method. This result shows that the underlying assumption of a topic distribution associated with each document (LDA model) is able to represent the data better as compared to the assumption that each document is represented by a single topic (MM model).

Compared to the topic model based approaches, the baseline method is: (i) more accurate and (ii) an order of magnitude faster. An inspection of outputs (other than just the segment boundaries provided by the two methods) gives a possible explanation for (i): there is a serious mismatch in vocabulary between *ReutersTrain* dataset (used for LDA and MM models training) and Choi's dataset used for testing.[7] Similar issues of semantic mismatch were highlighted, for example, in Bestgen (2001), where the authors used a generic latent semantic space while performing text segmentation. This issue is explained further by the statistics presented in Table 2. The baseline method utilizes the full available vocabulary and all the content words (stop words were removed before segmenting the text) for computing the score of a segment. In contrast, the vocabulary of the unsupervised topic models is defined by its training data. During test, the content words in the (test) data that are not present in

[6] In the case of LDA and MM models, the number of topics is 50 and the models are trained on the *ReutersTrain* dataset.

[7] It may be recalled that Choi's dataset is extracted from the Brown corpus which itself was published in 1961, has a small coverage and is relatively biased towards non-news English prose of that time (check http://129.177.24.52/icame/manuals/brown/ or http://en.wikipedia.org/wiki/Brown_Corpus#Sample_distribution, to see the sample distribution of prose of this corpus; the links active as on May 2010). Not only is there a mismatch in domain (news vs. mostly non-news prose) but also in time (1996 vs. 1961).

Table 2
The "average vocabulary size of documents" and "average number of content words in documents" when the three methods are used for computing scores.

	Vocabulary size			
	3–5	6–8	9–11	3–11
Baseline	398.4	641.0	853.3	651.6
LDA/MM	350.4	567.8	751.6	574.1
	Number of content words			
	3–5	6–8	9–11	3–11
Baseline	502.7	869.1	1220.3	869.4
LDA/MM	446.8	779.9	1090.1	776.5

Table 3
Text segmentation performance on Choi's dataset by the baseline, unadapted and adapted LDA topic model.

Method	P_k, %			
	3–5	6–8	9–11	3–11
Baseline	14.9	8.1	7.7	11.4
Unadapted LDA (ReutersTrain)	23.6	15.9	15.1	16.2
Adapted LDA (ReutersTrain+BrownTrain)	**16.3**	**10.9**	**10.9**	**12.1**

the training vocabulary are not used for computing the score of a segment. Comparing the baseline and topic model based methods, approximate loss in vocabulary and content words by topic models are 11.8% and 10.5% respectively.

Apart from this loss in vocabulary, the other aspect is the distribution of the content words present in the test set. We found that the high frequency content words of *ReutersTrain* set were mostly missing or under represented in the documents of *Choi's* set, and these documents often consist of low frequency content words of *ReutersTrain* set.

5.3.2. Experiments with an adapted LDA model

To overcome this problem of vocabulary, domain and temporal mismatches, we conducted a second series of complementary experiments using more appropriate training data. Given that Choi's pseudo-documents are generated using (some of) the first 11 sentences of documents of the Brown corpus, we built a new training set containing documents both from the Reuter corpus (as before) and 500 pseudo-documents from the Brown corpus. The last 40 sentences of each document in the Brown corpus were selected to generate these pseudo-documents. The new combined training set has a vocabulary size of approximately 109K and the number of content word tokens are approximately 3.9M, an increase of approximately 17.2% and 7.4% over *ReutersTrain* set respectively. Using this new dataset, we trained an "adapted" LDA model using the same value of hyper-parameters as before ($T = 50$, $\alpha = 1$ and $\beta = 0.01$) and used this adapted model to produce segmentations. These results are reported in Table 3. The results presented in Table 3 show that using a training corpus that resembles the test data more closely allows to significantly improve the performance across all the conditions, the improvement being especially large for shorter segments. With this new training dataset, our results are more comparable with the baseline.[8]

A part of this improvement is due to the reduction in vocabulary mismatch: with this new training data the average number of content word tokens taken into account during test increases by approximately 2.5%. This however also means that in spite of increasing the vocabulary size by approximately 17.2%, still about 8% of the content words present in the test corpus are not used. Therefore the main cause of this improvement can be attributed to the creation of topics that more appropriately describe the thematic content of the Brown corpus. Using these new topics, our algorithm is now in a position to identify segments boundaries between documents which previously were thematically indistinguishable.

