Modeling Latent Topics and Temporal Distance for Story Segmentation of Broadcast News

Hongjie Chen, *Student Member, IEEE*, Lei Xie, *Senior Member, IEEE*, Cheung-Chi Leung, *Member, IEEE*, Xiaoming Lu, *Student Member, IEEE*, Bin Ma, *Senior Member, IEEE*, and Haizhou Li, *Fellow, IEEE*

***Abstract*—This paper studies a strategy to model latent topics and temporal distance of text blocks for story segmentation, that we call graph regularization in topic modeling or GRTM. We propose two novel approaches that consider both temporal distance and lexical similarity of text blocks, collectively referred to as data proximity, in learning latent topic representation, where a graph regularizer is involved to derive the latent topic representation while preserving data proximity. In the first approach, we extend the idea of Laplacian probabilistic latent semantic analysis (LapPLSA) by introducing a distance penalty function in the affinity matrix of a graph for latent topic estimation. The estimated latent topic distributions are used to replace the traditional term-frequency vectors as the data representation of the text blocks and to measure the cohesive strength between them. In the second approach, we perform Laplacian eigenmaps, which makes use of the graph regularizer for dimensionality reduction, on latent topic distributions estimated by conventional topic modeling. We conduct the experiments on the automatic speech recognition transcripts of the TDT2 English broadcast news corpus. The experiments show the proposed strategy outperforms the conventional techniques. LapPLSA performs the best with the highest F1-measure of 0.816. The effects of the penalty constant in the distance penalty function, the number of latent topics, and the size of training data on the segmentation performances are also studied.**

***Index Terms*—Graph regularization, Laplacian probabilistic latent semantic analysis, Laplacian eigenmaps, topic segmentation, topic modeling.**

I. Introduction

With the explosive growth of multimedia content (i.e. audio, video, or text), it becomes a challenge for user to retrieve the exact information from a large database. Story segmentation is the task of breaking down a multimedia stream into homogenous units each embodying a main topic or coherent story [1] for ease of information access. Studies in automatic segmentation have borne fruit over the last few decades, facilitating applications including information retrieval, text summarization, segmentation of video feeds, etc.

A half-hour broadcast news program usually consists of multiple stories. Indexing such a broadcast news program, it is desirable to divide the news program into a number of independent stories. Manual segmentation is accurate but labor-intensive and costly both in time and effort. Therefore, automatic story segmentation is highly demanded. In this paper, we are interested in story segmentation using the spoken content in the audio.

While story segmentation can be done using acoustic/prosodic cues in the audio stream [2]–[4], e.g., speaker change, significant pause and pitch reset, the lexical-cohesion based approaches that originate from text segmentation have been shown to be effective for automatic segmentation of broadcast news [2], [4], [5].

In such approaches, in addition to performing segmentation using speech transcripts, working directly on acoustic features using segmental dynamic time warping [6], [7] has been explored. In this paper, we study the use of speech transcripts provided by automatic speech recognition (ASR).

An ASR transcript is first segmented into a sequence of text blocks, each having a fixed number of terms or/and being separated by pause interval, from which lexical cues are extracted. Such story segmentation task can be considered as detecting whether a text block involves the change of story.

By lexical cohesion, we believe that the terms in a coherent story are likely to be semantically related, and different stories tend to use different sets of terms. If we represent a text block as a term-frequency vector, which is called a data point, the cohesive strength between two text blocks can be measured by the lexical similarity (e.g. cosine similarity) between the two term-frequency vectors [8].

The traditional lexical-cohesion based approaches only consider term statistics, without taking polysemy and synonymy into consideration. To address the problems, topic modeling techniques were studied to provide semantic level similarity measurement. Probabilistic latent semantic analysis (PLSA) [9] and latent Dirichlet allocation (LDA) [10] have recently become the commonly used topic modeling techniques for story segmentation and related tasks [11]–[13]. Note that topic modeling can be viewed as a kind of dimensionality reduction techniques, in which the term statistics in a text block is projected to a vector of the distribution of latent topics.

Recently, Laplacian eigenmaps (LE) [14], as a manifold learning algorithm, has been adopted for dimensionality reduction in story segmentation [15]. Manifold learning is a class of

Manuscript received May 7, 2016; revised September 24, 2016; accepted October 24, 2016. Date of publication November 8, 2016; date of current version November 28, 2016. This work was supported by the National Natural Science Foundation of China under Grant 61571363. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Tatsuya Kawahara.

H. Chen, L. Xie, and X. Lu are with the School of Computer Science, Northwestern Polytechnical University, Xi'an 710129, China (e-mail: hjchen@nwpu-aslp.org; xielei21st@gmail.com; luxiaomingnpu@gmail.com).

C.-C. Leung and B. Ma are with the Institute for Infocomm Research, Agency for Science, Technology and Research (A*STAR), Singapore 138632 (e-mail: ccleung@i2r.a-star.edu.sg; mabin@i2r.a-star.edu.sg).

H. Li is with the Institute for Infocomm Research, Agency for Science, Technology and Research (A*STAR), Singapore 138632, and also with the Department of Electrical and Computer Engineering, National University of Singapore, Singapore 119077 (e-mail: haizhou.li@nus.edu.sg).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASLP.2016.2626965

dimensionality reduction algorithms based on the assumption that data come from a low-dimensional manifold embedded in a high-dimensional ambient space. Although the dimensionality of many natural data sets is usually high, they are often generated by systems with much fewer underlying degrees of freedom and thus have lower dimensionality. In the study for story segmentation, the cohesive strength between data points is typically defined by the lexical similarity between text blocks and represented by a graph. LE finds a low-dimensional representation for a text block through an affinity matrix of a time sequence of text blocks, that we call data points, aiming at preserving the intrinsic local geometric structure of the data. The low-dimensional data represented by LE has been shown more robust against ASR errors [15] than the original high dimensional term-frequency vector representation.

Considering the running broadcast news as a time sequence of text blocks, we observe that the temporal distance between text blocks plays an important role in story segmentation [15]. By incorporating the temporal distance into the LE similarity metric between data points, we achieved a better segmentation performance. Intuitively, this design is particularly suitable for story segmentation because if two text blocks are temporally far away from each other, their data points should differ very much. It is desired that such two text blocks are assigned to different stories even if they share similar lexical terms.

This observation motivates us to take into consideration both the cosine similarity and temporal distance of text blocks when we measure the data proximity, and encode the data proximity into an affinity matrix when learning the topics for story segmentation. We propose two approaches to achieve this. In the first approach, we extend the idea of Laplacian probabilistic latent semantic analysis (LapPLSA) [16] from topic modeling of text documents to that of sequential text blocks in transcripts [17]. This approach makes use of a graph Laplacian as a regularization term in the objective function when estimating the latent topic distributions. In the second approach, we perform LE, which uses the graph Laplacian to perform dimensionality reduction, on the latent topic distributions that are estimated using PLSA and LDA. In these two approaches, the common strategy is to use the graph Laplacian to preserve the data proximity. We conduct a comprehensive study on these two approaches. The contributions of this paper are summarized as follows:

1) We propose a measure of similarity that incorporates both cosine similarity and temporal distance between text blocks in topic modeling.
2) We propose to preserve the data proximity in learning latent topic representation for story segmentations, and the proposed approaches outperform conventional topic modeling.
3) We conduct a comparative study among LapPLSA, Laplacian eigenmaps and conventional topic modeling in story segmentation in face of ASR errors.

Note that there are other techniques, such as locally-consistent topic modeling (LTM) [18], which also consider the nearest neighbors of data points in topic modeling. LTM uses the Kullback-Leibler divergence instead of the Euclidean distance to describe the proximity between two latent topic distributions. The performance of LTM in document classification and clustering is similar to that of LapPLSA as reported in [19]. We would like to highlight that the focus of this study is the integration of lexical similarity and temporal distance of text blocks in story segmentation, thus LapPLSA here is used as a utility to investigate the effectiveness of our strategy.

We also take note that, apart from LE, other manifold learning algorithms, such as Isomap [20] and locality preserving projection (LPP) [21], also offer the low-dimensional data representation of text blocks. These manifold learning algorithms are related in the sense that, they represent the similarity of data points by a graph, and they obtain low-dimensional embedded representation by eigen-decomposition. However, they are motivated to address different research problems, and may not be suitable for this study, for example, Isomap is less robust to noisy data than LE [22], which is perhaps not adequate for imperfect ASR transcripts. LPP assumes that the mapping from the original data representation to the low-dimensional one is linear, which is not well grounded.