The main conclusion of this small study is the reiteration of the fact that topic models need to be trained on a dataset that is as similar as possible to the test documents to be segmented.

In the next section, we present the results on a bigger and more realistic dataset (RDS) to investigate other issues related with the task of text segmentation which were not observed while working with Choi's dataset. Finally, in Section 7, we evaluate all the algorithms on a dataset that was proposed for TRECVid 2003 news story segmentation evaluation task. The results obtained on TRECVid dataset again show the limited coverage of Choi's dataset. By using an adapted LDA model, an average improvement of approximately 4–7% absolute is obtained on Choi's dataset whereas improvement on TRECVid 2003 dataset is a modest (though significant) 2% absolute.

[8] In fact, additional experiments suggest that a more careful tuning of the various parameters, notably the number of themes and penalty factor, could yield further gains. Due to the peculiar nature of Choi's dataset, we think that trying to tune/optimize the segmentation performance on this test set is not very informative.

Table 4
The text segmentation performance on the Reuters database (along with the average time taken per file).

Method	P_k, % (time, s)				
	3–5	6–8	9–11	3–11	Fulltext
Set 1: Segments per document = 10					
Baseline	16	14	14	14	17
	(0.208)	(0.776)	(1.541)	(0.665)	(10.619)
LDA	6.5	4.5	5.6	5.9	11.1
	(200.3)	(801.9)	(2309.9)	(734.2)	(8750.8)
MM	14.8	13.3	12.1	13.8	11.5
	(8.4)	(14.4)	(24.3)	(13.7)	(94.5)
Set 2: Segments per document = 50					
Baseline	36	34	33	34	29
	(21.85)	(92.77)	(191.3)	(85.82)	(906.2)
LDA	9.6	5.8*	HCC	HCC	HCC
	(19,364)	(80,486)			
MM	19.4	14.7	13.1	15.6	12.1
	(200.3)	(832.8)	(1851.2)	(782.7)	(8846.7)
Set 3: Segments per document = 100					
Baseline	41	41	40	40	38
	(210.0)	(625.7)	(1285.6)	(400.0)	(5382.9)
LDA	6.6*	HCC	HCC	HCC	HCC
	(202,085)				
MM	21.0	16.2	14.4	16.5	13.4*
	(1454.7)	(6035.0)	(12828.8)	(5689.4)	(60165.3)

* The partial result is based only on a few documents, and "HCC" indicates that the segmentation could not be completed even for a single document in that subset because of high computation cost.

5.4. Reuters data: results and analysis

As explained earlier, *RDS* includes sets that contain a larger number of segments (**Set 2** and **Set 3** contain 50 and 100 segments per document respectively). In addition, a subset consisting of complete Reuters news stories was also considered ("Fulltext"). The results on this dataset are presented in Table 4. An analysis of these results unfolds a different story. For *RDS*, the performance of the baseline method drops significantly, especially when used to segment longer documents. In comparison, in spite of higher computational cost, the topic based approaches yield better results, at least for the test documents that could be segmented. This result shows that unsupervised topic model based approaches can produce segmentations that compare favorably with the baseline, provided the model is trained on a corpus that is similar in diversity to the test documents to be segmented.

As pointed out earlier on Choi's dataset, out of the two model based approaches, the performance of the LDA based approach is better than that of the MM based approach. On a complex dataset like *RDS*, better expressiveness of the LDA model is able to show its strength in data modeling. However, computational cost of the former is extremely high because of the iterative procedure for estimating the topic distribution.

The poor performance of Utiyama's method on longer pieces of text was also reported in Utiyama and Isahara (2001); Malioutov and Barzilay (2006). A possible explanation is as follows: The maximum-likelihood (ML) estimator in Utiyama and Isahara (2001), even with smoothing, is not robust for small segments. Based on the observation of a handful of words, Utiyama and Isahara (2001) claims to estimate a distribution over the entire vocabulary whose size increases rapidly as the number of segments in a document increase and could easily be a few thousand parameters for reasonably sized documents. By comparison, the LDA based method only attempts to estimate the Θ distribution, which corresponds to a much smaller number of parameters, and can thus be performed more robustly, even for small segments.

In Appendix A, we show the segmentation estimated by the proposed LDA based approach for a sample document. Along with the segmentation, we also show the topic assigned (only the topic with the highest probability is printed) by the LDA model to each segment. A comparison of "the top words in a topic" and "the topic assigned to a segment based on the words in that segment" gives a very promising picture. In the near future, we will study these results in greater detail for applications like discourse analysis and summarization.