The rest of this paper is organized as follows: in Section II, we give a brief overview on previous works related to the use of topic modeling for broadcast news story segmentation and provide the background of topic modeling.

Section III briefly presents the graph Laplacians for preserving locality of data. Section IV elaborates two ways to exploit graph Laplacians in topic modeling. We present the overall procedure of story segmentation in Section V. Experimental setup are presented in Section VI, and experimental results and analysis are provided in Section VII. Finally, we conclude in Section VIII.

II. Background of Topic Modeling

Recently many studies have evidenced a paradigm shift from bag-of-words modeling to topic modeling for data representation in lexical cohesion based story segmentation. The use of topic modeling is found in different kinds of lexical cohesion based approaches, which can be classified into local minimum search in lexical similarity [12], [23], [24], similarity based clustering [25]–[27] (also known as global optimization search) and probabilistic modeling [13].

Topic model is a statistical model for discovering the latent topic representations from a collection of documents. Latent semantic analysis (LSA) [28] is the earliest topic modeling approach to the segmentation of text [25] or spoken documents [24]. PLSA [9] is a probabilistic variant of LSA that offers a principled statistical formulation. PLSA has been shown to outperform LSA in story segmentation [11]. In probabilistic topic modeling, each document is represented as a weighted mixture of latent topics, and each latent topic is represented as a weighted mixture of terms. The weights in a mixture of latent topics form a latent topic distribution, which represents the document. The representation with latent topic distribution facilitates topic level comparison between documents. Generally the dimension of latent topic representation is significantly lower than that of term-frequency vector representation.

Despite the remarkable success of PLSA in broadcast news story segmentation, one notices that the number of parameters in PLSA grows linearly with the size of the corpora. This is not desirable especially when a considerable amount of data are involved [29]. To overcome this, Blei et al. proposed the latent Dirichlet allocation (LDA) technique [10], in which the topics of each document are described by a multinomial distribution with a set of parameters generated from a Dirichlet distribution. LDA has been proved to be effective in many related tasks [13], [30], [31]. In [13], it was studied to jointly carry out story segmentation and LDA topic modeling without the need of explicit story boundaries in training data.

In topic modeling, typically a text block is treated as a document, while the temporal distance between the text blocks is not considered. Let's first revisit some basic probabilistic topic models. Formally, we define the following notations which will be used throughout this section:

1) N text blocks are denoted by $D = \{d_1, d_2, \ldots, d_N\}$;
2) M unique terms that appear in the text blocks are denoted by $W = \{w_1, w_2, \ldots, w_M\}$;
3) K latent topics are denoted by $Z = \{z_1, z_2, \ldots, z_K\}$.

A. Probabilistic Latent Semantic Analysis

Given the document corpus D and the term set W, probabilistic latent topic analysis (PLSA) considers each term-document co-occurrence, i.e., the occurrence of a term $w_m \in W$ in a particular document $d_n \in D$, is associated with latent variables $z_1, z_2, \ldots, z_K$. These latent variables can be considered as class labels or latent topics.

The joint probability of co-occurrence pair $(d_n, w_m) \in (D \times W)$ is defined as follows:

$$P(d_n, w_m) = P(d_n) \sum_{k=1}^{K} P(w_m|z_k) P(z_k|d_n) \tag{1}$$

in which the conditional probabilities $P(w_m|z_k)$ and $P(z_k|d_n)$ can be estimated by maximizing the log-likelihood:

$$\zeta_{\mathrm{PLSA}} = \sum_{m=1}^{M} \sum_{n=1}^{N} \#(d_n, w_m) \log P(d_n, w_m) \tag{2}$$

where $\#(d_n, w_m)$ is the number of occurrences of term w_m in document d_n.

The estimation of the conditional probabilities is performed using expectation maximization (EM). EM alternates two steps [32], [33]: (i) an expectation (E) step where posterior probabilities are computed for the latent variables according to the following equation:

$$P(z_k|d_n, w_m) = \frac{P(w_m|z_k)P(z_k|d_n)}{\sum_{l=1}^{K} P(w_m|z_l)P(z_l|d_n)}; \tag{3}$$

(ii) an maximization (M) step, where parameters $P(w_m|z_k)$ and $P(z_k|d_n)$ in Eq.(3) are updated according to the following formulae:

$$P(w_m|z_k) = \frac{\sum_{n=1}^{N} \#(d_n, w_m) P(z_k|d_n, w_m)}{\sum_{n=1}^{N} \sum_{m=1}^{M} \#(d_n, w_m) P(z_k|d_n, w_m)}, \tag{4}$$

$$P(z_k|d_n) = \frac{\sum_{m=1}^{M} \#(d_n, w_m) P(z_k|d_n, w_m)}{\sum_{l=1}^{K} \sum_{m=1}^{M} \#(d_n, w_m) P(z_l|d_n, w_m)}. \tag{5}$$

PLSA alternately applies the E-step and M-step until a convergence threshold is met.

B. Latent Dirichlet Allocation

Latent Dirichlet allocation (LDA) [10] is a generative probabilistic model of a document collection. LDA considers that documents are represented as random mixtures (drawn from a Dirichlet distribution) over latent topics, though both PLSA and LDA consider each topic as a distribution over terms.

The generative process of LDA can be summarized as follows:

1) For each document $d_n \in D$, pick a topic distribution $\boldsymbol{\theta}_n = (\theta_{n1}, \theta_{n2}, \ldots, \theta_{nK})$ from a Dirichlet distribution $Dir(\boldsymbol{\alpha})$ (denoted as $\boldsymbol{\theta}_n|\boldsymbol{\alpha} \sim Dir(\boldsymbol{\alpha})$) according to the following probability density function:

$$P(\boldsymbol{\theta}_n|\boldsymbol{\alpha}) = \frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \theta_{n1}^{\alpha_1 - 1} \cdots \theta_{nK}^{\alpha_K - 1}, \tag{6}$$

where $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_K)^T$ is a Dirichlet prior on the topic distributions with components $\alpha_k > 0$, and $\Gamma(\cdot)$ is the Gamma function.

2) For each term in document d_n, select a topic z_k from the $\boldsymbol{\theta}_n$-specific multinomial distribution, denoted as $z_k|\boldsymbol{\theta}_n \sim Multi(\boldsymbol{\theta}_n)$.

3) Select a term w from $P(w|z_k, \boldsymbol{\beta})$, which is a multinomial distribution over terms in W. Here $\boldsymbol{\beta}$ is a $K \times M$ matrix which defines the term distributions; $\boldsymbol{\beta}_k = (\beta_{k1}, \beta_{k2}, \ldots, \beta_{kM})$ is the distribution over terms in W for the latent topic z_k .

Variational inference [10] or Gibbs sampling [34] can be used to estimate the parameters in the LDA model. LDA assumes that the per-document topic distribution $\boldsymbol{\theta}_n$ is generated from the K-dimensional Dirichlet distribution rather than a large set of individual parameters, which are directly linked to the training data. In this way, LDA overcomes the overfitting problem.

III. Background of Graph Laplacians

We have seen the use of graph Laplacians [35] in various tasks in machine learning, including clustering [36], [37], manifold learning [38] and semi-supervised learning [39], [40]. Graph Laplacians are widely used to preserve the local geometric structure of data in an optimization task. Suppose that we have N data points $\{\mathbf{x}_n\}_{n=1}^{N}$, where $\mathbf{x}_n \in \mathbb{R}^M$. For the graph of these data points, we define an affinity matrix S whose elements are pair-wise similarity among them. Given the affinity matrix S, a corresponding graph Laplacian $\mathbf{L}$ can be defined typically as $\mathbf{L} = \mathbf{C} - \mathbf{S}$, where $\mathbf{C}$ is a diagonal matrix with elements $c_{ii} = \sum_{j=1}^{N} s_{\mathrm{ij}}$.

To find a data representation to preserve the local geometric structure of data, a function is usually defined as:

$$\mathfrak{L} = \frac{1}{2}\sum_{i,j=1}^{N} \parallel \mathbf{y}_i - \mathbf{y}_j \parallel^2 s_{\rm ij} = \sum_{q=1}^{Q} \mathbf{f}_q^T \mathbf{L}\mathbf{f}_q, \tag{7}$$

where $\mathbf{y}_i$ is the Q-dimensional output representation for $\mathbf{x}_i$, $\mathbf{f}_q = [y_1^q, y_2^q, ..., y_N^q]^T$, and y_n^q is the q-th element of $\mathbf{y}_n$.

The locality preserving property is obtained when $\mathfrak{L}$ is minimized (e.g. $\parallel \mathbf{y}_i - \mathbf{y}_j \parallel^2 s_{\rm ij}$ is small for all i, j). Intuitively speaking, the minimization of $\mathfrak{L}$ ensures that when the similarity connection $s_{\rm ij}$ between $\mathbf{x}_i$ and $\mathbf{x}_j$ is strong (i.e. the value is large), the distance between their low-dimensional representations $\mathbf{y}_i$ and $\mathbf{y}_j$ remains small. Hence, we maintain that the geometrical relationship between $\mathbf{y}_i$ and $\mathbf{y}_j$ is similar to that between $\mathbf{x}_i$ and $\mathbf{x}_j$.

In Laplacian eigenmaps, as a geometrically motivated algorithm for data representation, $\mathfrak{L}$ serves as an objective function that is minimized through eigen-decomposition, as will be described in Section IV-B. Note that there exist other manifold learning algorithms, such as locally linear embedding [41] and locality preserving projection [21]. They are related to Laplacian eigenmaps in one way or another.

The function $\mathfrak{L}$ can also be used as a regularization term in many objective functions for various tasks. In topic modeling, Laplacian PLSA [16] uses $\mathfrak{L}$ as a regularization term to be minimized when maximizing the log-likelihood of a data set. The optimization of its objective function is performed by the generalized expectation maximization (GEM) algorithm [32], which will be briefly presented in Section IV-A. It is worth noting that the regularization term has also been used in other data representation algorithms [42]–[44] for similar purposes.

IV. Graph Regularization in Topic Modeling

In this paper, we study a Graph Regularization strategy in Topic Modeling (GRTM) to learn the latent topic representation for story segmentation. In the two approaches under the GRTM strategy, we consider both temporal distance and lexical similarity of text blocks (collectively referred to as data proximity) in learning latent topic representation. A graph regularizer is involved to derive the latent topic representation of the text blocks in each type of approach.

The relationship between the text blocks is represented by a graph. In the graph, each vertex represents a text block, and an edge between two vertices represents the proximity between the two text blocks. Then a document manifold can be approximated through the graph.

We define the affinity matrix $\mathbf{S} = \{s_{\rm ij}\}_{i,j=1}^{N}$ to denote the proximity between the text block pairs. We adopt the cosine measure between text block pairs to depict their lexical similarity, and we incorporate the temporal distance between text blocks into the affinity matrix. If two text blocks d_i and d_j in the training data come from the same story, we put an edge between nodes i and j and define $s_{\rm ij}$ as follows:

$$s_{\rm ij} = \cos(\mathbf{x}_i, \mathbf{x}_j)\mu^{|i-j|} \tag{8}$$

where $\mathbf{x}_i$ and $\mathbf{x}_j$ represent the term-frequency vectors of the text block $\mathbf{d}_i$ and $\mathbf{d}_j$, $\cos(\mathbf{x}_i, \mathbf{x}_j)$ is the cosine similarity measure between $\mathbf{x}_i$ and $\mathbf{x}_j$, $\mu^{|i-j|}$ is the penalty as a function of the temporal distance between i and j, and μ is a penalty constant smaller than 1.0 that is tuned from a set of development data. The penalty function is expected to reduce $s_{\rm ij}$ dramatically when the temporal distance is large. If they are not from the same story, we set $s_{\rm ij}$ to zero.

Moreover, the affinity matrix is inevitably affected by noise because of lexical choice, even if there are no ASR errors. Such noise can be significantly reduced by adding the distance penalty factor to the affinity matrix. An affinity matrix can be visualized using a two dimensional dotplot as shown in Fig. 1. Higher similarity values are represented by darker dots in the figure. For brevity, Fig. 1(a) and (b) plot the affinity matrices of a broadcast news program. In the ideal case, we expect some dark squares along the diagonal of the dotplot, and the edges of such squares suggest the story boundaries.

However, in reality as illustrated in Fig. 1(a), we observe that the dark squares are not salient enough for direct story boundary detection. There are also light dots in the dark squares and considerable dark dots off the diagonal. As shown in Fig. 1(b), the distance penalty factor effectively suppresses many dark dots off the diagonal.

A. LapPLSA: Graph Laplacian as Regularization Term in Objective Function

Neither PLSA nor LDA considers the local geometrical structure in the documents. In the generative process of topic modeling, there is no decipherable relation between $P_D = \{P(d_n)\}_{n=1}^{N}$ and the conditional probability distribution $P(z|d_n)$. In contrast, Laplacian probabilistic latent semantic analysis (LapPLSA) [16] makes a specific assumption about the connection between P_D and $P(z|d_n)$. If two text blocks $d_i, d_j \in D$ are close in the intrinsic geometry of P_D space, LapPLSA forces the conditional probability distributions $P(z|d_i)$ and $P(z|d_j)$ similar to each other.

It has been shown that LE approach to story segmentation benefits consistently from modeling the temporal distance [15]. In this paper, we continue to explore the use of temporal distance under the LapPLSA framework [16]. We are inspired by the fact that the regularization framework in LapPLSA has been proven successful in information retrieval [45], [46] and social network analysis [47].

The parameters of LapPLSA are estimated by maximizing the following regularized log-likelihood:

$$\zeta_{\rm LapPLSA} = \zeta_{\rm PLSA} - \lambda \sum_{k=1}^{K} \Re_k, \tag{9}$$

$$\Re_k = \frac{1}{2}\sum_{i,j=1}^{N} \Big(P(z_k|d_i) - P(z_k|d_j) \Big)^2 s_{\rm ij}, \tag{10}$$

where λ is a regularization weight, $\zeta_{\rm PLSA}$ is the log-likelihood defined as in Eq.(2) for PLSA. Note that the regularization term $\Re_k$ that corresponds to the latent topic k can be

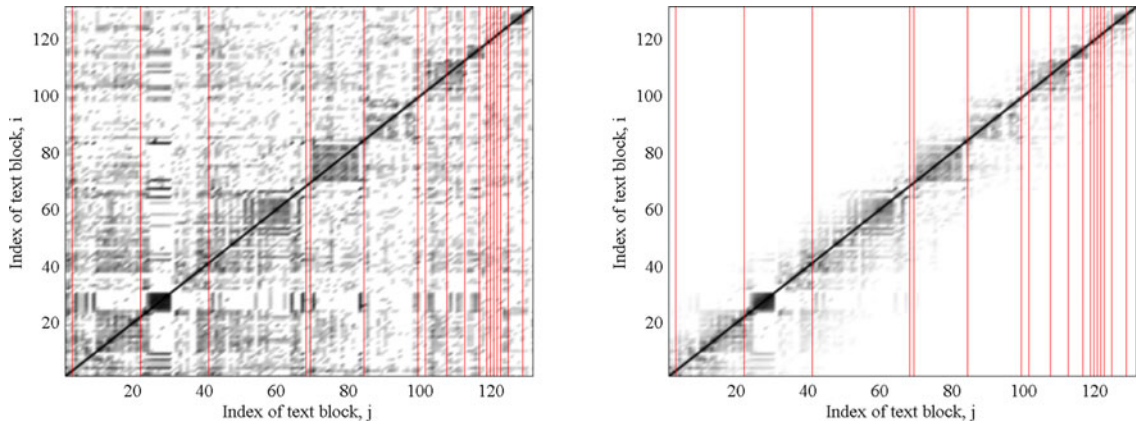

Fig. 1. Similarity between text blocks represented in an affinity matrix S: (a) only cosine similarity $\cos(\mathbf{x}_i, \mathbf{x}_j)$ is used; (b) cosine similarity $\cos(\mathbf{x}_i, \mathbf{x}_j)$ and penalty function of temporal distance $\mu^{|i-j|}$ are used.

written in the form of $\mathbf{f}_k^T \mathbf{L} \mathbf{f}_k$ similar to Eq.(7) where $\mathbf{f}_k = [P(z_k|d_1), P(z_k|d_2), ..., P(z_k|d_N)]^T$. The regularization term ensures that the geometrical relationship between $P(z_k|d_i)$ and $P(z_k|d_j)$ is similar to that between d_i and d_j.

It should be noted that data points in [16] are independent text documents. But in this paper, a document is a text block in the running text stream. Furthermore, the regularization term in this paper involves both the temporal distance and lexical similarity of text blocks as in Eq.(8), which is different from [16] that only considers the lexical similarity in the geometrical relationship of documents.