Encouraged by these results, we analyzed the DP algorithm to alleviate the issue of high computation cost associated with the topic based segmentation algorithm.

6. Sparse path and modified DP

The main difficulty of DP based segmentation stems from its requirement to score all possible node pairs as segments, whose number grows quadratically with the number of sentences in a document. In LDA, scoring a segment requires an iterative estimation of its topic distribution. An obvious remedy to this problem is computing the score of only a subset of the

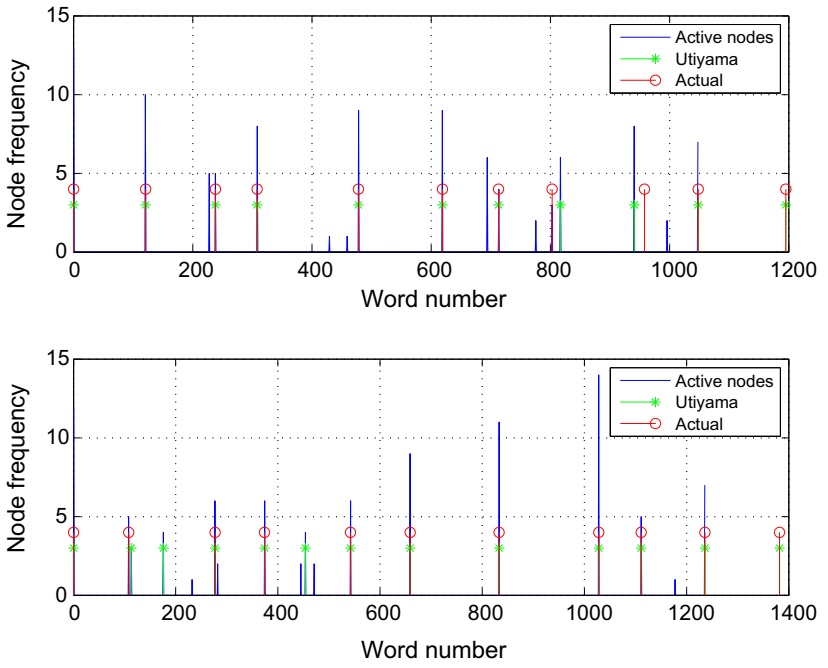

Fig. 3. Frequency of a node as an active *B* node for two different test files.

most relevant segments. In Malioutov and Barzilay (2006), for instance, the authors used a threshold to prune/filter the segments that were too long (their *B* and *E* nodes were too far apart), a strategy that proved effective in reducing the computational cost of their technique. Such distance thresholds were also used in Stokes et al. (2004) in order to reduce spurious chains and also to avoid very long segments.

Our proposed modification is based on the observation that during the forward-pass, most nodes are never active as *B* nodes. In other words, for most of the nodes, we cannot find a single *E* node whose best starting node is among these nodes. As a consequence, these nodes are never on the maximum score path. This observation reflects in Fig. 3, where we plot, for two different test files, the frequency count of active nodes during the forward-pass; more precisely, we display the number of times each node $B = i$ is an active begin node for any of the possible nodes $E = j$, with $j > i$. The figure also plots the actual segment boundaries and the segment boundaries obtained by the baseline method after back-tracing. As shown in Fig. 3, only a few nodes are active *B* nodes. Conversely, *B* nodes that are highly active (meaning that they are the best starting point of many segments) are very often a good candidate for segment boundaries. *This also means that the most likely candidates for segment boundaries can be found in the forward-pass itself, thus being in a position to segment in a streamwise manner*. Moreover, once an active *B* node is crossed, the segments starting from before this *B* node are mostly not in the maximum score path. Based on this observation, we suggest a solution to reduce the computational cost of DP: disregard the segments whose starting node lies before the last active *B* node.

This heuristic, which holds irrespective of the underlying probabilistic model, does not require defining a threshold for segment size as suggested in Malioutov and Barzilay (2006). In fact, this sparsity of data for text segmentation task is visible in many previous publications, though in a different format called as similarity or dot plot (for example, Fig. 1 in Malioutov and Barzilay (2006). Fig. 1 in Fragkou et al. (2004), Fig. 1 and 3 in Choi (2000)). However, it was never recognized, discussed or put to use. This information is present during the forward-pass of the DP itself and the saving in computational cost could be significant if it is used wisely.