We follow the generalized EM (GEM) algorithm in [32] for parameter estimation. Similar to the standard EM for the parameter estimation of PLSA, the GEM algorithm of LapPLSA are updated iteratively by an E-step and an M-step until a convergence threshold is met. The E-step in LapPLSA is the same as that in PLSA. However, the M-step of the GEM algorithm finds parameters that merely increase the expected data log-likelihood rather than maximizing it.

B. Dimensionality Reduction on Latent Topic Distributions With Graph Laplacian

We can also encode the data proximity of text blocks in a new way by performing Laplacian eignmaps (LE) on the data representations in form of latent topic distributions, which exploits LE to conduct graph Laplacian regularization directly on latent topic distributions. This approach works for any specific topic modeling algorithms, and comes in handy when a topic model is readily available. Moreover, LE has been proven effective in characterizing times sequence of text blocks. Especially, the low-dimensional data representation obtained from LE is relatively more robust against ASR errors [15]. This motivates us to study its interaction with two commonly used topic modeling algorithms, namely PLSA and LDA, in this section.

We use LE to process the text blocks, which are represented using latent topic distributions. We use $\mathbf{X} = \{\mathbf{x}_n\}_{n=1}^N$ where $\mathbf{x}_n \in \mathbb{R}^K$ to represent the vectors of latent topic distributions as input data. We use $\mathbf{Y} = \{\mathbf{y}_n\}_{n=1}^N$ ($\mathbf{y}_n \in \mathbb{R}^Q, Q \leq K < M$, and $\mathbf{Y} \in \mathbb{R}^{N \times Q}$) to represent the output embedded representation.

In LE, the following function similar to Eq.(7) is used as the objective function to be minimized:

$$\frac{1}{2} \sum_{i,j=1}^{N} \| \mathbf{y}_i - \mathbf{y}_j \|^2 s_{\mathrm{ij}} = \sum_{q=1}^{Q} \mathbf{f}_q^T \mathbf{L} \mathbf{f}_q \tag{11}$$

$$= \mathrm{tr}(\mathbf{Y}^T \mathbf{L} \mathbf{Y}), \tag{12}$$

where $\mathbf{f}_q = [y_1^q, y_2^q, ..., y_N^q]^T$ and y_n^q is the q-th element of $\mathbf{y}_n$. Again we define s_{ij} as in Eq.(8) in order to include both temporal distance and lexical similarity of text blocks in the affinity matrix. Note that Eq.(12) rewrites the objective function in form of a graph regularizer, $\sum_{q=1}^{Q} \mathbf{f}_q^T \mathbf{L} \mathbf{f}_q$, which preserves the data proximity of text blocks within the output embedded representation.

A zero matrix and other matrices with the rank less than Q are also the solutions to minimize $\mathrm{tr}(\mathbf{Y}^T \mathbf{L} \mathbf{Y})$ in Eq.(11), but they are meaningless for the task. To prevent this from happening, we impose the constraint $\mathbf{Y}^T \mathbf{C} \mathbf{Y} = \mathbf{I}$ where $\mathbf{I}$ is an identity matrix. By the Rayleigh-Ritz theorem [48], we can obtain the solution by using the matrix of eigenvectors corresponding to the Q smallest eigenvalues of the generalized eigenvector problem:

$$\mathbf{L}\mathbf{v} = \lambda \mathbf{C}\mathbf{v}. \tag{13}$$

With this formula, we can stack the N-dimensional eigenvectors $\mathbf{v}'_1, \mathbf{v}'_2, \ldots, \mathbf{v}'_Q$ in the order of their eigenvalues $\lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_Q$ to approximate the mapping matrix $\mathbf{Y}$.

V. Practical Issues in Implementation

The architecture of our story segmentation system is illustrated in Fig. 2. The ASR transcript is first segmented into a sequence of text blocks. The starting point of each text block marks a candidate of story boundary. The story segmentation can be formulated as a detection task to determine whether those candidates are story boundaries.

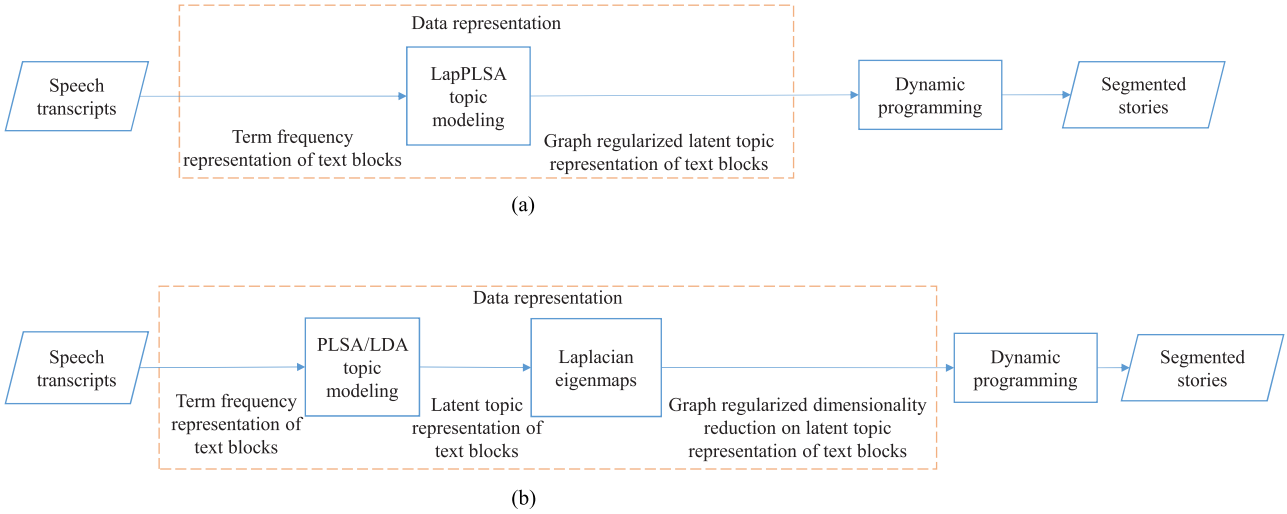

Fig. 2. Architecture of the proposed systems that segment automatic speech recognition (ASR) transcripts or manual transcripts into individual stories with (a) topic modeling that is based on the LapPLSA framework and (b) topic modeling (e.g. PLSA and LDA) followed by LE.

We are interested in how to effectively characterize text blocks with latent topic distributions. As illustrated in Fig. 2, the term-frequency vector of each text block is projected to a vector of latent topic distribution in both approaches. Note that the parameter estimation in topic modeling from a set of training data is not shown in the figure. The story boundary detection is performed on the low dimensional vectors, each of which represents a text block.

A. Estimation of Latent Topic Distribution

In PLSA and LapPLSA, the term-frequency vectors of the text blocks from a set of training data are used for parameter estimation as described in Section II-A and Section IV-A respectively. While each text block in the training set is a complete story, a text block in the test set contains a sequence of terms segmented by pauses or a predefined number of terms, which may or may not constitute a complete story. The parameter estimation yields two sets of parameters: the term distributions over latent topics $P(w_m|z_k)$, and the topic distributions of training text blocks $P(z_k|d_n)$. The former is used to perform a folding-in process to obtain $P(z_k|d')$ for an unseen text block d' in the test set.

In LDA, the hyperparameters (including $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$) can be estimated using the text blocks from the training data set. Given the hyperparameters, the topic distribution $\boldsymbol{\theta}'$ of each unseen text block d' from the test set is estimated. The above parameter estimation can be done by variational inference [10] or Gibbs sampling [34]. We employed a C implementation[1] of variational inference for LDA with the convergence threshold, $1 \times \mathrm{e}^{-4}$, to estimate the parameters.

B. Story Boundary Detection

The data representation only provides a means of measuring pairwise similarity between text blocks. We need a strategy to find a set of story boundaries in a running news program. A simple strategy could be that we locate the valleys in the sequence of cohesive strength scores between adjacent text blocks. Story boundaries are then identified at the positions where the cohesive strength are weaker than a pre-set threshold. One of such implementations is TextTiling, which performs well when there are salient changes in the sequence of cohesive strength scores. However, the topic transition between two adjacent news can be gradual. In such cases, the changes in the score sequence are subtle. The dynamic programming (DP) algorithm [6] is known to partly address this problem by making a global decision for a set of boundaries over the entire news program. As reported in [15], DP consistently outperforms TextTiling for news segmentation. Thus next we only use DP as the boundary detector. With DP, the global optimal solution can be obtained by minimizing:

$$\Im = \sum_{t=1}^{N_s} \Big(\sum_{i,j \in Seg_t} \| \mathbf{u}_i - \mathbf{u}_j \|^2 \Big) \tag{14}$$

where $\mathbf{u}_i$ and $\mathbf{u}_j$ are the low-dimensional representation of text block i and j respectively, Seg_t indicates the text blocks assigned to a hypothesized story, and N_s is the number of hypothesized stories. Imposing a linear constraint on the story segmentation [6], we can obtain the global minimization of Eq.(14) using a DP algorithm in polynomial time [15]. The DP procedure includes a forward process in which we compute the accumulated scores of all possible segmentation paths within the search area, and a back-tracing process in which we recover the best set of segmentation boundaries. According to the principle of Occam's razor [49], if there are multiple solutions of Eq.(14), we choose the one with the fewest segments.

VI. Experimental Setup

To evaluate the approaches using our proposed GRTM strategy, we carried out story segmentation experiments on the TDT2

[1]https://www.cs.princeton.edu/ blei/lda-c/index.html

English broadcast news corpus[2], which consists of 1,033 news programs collected from VOA World News, PRI The World, CNN Headline News and ABC World News Tonight. We divided this corpus into three non-overlapping sets:

1) a training set of 500 programs, which was collected between January and March in 1998, for parameter estimation in topic modeling and LE.
2) a development set of 133 programs, which was collected in April of 1998, for empirical parameter tuning.
3) a test set of 400 programs, which was collected between the end of April and June in 1998, for performance evaluation.

To study the adverse effect due to ASR errors, we report experiments on both manual transcripts and ASR transcripts provided in the corpus. The word error rate of the ASR transcripts is around 30%. Story boundary tags are available in the both types of transcripts. However, the time alignment information of words and pauses is only available in the ASR transcripts. At the training stage, the preprocessing steps for the ASR transcripts and the manual transcripts were the same. Word streams were divided into text blocks using the time labels of pauses in the ASR transcripts. We used word unigram as the basic term unit in the aforementioned topic models. If the pause duration between two blocks was more than 1.0 second, it was considered as a boundary candidate.

We conducted comprehensive experiments that covered seven types of data representations, as summarized in Table I. LapPLSA-DP, PLSA-LE-DP and LDA-LE-DP are proposed in this paper that use graph regularization to model data proximity. PLSA-DP and LDA-DP use conventional topic modeling without considering the data proximity.

Three widely-used evaluation metrics, including F1-measure [50], P_k metric [51] and WindowDiff [52], were used to evaluate the story segmentation performance. Higher values of F1-measure or lower values of P_k and WindowDiff indicate better segmentation performance. F1-measure is the harmonic mean of precision and recall. When using F1-measure on ASR transcripts, we followed the TDT2 evaluation rule: a detected story boundary is considered correct if it falls within a 15-second tolerant window on each side of a reference boundary. When the evaluation was on manual transcripts, we used a 40-word tolerant window on each side of a reference boundary instead. P_k measures the probability that a hypothesized segmentation is inconsistent with a reference segmentation when moving a fixed-width window. Here we set k to be half of the average reference segment length as [51] does. WindowDiff is a modification of P_k metric, which measures the probability that the number of hypothesized boundaries is not the same as the number of reference boundaries when moving a fixed-width window.

A number of parameters were set empirically based on the development set. The convergence threshold in LapPLSA, PLSA and LDA was set to 1.0×10^{-4}. Unless stated otherwise, μ in the penalty function was set to 0.9, the number of latent topics in topic modeling was set to 64, and the dimensionality after LE projection was set to 32. For a fair comparison with PLSA-LE-DP and LDA-LE-DP, the number of latent topics in LapPLSA-DP was also set to 32.

[2]http://projects.ldc.upenn.edu/TDT2/

VII. Experimental Results and Analysis

We first compare the proposed GRTM data representations with the baseline reference systems. We then study the effect of the penalty constant μ, the number of latent topics, and the size of training data during topic modeling. Through the study, we review the training and runtime behaviors of different data representation approaches, and suggest a way of setting parameters in system development.

A. Comparison of Data Representations

We report the results of seven data representation approaches in Fig. 3.

Firstly, the GRTM approaches consistently outperformed the baseline approaches on both ASR and manual transcripts in term of F1-measure, P_k and WindowDiff. And they greatly narrowed the performance difference between ASR and manual transcripts with relative 0.8–2.0% in F1-measure, relative 6.8–12.5% in P_k and relative 3.5–8.7% in WindowDiff. This suggests that the combination of latent topics and temporal distance carries complementary information in the data representation for story segmentation. Furthermore, LapPLSA outperformed PLSA/LDA-LE. Note that LapPLSA considers latent topics and temporal distance jointly in parameter estimation, while PLSA/LDA-LE models latent topics and temporal distance in a sequential fashion. Therefore, a joint consideration of latent topics and temporal distance is seen as beneficial.

Secondly, LapPLSA and PLSA/LDA-LE outperformed TF-LE, and PLSA/LDA outperformed TF on both ASR and manual transcripts. This validates the finding that modeling of latent topics is helpful in story segmentation [11], [17]. We believe that topic modeling is less sensitive to lexical variations and it handles polysemous words and synonyms in a better way.

Thirdly, PLSA/LDA-LE prevailed over PLSA/LDA for both ASR and manual transcripts. This suggests that the graph Laplacian with the consideration of temporal distance contributes to a better story segmentation. We will discuss the effect of the penalty function in the graph Laplacian, in Section VII-B. Similarly, TF-LE outperformed TF, which was consistent with the observation in [15].

In summary, the proposed GRTM strategy differs from TF-LE [15] by introducing topic modeling into data representation. It models both latent topics and temporal distance and consistently outperformed other baseline reference systems.

B. Effect of the Penalty Constant μ

Fig. 4 illustrates the effect of the penalty constant μ in PLSA-LE-DP, LDA-LE-DP and LapPLSA-DP on the development set. We observed that LapPLSA-DP consistently outperformed PLSA-LE-DP and LDA-LE-DP at different values of μ. The results on the development set suggest that $\mu = 0.9$ would be a good setting. Note that $\mu = 1.0$ is a special case when

TABLE I
A SUMMARY OF SEVEN DATA REPRESENTATIONS FOR COMPUTING THE COHESIVE STRENGTH BETWEEN TEXT BLOCKS IN THE BASELINE REFERENCE SYSTEMS, AND THE PROPOSED GRAPH REGULARIZATION IN TOPIC MODELING STRATEGY AS ILLUSTRATED IN FIG. 2

	Systems	Data Representations	Latent Topic Modeling	Temporal Distance Modeling
Baseline Systems	TF-DP [6]	Term Frequencies	No	No
	PLSA-DP	Latent Topic Distributions	Yes	No
	LDA-DP	Latent Topic Distributions	Yes	No
	TF-LE-DP [15]	Graph Regularized Dimensionality Reduction (LE) on Term Frequencies	No	Yes
Graph Regularization in Topic Modeling (GRTM)	PLSA-LE-DP	Graph Regularized Dimensionality Reduction (LE) on Latent Topic Distributions	Yes	Yes
	LDA-LE-DP	Graph Regularized Dimensionality Reduction (LE) on Latent Topic Distributions	Yes	Yes
	LapPLSA-DP	Graph Regularized Latent Topic Distributions	Yes	Yes

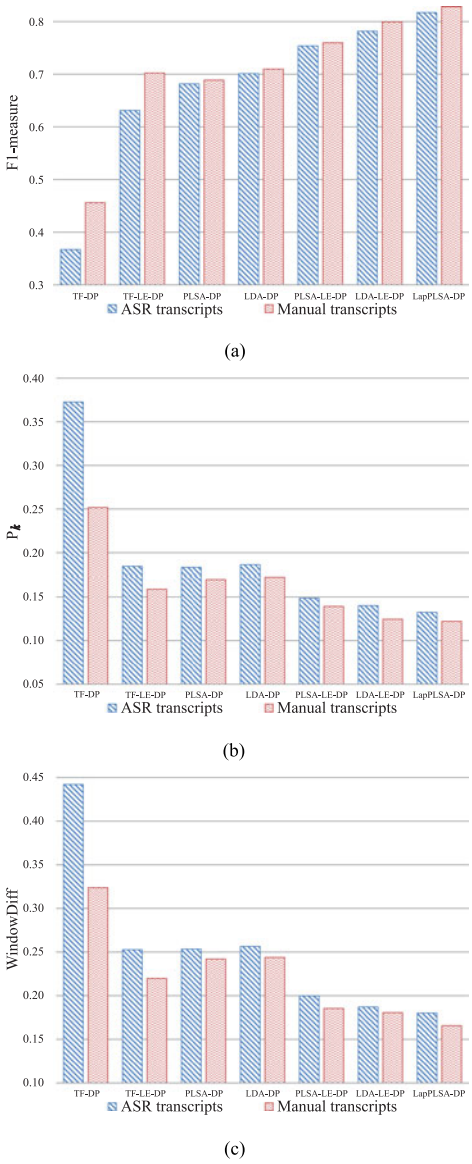

Fig. 3. Segmentation results using different data representations on ASR transcripts (blue bars) and manual transcripts (red bars) on the test set in terms of different evaluation metrics: (a) F1-measure; (b) P_k; (c) WindowDiff.