As explained before, the DP path is sparse and most nodes are never active *B* nodes for any *E* node. If an active node ($B = X$) is found while doing the forward-pass, the paths originating from the nodes before this active node ($B < X$) are not considered for score computation.

In Fig. 4, for example, if $B = 3$ is an active node (gives the maximum cumulative score for $E = 4$), for $E = 5$ there is no need to compute the scores for segments *2–5, 1–5, and 0–5*. As the DP progresses, the last active node keeps getting updated. For example, if towards the end of the document $B = N - 4$ is an active node, scores of only a few segments need to be computed. In order to make this algorithm more robust, a node is considered active only when it has been active for at least two consecutive segments. In the discussion above, $B = 3$ is considered an active node only if it is the best starting point for $E = 4$ and $E = 5$. If $B = 3$ was an active node for at least 2 times, then for $E = 6$ the scores for segments *2–6, 1–6, and 0–6* need not be computed.

```
0-1
1-2 0-2
2-3 1-3 0-3
3-4 2-4 1-4 0-4
4-5 3-5
5-6 4-6 3-6
..........
(N-3)-(N-1) (N-4)-(N-2)
(N-2)-(N-1) (N-3)-(N-1) (N-4)-(N-1)
(N-1)-N (N-2)-N (N-3)-N (N-4)-N
```

Fig. 4. Relevant node pairs (segments) in modified DP (MDP). Compare with Fig. 2.

Table 5
The text segmentation performance on Reuters database (along with the average time taken per file) by Modified DP "MDP". Compare these results with the performance and processing time of the topic models in Table 4.

Method	P_k, % (time, s)				
	3–5	6–8	9–11	3–11	Fulltext
Set 1: Segments per document = 10					
LDA MDP	6.5 (33.3)	4.4 (57.5)	5.4 (91.9)	5.9 (58.3)	11.9 (349.6)
MM MDP	14.6 (7.2)	12.2 (7.7)	11.6 (8.2)	13.1 (7.7)	12.0 (12.3)
Set 2: Segments per document = 50					
LDA MDP	11.2 (172.8)	6.0 (320.0)	5.5 (545.0)	7.3 (332.4)	8.4 (2054.7)
MM MDP	18.5 (8.8)	13.8 (11.3)	12.3 (13.9)	14.6 (11.3)	12.0 (64.0)
Set 3: Segments per document = 100					
LDA MDP	13.6 (548.9)	7.8 (1009.8)	6.4 (1540.1)	8.4 (1030.6)	8.4 (7950.1)
MM MDP	19.8 (11.6)	15.3 (16.6)	13.7 (22.7)	15.5 (16.8)	12.9 (145.1)

This simple scheme, which trades-off the exactness of search with speed, proves to be very efficient in terms of reducing the computational load, especially for longer texts, with almost no loss in performance in most of the cases. This is confirmed by the results achieved by modified DP (MDP) algorithm (Table 5).

The results in Table 5 reflect the effectiveness of this simple heuristics, both in terms of speed and accuracy. For all segment sizes, the time needed to segment documents is divided by more than 10, with even larger gains for long documents. For test **Set 1** that contains 10 segments per file, we can actually compare the performance of exact and approximate search. We find that both the methods (DP and MDP) for both the topic models (LDA and MM) yield very comparable results. In fact, the modified algorithm proves slightly better in several cases as it discards solutions that are optimal for the model but would yield segments that are too long. For test **Set 2** and **Set 3** containing more segments (50 and 100 segments respectively), our topic model based methods outperform baseline by a wide margin. In fact, contrary to the baseline performance, the performance of the topic model based methods degrades very gracefully with an increase in the number of segments.

The performance of the two topic model based approaches can also be viewed from a different perspective. If time complexity is an important criterion for performance and one is willing to compromise the accuracy of the segmentation, it is reasonable to use MM based approach for the segmentation task. Nevertheless, LDA based method gives better performance than MM based method and in future we would like to focus on the iterative procedure of estimating the topic distribution in order to reduce the time complexity of the LDA based method.

7. A real application: news story segmentation task

In TRECVid 2003 (TREC Video Retrieval Evaluation, 2003), news story segmentation was defined as a separate evaluation task. In this section, we present the performance of the proposed LDA based method on the dataset proposed for TRECVid 2003 news story segmentation evaluation task. The dataset consists of 120 h of video news story collected from ABC and CNN news channels in the year 1998. Approximately 60 h of data was set aside for training or development (*TRECVidTrain*) and the rest for testing (*TRECVidTest*). The close caption text for these news broadcasts was also provided. The participants were expected to utilize multi-modal features from text, audio and video streams to perform the segmentation task.