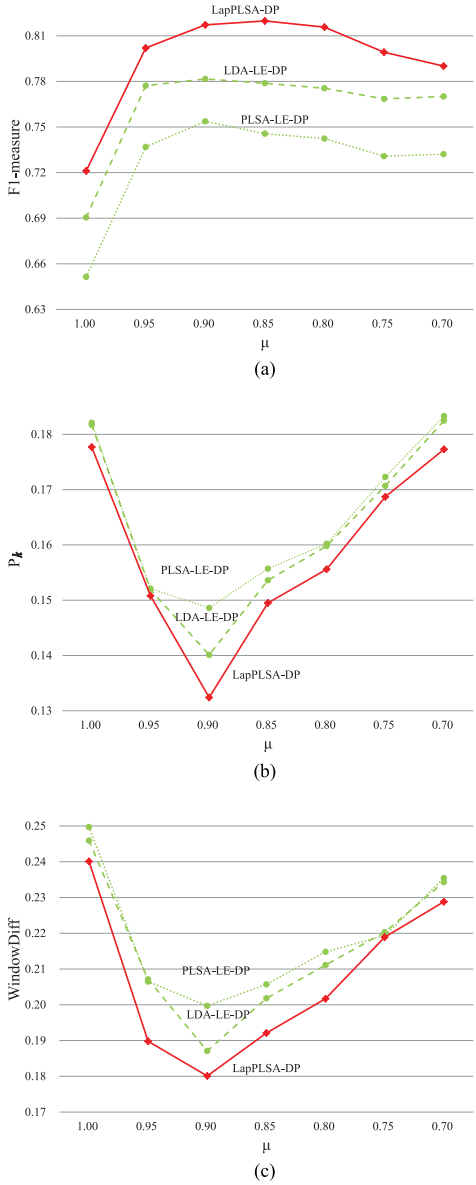

Fig. 4. Effect of the penalty constant μ on the development set. (a) F1-measure; (b) P_k; (c) WindowDiff.

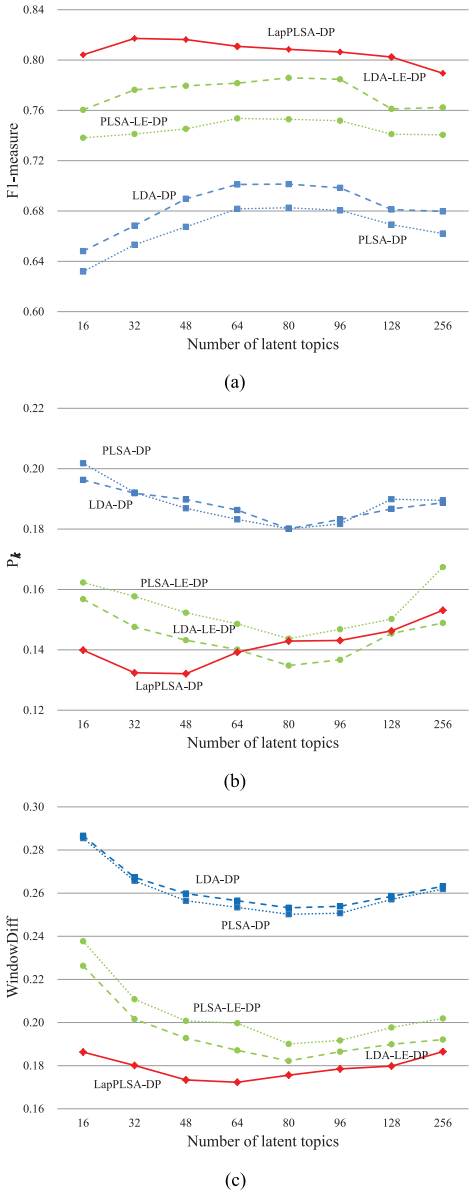

Fig. 5. Segmentation results on the development set with different number of latent topics. (a) F1-measure; (b) P_k; (c) WindowDiff.

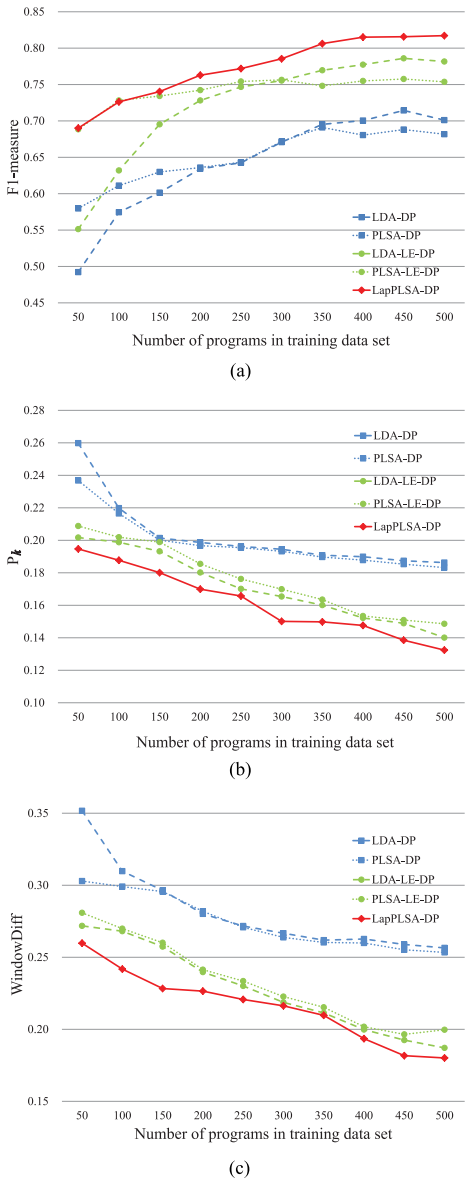

Fig. 6. Segmentation results on the development set with different amounts of training data. (a) F1-measure; (b) P_k; (c) WindowDiff.

the distance penalty function has no effect on the cosine similarity in Eq.(8) between two text blocks. For each approach illustrated in Fig. 4, comparing the performance with $\mu = 1.0$ and that with $\mu < 1.0$, we can be sure that the temporal distance is beneficial to story segmentation. The observation was similar to that in the experiments on the TDT2 Mandarin and CCTV Mandarin broadcast news corpora in [15]. This experiment also shows that the choice of μ is important under the GRTM strategy.

C. Effect of the Number of Latent Topics

Fig. 5 reports the effect of the number of latent topics on the development set for the five approaches that involve topic modeling. We observed that LE projection improved PLSA or LDA consistently at different numbers of latent topics. In terms of F1-measure and WindowDiff, LapPLSA-DP outperformed other topic modeling approaches in general.

It is worth noting that the PLSA and LDA based approaches gained optimal results when the number of latent topics was set between 64 and 96. However, LapPLSA was optimal when a smaller number of latent topics (between 32 and 64) was used. This observation suggests that LapPLSA can potentially provide a more compact, i.e. low-dimensional, data representation than PLSA and/or LDA.

D. Effect of the Size of Training Dataset

We increased the training set from 50 programs to 500 programs (by 50 programs at a time) to evaluate the effects of different of size of training data in the topic modeling related approaches. Fig. 6 illustrates the results on the development set.

We observed that all approaches in general benefited from an increasing size of training data. LapPLSA-DP achieved the best performance (F1-measure of 0.8172, P_k of 0.1324 and WindowDiff of 0.1801), when trained on 500 programs of data. It was encouraging to see that LapPLSA-DP outperformed the other four approaches in most of the cases when different sizes of training data were used. Moreover, when applying LE projection on PLSA or LDA topic distributions, we achieved consistent performance gains over the PLSA or LDA baseline counterparts.

VIII. Conclusion and Future Works

We have proposed to incorporate the temporal distance between text blocks into the similarity metric in topic modeling for text block representation in story segmentation. Two approaches under the proposed GRTM strategy consistently outperform PLSA and LDA. We find that the distance penalty function in the affinity matrix is crucial to the segmentation performance. We also find that LapPLSA provides a more compact data representation than PLSA and LDA. Although we used the DP algorithm in all experiments as a boundary detection search strategy, we believe that the proposed GRTM strategy also works for other boundary detection strategies.