In the actual evaluation, the participants were allowed to use specific cues like "good evening", "*xyz* (person) reporting from *abc* (place)", "coming next" etc from the text stream to facilitate the task. Such cues are assumed to be closely associated with news broadcasts and typically suggest a story change. These cues were obtained from the TRECVidTrain data. The TRECVid results are typically reported in terms of precision, recall and F1 measures. These measures are defined as:

Table 6
The text segmentation performance in terms of P_k and F1 measures on TRECVidTest dataset by the baseline, unadapted LDA (ReutersTrain) and adapted LDA (ReutersTrain+TRECVidTrain) methods.

Method	P_k, %	F1
Baseline	22.3	0.39
Unadapted LDA	22.7	0.40
Adapted LDA	20.7	0.44

$$\text{precision} = \frac{\text{no. of estimated boundaries that are actual}}{\text{no. of estimated boundaries}}$$
$$\text{recall} = \frac{\text{no. of estimated boundaries that are actual}}{\text{no. of actual boundaries}}$$
$$\text{F1} = \frac{2 * \text{precision} * \text{recall}}{\text{precision} + \text{recall}}$$

In this paper, we approach news story segmentation task as a text segmentation problem. Only the close caption text was utilized for performing this task. It is worth mentioning that *no cues specific to the news broadcast were employed to facilitate the task*.

We show the performance of the baseline and two LDA based methods in Table 6. In the first LDA based method, the model is trained on *ReutersTrain* dataset while the second LDA based method utilizes *ReutersTrain and TRECVidTrain* datasets for training the LDA model. The results are presented in terms of two measures, P_k and F1. It must be noted that a lower value of P_k indicates a better performance, whereas a higher value of F1 measure indicates a better performance. A comparison of the results shows that the baseline and unadapted LDA methods yield similar performances, and the performance of the LDA method can be improved by adaptation. These results validate the earlier results obtained on Choi's dataset that adaptation improves the performance of LDA based method.

These encouraging results on a real application by the unsupervised topic modeling approach (the proposed LDA based method in our case) suggest the following: text segmentation from unsupervised topic modeling perspective is an alternative and a competitive tool which gives a reasonable performance on several different datasets.

Still, the text segmentation problem is far from being solved and below are some of the observations from TRECVid 2003 dataset: In this dataset, a typical news broadcast of CNN and ABC started with "main headlines" followed by the actual news where the headlines were presented in detail. In the "main headlines", each story was represented by a single sentence. In most of these cases, it was not possible to estimate the boundaries correctly in the "main headlines" section. This highlights the problem of unsupervised approaches which need some minimum amount of data to perform reliable estimates of the score.

We expect our LDA based approach to be complementary to the existing approaches. In our last experiment we investigated the complementarity of the two approaches by combining the segment boundaries estimated by the baseline and the LDA methods by simple *OR* operation. That is, if any one of the methods suggested a segment boundary, we accepted it as a segment boundary. This combination was performed because an analysis of the segment boundaries estimated by the two methods revealed that: (a) a boundary estimated by these methods was mostly an actual boundary, (b) both the methods do undersampling, that is, the number of boundaries estimated by individual methods is always less than the number of actual boundaries, and (c) the boundaries estimated by the two methods are not the same. These observations suggest that the outputs provided by the two methods are complementary. This is expected because both the methods rely on different techniques to compute the segment scores and as a consequence have different estimates for segment boundaries. This simple combination yields an F1 measure of 0.52 and 0.55 for *Baseline+unadapted LDA* and *Baseline+adapted LDA* respectively.

8. Conclusions

In this study, we proposed an application of two well established unsupervised topic models (LDA and MM) for the task of text segmentation. Out of the two approaches proposed in this paper, the LDA based method was able to estimate the segment boundaries with higher accuracy as compared to the boundaries estimated by the MM based method.

Another advantage of the proposed LDA based method is that it computes topic distributions (Appendix A) jointly with segmentation, thus allowing one to collect information about the thematic content of each segment. This information can be used to keep track of recurring topics.

We investigated and compared the performance of our methods with a standard approach often used as a baseline for the text segmentation task (Utiyama & Isahara, 2001), and analyzed their potential strengths and weaknesses. The LDA based method gave a better performance than that of the baseline in matched conditions (train and test data from the same domain). As expected, when topic models are trained with data that substantially differs from the test set, the performance is less favorable; we show that the performance can be easily improved using a small amount of adaptation data. This trend was observed in all the adaptation experiments reported in this paper.

One drawback of the proposed LDA based approach was its high computational cost. We proposed a modification to the DP algorithm that effectively reduces the computational cost. The unsupervised method of discarding the unimportant segments from the process of score computation brings a substantial reduction in computational cost without significantly affecting the performance. The proposed modification in the DP algorithm is not only useful for text segmentation task but can also be used in many other situations where a similar sparsity in data is encountered.

LDA based approach remains an order of magnitude slower than the baseline, indicating that there is scope for improvement. One possible solution is to combine both the techniques, that is, using the baseline method to locate the most promising boundaries, and then rescoring them using the topic based methods. The second issue with the topic based approaches is related to mismatch between the train and the test domains. This could be alleviated, for instance, by adapting the parameters of the LDA model via a first pass on the test data, during which the topic assignments for the whole document would be computed and used to revise the parameters.

Acknowledgments

This research was partly supported by the European Commission under the Contracts *FP6-027122-SALERO* and *FP7-231854-SYNC3*.

Appendix A. Text segmentation by LDA: Sample output

The results previously published in the literature typically concentrated on the segmentation performance (either some error metric or time complexity). Though estimating the segment boundaries is important, yet if the segments can be identified by a topic (or topic distribution), this information can have profound impact in several other applications such as discourse analysis and information retrieval. LDA being a topic model is in a position to provide this information along with the segment boundaries; this very aspect of the LDA model is shown in this section.

A.1. An analysis of LDA output

In this section, we show the following two complementary outputs of the LDA model:

- Text segmentation phase output [Section A.2]: For a sample document segmented by LDA based approach, the top topic label assigned by LDA model to each segment. To save space, long sentences were terminated by "..." to reflect continuity beyond the printed words.
- Training phase output [Section A.3]: Top ten words of the topics that are found in the example segmented document in Section A.2.

A.2. Segmented text output

indian finance minister p chidambaram said on friday india was concerned over the impact ... look at what happened over the last two or three ... what happens to our oil import bill **MISSED BOUNDARY** pakistani officials said on saturday a war like situation existed ... they accused indian forces of starting the firing in the ... public rallies and marches were held in azad kashmir to ... at least one civilian was killed and two were wounded ... **ESTIMATED BOUNDARY is CORRECT: TOPIC 46 has highest probability (0.28)** shares in industrial conglomerate btr plc rose in early trading ... the stock was up 5p at two hundred sixty five ... british newspaper reports over the weekend had anticipated the sale ... **ESTIMATED BOUNDARY is CORRECT: TOPIC 50 has highest probability (0.26)** the washington post carried the following stories on its front ... dukan iraq government backed kurdish guerrillas overran sulaimaniya and captured ... washington more than hundred iraqi dissidents and military officers associated ... little rock a defiant susan mcdougal reported to jail this ... **ESTIMATED BOUNDARY is CORRECT: TOPIC 35 has highest probability (0.44)** south africa's chamber of mines said on wednesday it had agreed wage increases ... the national union of mineworkers num said earlier it had ... minimum wage rates for different mining houses have been adjusted ... **ESTIMATED BOUNDARY is CORRECT: TOPIC 32 has highest probability (0.41)** the czech government plans a balanced thousand nine hundred ninty ... the cabinet did not find bigger room for tax cuts ... the budget plan is expected to go to a final vote ... **INSERTED BOUNDARY: TOPIC 25 has highest probability (0.54)** the thousand nine hundred ninty seven budget assumes growth in gross domestic ... **ESTIMATED BOUNDARY is CORRECT: TOPIC 42 has highest probability (0.70)** deutsche bahn chairman heinz duerr said on wednesday that the current building ... the construction industry must be creative costs have to be reduced by at least fifty percent duerr told ... duerr cited infrastructure as one of a number of investment challenges facing the german railway ... **ESTIMATED BOUNDARY is CORRECT: TOPIC 19 has highest probability (0.19)** california state treasurer matt fong has called on the securities and exchange commission ... fong called for the investigations in letters to sec chairman arthur ... the letters were released thursday at a meeting of the california ... **ESTIMATED BOUNDARY is CORRECT: TOPIC 26 has highest probability (0.60)** about fifty containers of aluminium phosphate have been found on the ... we've found about fifty small tubes with the chemical inside them none were open but there's always a risk that some of the poison ... aluminium phosphate is used as rat poison and produces a toxic gas on