While both topic modeling and manifold-based dimensionality reduction, such as LE, are robust against ASR errors, this study suggests that topic modeling is more effective than LE when the two techniques act alone. We note that the choice of penalty constant, the number of latent topics, and the size of training data are task-dependent. This study suggests a way to configure a system in a specific task.

In the future, we plan to investigate several aspects of the proposed strategy. 1) Extension to Bayesian non-parametric topic models. In this study, we assume that a development set is available for setting the parameters. As a future work, we hope to investigate how to integrate the proposed strategy into a Bayesian non-parametric framework [53], [54], when a development set is not available. 2) Generalization to online topic modeling. In this study, topic models are studied a offline batch process. We are interested in further this study for online topic modeling and story segmentation. 3) Extension to deep neural network based text block representation. The recent research on distributed representations of text based on deep neural network (DNN) models, e.g. the continuous bag-of-word model (CBOW) and the skip-gram model [55], sentence and document embedding [56]–[58], etc., represents a new way of semantic representation. It would be interesting to study the interaction between our strategy and those semantic representations for story segmentation.

Acknowledgment

The authors would like to thank the reviewers for their very constructive and helpful suggestions on the paper.

References

[1] J. Allan, *Topic Detection and Tracking: Event-Based Information Organization*. New York, NY, USA: Springer, 2002.

[2] E. Shriberg, A. Stolcke, D. Hakkani-Tür, and G. Tür, "Prosody-based automatic segmentation of speech into sentences and topics," *Speech Commun.*, vol. 32, no. 1, pp. 127–154, 2000.

[3] G. Tür, D. Hakkani-Tür, A. Stolcke, and E. Shriberg, "Integrating prosodic and lexical cues for automatic topic segmentation," *Comput. Linguistics*, vol. 27, no. 1, pp. 31–57, 2001.

[4] L. Xie, C. Liu, and H. Meng, "Combined use of speaker- and tone-normalized pitch reset with pause duration for automatic story segmentation in Mandarin broadcast news," in *Proc. Human Lang. Technol. Conf. North Amer. Assoc. Comput. Linguistics*, 2007, pp. 193–196.

[5] F. Y. Choi, "Advances in domain independent linear text segmentation," in *Proc. 1st North Amer. Chapter Assoc. Comput. Linguistics*, 2000, pp. 26–33.

[6] I. Malioutov and R. Barzilay, "Minimum cut model for spoken lecture segmentation," in *Proc. 21st Int. Conf. Comput. Linguistics 44th Annu. Meeting Assoc. Comput. Linguistics*, 2006, pp. 25–32.

[7] L. Zheng, C.-C. Leung, L. Xie, B. Ma, and H. Li, "Acoustic texttiling for story segmentation of spoken documents," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2012, pp. 5121–5124.

[8] G. Salton and M. J. McGill, *Introduction to Modern Information Retrieval*. New york, NY, USA: McGraw-Hill, 1986.

[9] T. Hofmann, "Probabilistic latent semantic indexing," in *Proc. 22nd Annu. Int. ACM/SIGIR Conf. Res. Develop. Inf. Retrieval*, 1999, pp. 50–57.

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

[11] M. Lu, C.-C. Leung, L. Xie, B. Ma, and H. Li, "Probabilistic latent semantic analysis for broadcast news story segmentation," in *Proc. Interspeech*, 2011, pp. 1109–1112.

[12] M. Riedl and C. Biemann, "Text segmentation with topic models," *J. Lang. Technol. Comput. Linguistics*, vol. 27, pp. 47–69, 2012.

[13] J.-T. Chien and C.-H. Chueh, "Topic-based hierarchical segmentation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 1, pp. 55–66, Jan. 2012.

[14] M. Belkin and P. Niyogi, "Laplacian eigenmaps for dimensionality reduction and data representation," *Neural Comput.*, vol. 15, no. 6, pp. 1373–1396, 2003.

[15] L. Xie, L. Zheng, Z. Liu, and Y. Zhang, "Laplacian eigenmaps for automatic story segmentation of broadcast news," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 1, pp. 276–289, Jan. 2012.

[16] D. Cai, Q. Mei, J. Han, and C. Zhai, "Modeling hidden topics on document manifold," in *Proc. 17th ACM Conf. Inf. Knowl. Manage.*, 2008, pp. 911–920.

[17] X. Lu, C.-C. Leung, L. Xie, B. Ma, and H. Li, "Broadcast news story segmentation using latent topics on data manifold," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2013, pp. 8465–8469.

[18] D. Cai, X. Wang, and X. He, "Probabilistic dyadic data analysis with local and global consistency," in *Proc. Int. Conf. Mach. Learn.*, 2009, pp. 105–112.

[19] S. Huh and S. E. Fienberg, "Discriminative topic modeling based on manifold learning," *ACM Trans. Knowl. Discovery Data*, vol. 5, no. 4, pp. 1–25, 2012.

[20] J. B. Tenenbaum, V. De Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, no. 5500, pp. 2319–2323, 2000.

[21] X. He and P. Niyogi, "Locality preserving projections," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, 2004, pp. 153–160.

[22] S. Gerber, T. Tasdizen, and R. T. Whitaker, "Robust non-linear dimensionality reduction using successive 1-dimensional Laplacian eigenmaps," in *Proc. Int. Symp. Chin. Spoken Lang. Process.*, 2007, pp. 281–288.

[23] M. Riedl and C. Biemann, "TopicTiling: A text segmentation algorithm based on LDA," in *Proc. ACL Student Res. Workshop*, 2012, pp. 37–42.

[24] Y. Yang and L. Xie, "Subword latent semantic analysis for texttiling-based automatic story segmentation of Chinese broadcast news," in *Proc. Int. Symp. Chin. Spoken Lang. Process.*, 2008, pp. 1–4.

[25] F. Y. Y. Choi, P. Wiemer-Hastings, and J. Moore, "Latent semantic analysis for text segmentation," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2001, pp. 109–117.

[26] Q. Sun, R. Li, D. Luo, and X. Wu, "Text segmentation with LDA-based fisher kernel," in *Proc. Annu. Meeting Assoc. Comput. Linguistics*, 2008, pp. 269–272.

[27] X. Lu, L. Xie, C.-C. Leung, B. Ma, and H. Li, "Broadcast news story segmentation using manifold learning on latent topic distributions," in *Proc. Annu. Meeting Assoc. Comput. Linguistics*, 2013, pp. 190–195.