... **ESTIMATED BOUNDARY is CORRECT: TOPIC 16 has highest probability (0.38)** contract talks continued saturday between ford motor co and the united auto workers ... bargainers were optimistic according to ford spokesman jon harmon but ... they're not there yet on all the issues it's going to take a while to get through harmon told reporters ... **ESTIMATED BOUNDARY is CORRECT: TOPIC 32 has highest probability (0.29)**

A quick analysis of the segmented output shows that the topic associated with a segment is mostly relevant to the words present in that segment. In particular, the **MISSED BOUNDARY** occurs because both the stories have a common theme **india**. On the other hand, **INSERTED BOUNDARY** divides a story into two segments, one related to "government economic policy (TOPIC 25)" and the other related to "gdp growth (TOPIC 42)".

A.3. Φ Matrix: top 10 words of relevant topics

TOPIC 46: 'rupees' 'india' 'indian' 'pakistan' 'bombay' 'new' 'million' 'sri' 'india's' 'market'
TOPIC 50: 'plc' 'stg' 'group' 'london' 'pounds' 'british' 'million' 'plus' 'investment' 'uk'
TOPIC 35: 'iraq' 'iraqi' 'u.s.' 'united' 'kurdish' 'northern' 'military' 'states' 'iran' 'gulf'
TOPIC 32: 'union' 'workers' 'strike' 'government' 'percent' 'italian' 'wage' 'state' 'pay' 'sao'
TOPIC 25: 'government' 'billion' 'budget' 'tax' 'state' 'minister' 'year' 'finance' 'economic' 'ministry'
TOPIC 42: 'percent' 'point' 'year' 'july' 'june' 'billion' 'month' 'rose' 'growth' 'august'
TOPIC 19: 'german' 'marks' 'tobacco' 'car' 'germany' 'industry' 'sales' 'ag' 'million' 'new'
TOPIC 26: 'court' 'u.s.' 'lloyd's' 'case' 'plan' 'judge' 'names' 'legal' 'law' 'investors'
TOPIC 16: 'people' 'officials' 'police' 'plane' 'miles' 'killed' 'passengers' 'spokesman' 'flight' 'airport'

References

Allan, J., Carbonell, J., Doddington, G., Yamron, J., & Yang, Y. (1998). Topic detection and tracking pilot study: Final report. In *Proceedings of the DARPA broadcast news transcription and understanding workshop, Lansdowne, VA, USA* (pp. 194–218).

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

Bestgen, Y. (2001). Improving text segmentation using latent semantic analysis: A reanalysis of Choi, Wiemer-Hastings, and Moore (2001). *Computational Linguistics*, 0891-2017, *32*(1), 5–12.

Blei, D. M., & Moreno, P. J. (2001). Topic segmentation with an aspect hidden Markov model. In *Proceedings of ACM special interest group on information retrieval, New Orleans, Louisiana, USA* (pp. 343–348).

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2002). Latent Dirichlet allocation. In T. G. Dietterich, S. Becker, & Z. Ghahramani (Eds.). *Advances in neural information processing systems (NIPS)* (Vol. 14, pp. 601–608). Cambridge, MA: MIT Press.

Brants, T., Chen, F., & Tsochantaridis, I. (2002). Topic-based document segmentation with probabilistic latent semantic analysis. In *Proceedings of the international conference on information and knowledge management* (pp. 211–218). McLean, Virginia, USA: ACM.

Choi, F. Y. Y. (2000). Advances in domain independent linear text segmentation. In *Proceedings of the conference of north american chapter of the ACL, Seattle, WA, USA* (pp. 26–33).

Choi, F., Wiemer-Hastings, P., & Moore, J. (2001). Latent semantic analysis for text segmentation. In *Proceedings of EMNLP, Pittsburgh, PA, USA* (pp. 109–117).

Chua, T.-S., Chang, S.-F., Chaisorn, L., & Hsu, W. (2004). Story boundary detection in large broadcast news video archives: Techniques, experience and trends. In *MM*. 1-58113-893-8 (pp. 656–659). New York, NY, USA: ACM.

Eisenstein, J., Barzilay, R., & Davis, R. (2008). Gestural cohesion for topic segmentation. In *Proceedings of ACL-08: HLT, Columbus, Ohio, USA* (pp. 852–860).

Fragkou, P., Petridis, V., & Kehagias, A. (2004). A dynamic programming algorithm for linear text segmentation. *Journal of Intelligent Information System*, 0925-9902, *23*(2), 179–197.

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences, 101*(supl. 1), 5228–5235.

Halliday, M. A. K., & Hasan, R. (1976). *Cohesion in English*. London: Longman.

Hearst, M. (1997). TextTiling: Segmenting texts into multi-paragraph subtopic passages. *Computational Linguistics, 23*(1), 33–64.

Heidel, A., an Chang, H., & shan Lee, L. (2007). Language model adaptation using latent Dirichlet allocation and an efficient topic inference algorithm. In *Proceedings of EuroSpeech, Antwerp, Belgium*.

Hofmann, T. (2001). Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning Journal, 42*(1), 177–196.

Kehagias, A., Pavlina, F., & Petridis, V. (2003). Linear text segmentation using a dynamic programming algorithm. In *Proceedings of the tenth conference on European chapter of the association for computational linguistics, Budapest, Hungary* (pp. 171–178).

Kozima, H. (1993). Text segmentation based on similarity between words. In *Meeting of the association for computational linguistics, Ohio, USA* (pp. 286–288).

Lewis, D. D., Yang, Y., Rose, T., & Li, F. (2004). RCV1: A new benchmark collection for text categorization research. *Machine Learning Research, 5*, 361–397.

Malioutov, I., & Barzilay, R. (2006). Minimum cut model for spoken lecture segmentation. In *Proceedings of the 21st international conference on computational linguistics and 44th annual meeting of the association for computational linguistics, Sydney, Australia* (pp. 25–32).

Minka, T., & Lafferty, J. (2002). Expectation–propagation for the generative aspect model. In *Proceedings of the conference on uncertainty in artificial intelligence, Edmonton, Alberta, Canada* (pp. 352–359).

Misra, H., Cappé, O., & Yvon, F. (2008). Using LDA to detect semantically incoherent documents. In *Proceedings of CoNLL, Manchester, UK* (pp. 41–48).

Morris, J., & Hirst, G. (1991). Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational Linguistics, 17*(1), 21–48.

Mulbregt, P. V., Carp, I., Gillick, L., Lowe, S., & Yamron, J. (1998). Text segmentation and topic tracking on broadcast news via a hidden Markov model approach. In *Proceedings of ICSLP, Sydney, Australia* (pp. 2519–2522).

Nigam, K., McCallum, A. K., Thrun, S., & Mitchell, T. M. (2000). Text classification from labeled and unlabeled documents using EM. *Machine Learning, 39*(2/3), 103–134.

Ponte, J. M., & Croft, W. B. (1997). Text segmentation by topic. In *European conference on digital libraries, Pisa, Italy* (pp. 113–125).

Reynar, J. C. (1998). Topic segmentation: Algorithms and applications. Ph.D. thesis, University of Pennsylvania.

Rigouste, L., Cappé, O., & Yvon, F. (2007). Inference and evaluation of the multinomial mixture model for text clustering. *Information Processing and Management, 43*(5), 1260–1280.

Stokes, N., Carthy, J., & Smeaton, A. F. (2004). SeLeCT: A lexical cohesion based news story segmentation system. *Journal of AI Communications, 17*(1), 3–12.

Sun, Q., Li, R., Luo, D., & Wu, S. (2008). Text segmentation with LDA-based Fisher kernel. In *Proceedings of ACL-08: HLT, Short Papers, Association for Computational Linguistics, Columbus, Ohio* (pp. 269–272).

Tam, Y-C., & Schultz, T. (2006). Unsupervised language model adaptation using latent semantic marginals. In *Proceedings of ICSLP, Pittsburgh, PA, USA*.

TREC Video Retrieval Evaluation (2003). <http://wwwnlpir.nist.gov/projects/tv2003/tv2003.html>.

Utiyama, M., & Isahara, H. (2001). A statistical model for domain-independent text segmentation. In *Meeting of the association for computational linguistics, Bergen, Norway* (pp. 491–498).
Wilkinson, R. (1994). Effective retrieval of structured documents. In *Proceedings of ACM special interest group on information retrieval, Dublin, Ireland* (pp. 311–317).
Xing, D., & Girolami, M. (2007). Employing latent Dirichlet allocation for fraud detection in telecommunications. *Pattern Recognition Letters, 28*(13), 1727–1734.
Youmans, G. (1991). A new tool for discourse analysis: The vocabulary-management profile. *Language, 67*, 763–789.