[28] S. C. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas, and R. A. Harshman, "Indexing by latent semantic analysis," *J. Amer. Soc. Inf. Sci.*, vol. 41, no. 6, pp. 391–407, 1990.
[29] D. M. Blei, "Probabilistic topic models," *Commun. ACM*, vol. 55, no. 4, pp. 77–84, 2012.
[30] R. Arora and B. Ravindran, "Latent Dirichlet allocation based multi-document summarization," in *Proc. 2nd Workshop Anal. Noisy Unstructured Text Data*, 2008, pp. 91–97.
[31] M. Morchid, M. Bouallegue, R. Dufour, G. Linars, D. Matrouf, and R. D. Mori, "Compact multiview representation of documents based on the total variability space," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 8, pp. 1295–1308, Aug. 2015.
[32] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Roy. Statist. Soc.*, vol. 39, pp. 1–38, 1977.
[33] R. M. Neal and G. E. Hinton, "A view of the EM algorithm that justifies incremental, sparse, and other variants," in *Learning in Graphical Models*. Dordrecht, The Netherlands: Kluwer, 1998, pp. 355–368.
[34] T. Griffiths and M. Steyvers, "A probabilistic approach to semantic representation," in *Proc. 24th Annu. Conf. Cogn. Sci. Soc.*, 2002, pp. 381–386.
[35] F. Chung, *Spectral Graph Theory*. Providence, RI, USA: AMS, 1997.
[36] J. Shi and J. Malik, "Normalized cuts and image segmentation," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, 1997, pp. 731–737.
[37] A. Y. Ng, M. I. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, 2001, pp. 849–856.
[38] M. Belkin and P. Niyogi, "Towards a theoretical foundation for Laplacian-based manifold methods," *J. Comput. Syst. Sci.*, vol. 74, pp. 1289–1308, 2005.
[39] O. Chapelle, B. Schlkopf, and A. Zien, *Semi-Supervised Learning*. Cambridge, MA, USA: MIT Press, 2010.
[40] Y. Liu and K. Kirchhoff, "Graph-based semisupervised learning for acoustic modeling in automatic speech recognition," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 11, pp. 1946–1956, Nov. 2016.
[41] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5500, pp. 2323–2326, 2000.
[42] M. Zheng *et al.*, "Graph regularized sparse coding for image representation," *IEEE Trans. Image Process.*, vol. 20, no. 5, pp. 1327–1336, May 2011.
[43] D. Cai, X. He, J. Han, and T. S. Huang, "Graph regularized nonnegative matrix factorization for data representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 8, pp. 1548–1560, Aug. 2011.
[44] M. Belkin, P. Niyogi, and V. Sindhwani, "Manifold regularization: A geometric framework for learning from labeled and unlabeled examples," *J. Mach. Learn. Res.*, vol. 7, pp. 2399–2434, 2006.
[45] J. Guo, X. Cheng, G. Xu, and X. Zhu, "Intent-aware query similarity," in *Proc. 20th ACM Int. Conf. Inf. Knowl. Manage.*, 2011, pp. 259–268.
[46] J. He, V. Hollink, and A. de Vries, "Combining implicit and explicit topic representations for result diversification," in *Proc. 35th Int. ACM/SIGIR Conf. Res. Develop. Inf. Retrieval*, 2012, pp. 851–860.
[47] Q. Mei, D. Cai, D. Zhang, and C. Zhai, "Topic modeling with network regularization," in *Proc. 17th Int. Conf. World Wide Web*, 2008, pp. 101–110.
[48] H. Lütkepohl, *Handbook of Matrices*. Hoboken, NJ, USA: Wiley, 1996.
[49] W. M. Thorburn, "The myth of Occam's razor," *Mind*, vol. 27, pp. 345–353, 1918.
[50] R. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval*. Reading, MA, USA: Addison-Wesley, 1999.
[51] D. Beeferman, A. Berger, and J. Lafferty, "Statistical models for text segmentation," *Mach. Learn.*, vol. 34, nos. 1–3, pp. 177–210, 1999.
[52] L. Pevzner and M. A. Hearst, "A critique and improvement of an evaluation metric for text segmentation," *Comput. Linguistics*, vol. 28, no. 1, pp. 19–36, 2002.
[53] S. J. Gershman and D. M. Blei, "A tutorial on Bayesian nonparametric models," *J. Math. Psychol.*, vol. 56, no. 1, pp. 1–12, 2012.
[54] J.-T. Chien, "Hierarchical Pitman-Yor-Dirichlet language model," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 8, pp. 1259–1272, Aug. 2015.
[55] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *CoRR*, arXiv:1301.3781, vol. 3, 2013.
[56] Q. V. Le and T. Mikolov, "Distributed representations of sentences and documents," in *Proc. Int. Conf. Mach. Learn.*, 2014, pp. 1188–1196.
[57] A. M. Dai, C. Olah, and Q. V. Le, "Document embedding with paragraph vectors," *CoRR*, arXiv:1507.07998, vol. 1, 2015.
[58] H. Palangi *et al.*, "Deep sentence embedding using long short-term memory networks: Analysis and application to information retrieval," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 4, pp. 694–707, Apr. 2016.

Hongjie Chen (S'16) received the B.Eng. degree from the Northwestern Polytechnical University (NPU), Xi'an, China, in 2013. He is currently working toward the Ph.D. degree in computer science and technology in the Audio, Speech & Language Processing Group (ASLP), Shaanxi Provincial Key Laboratory of Speech & Image Information Processing, School of Computer Science, NPU. In 2014, he visited the Institute for Infocomm Research in Singapore as an intern. His current research interests include automatic speech recognition, natural language processing, spoken term detection, and unsupervised speech modeling.

Lei Xie (M'07–SM'15) received the Ph.D. degree in computer science from Northwestern Polytechnical University (NPU), Xi'an, China, in 2004. He is currently a Professor in the School of Computer Science, NPU. From 2001 to 2002, he was with the Department of Electronics and Information Processing, Vrije Universiteit Brussel (VUB), Brussels, Belgium, as a Visiting Scientist. From 2004 to 2006, he was a Senior Research Associate in the Center for Media Technology (RCMT), School of Creative Media, City University of Hong Kong, Kowloon Tong, Hong Kong. From 2006 to 2007, he was a Postdoctoral Fellow in the Human-Computer Communications Laboratory (HCCL), Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong, Shatin, Hong Kong. His current research interests include speech and language processing, multimedia, and human–computer interaction. He has published more than 120 papers in major journals and proceedings, such as the IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, IEEE TRANSACTIONS ON MULTIMEDIA, INFORMATION SCIENCES, PATTERN RECOGNITION, ACM Multimedia, ACL, INTERSPEECH, and ICASSP.

Cheung-Chi Leung (M'10) received the B.Eng. degree from the Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong, in 1999, and the M.Phil. and Ph.D. degrees from the Chinese University of Hong Kong, Shatin, Hong Kong, in 2001 and 2004, respectively. From 2004 to 2008, he was with the Spoken Language Processing Group, CNRS-LIMSI in France as a Postdoctoral Researcher. In 2008, he joined the Institute for Infocomm Research (I^2R), Singapore, where he is currently a Scientist at the Human Language Technology Department. He is a key member for the development of I^2R's state-of-the-art automatic speech recognition systems. His current research interests include automatic speech recognition, spoken document retrieval, spoken language recognition, and speaker recognition.

Xiaoming Lu (S'12) received the B.Eng. and M.Sc. degrees in computer science from Northwestern Polytechnical University (NPU), Xi'an, China. He visited the Institute for Infocomm Research (I^2R) in Singapore as an intern in 2013. In 2014, he joined Robosay Intelligent Technology Co. Ltd. as an Engineer working on natural language processing. His research interests include machine learning and natural language processing.

Bin Ma (M'00–SM'06) received the B.Sc. degree in computer science from Shandong University, Jinan, China, in 1990, the M.Sc. degree in pattern recognition & artificial intelligence from the Institute of Automation, Chinese Academy of Sciences (IACAS), Beijing, China, in 1993, and the Ph.D. degree in computer engineering from The University of Hong Kong, Pokfulam, Hong Kong, in 2000. He was a Research Assistant from 1993 to 1996 in the National Laboratory of Pattern Recognition, IACAS. In 2000, he joined Lernout & Hauspie Asia Pacific as a Researcher working on speech recognition. From 2001 to 2004, he worked for InfoTalk Corp., Ltd, as a Senior Researcher and a Senior Technical Manager for speech recognition. He joined the Institute for Infocomm Research, Singapore, in 2004 and is now working as a Senior Scientist and the Lab Head of Automatic Speech Recognition. His current research interests include robust speech recognition, speaker & language recognition, spoken document retrieval, natural language processing and machine learning. He has served as a Subject Editor for Speech Communication in 2009–2012, the Technical Program Co-Chair for INTERSPEECH 2014, and is now serving as an Associate Editor for IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING.

Haizhou Li (M'91–SM'01–F'14) received the B.Sc., M.Sc., and Ph.D. degrees in electrical and electronic engineering from South China University of Technology, Guangzhou, China, in 1984, 1987, and 1990, respectively. He is a Professor in the Department of Electrical and Computer Engineering, National University of Singapore, Singapore, and a Conjoint Professor with the University of New South Wales, Sydney, NSW, Australia. He is also a Principal Scientist at Human Language Technology in the Institute for Infocomm Research (I^2R), Singapore. His research interests include automatic speech recognition, speaker and language recognition, and natural language processing. Prior to joining I^2R, he taught in the University of Hong Kong (1988–1990) and South China University of Technology (1990–1994). He was a Visiting Professor at CRIN in France (1994–1995), a Research Manager at the Apple-ISS Research Centre (1996–1998), a Research Director in Lernout & Hauspie Asia Pacific (1999–2001), and the Vice President in InfoTalk Corp. Ltd. (2001–2003). He is currently the Editor-in-Chief of IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING (2015–2017), a Member of the Editorial Board of Computer Speech and Language (2012–2015). He is an elected Member of IEEE Speech and Language Processing Technical Committee (2013–2015), the President of the International Speech Communication Association (2015–2017), and the President of Asia Pacific Signal and Information Processing Association (2015–2016). He was the General Chair of ACL 2012 and INTERSPEECH 2014. He received the National Infocomm Award 2002 and the Presidents Technology Award 2013 in Singapore. He was named one of the two Nokia Visiting Professors in 2009 by the Nokia Foundation.