# BROADCAST NEWS STORY SEGMENTATION USING LATENT TOPICS ON DATA MANIFOLD

*Xiaoming Lu*[1, 2]*, Cheung-Chi Leung*[2]*, Lei Xie*[1]*, Bin Ma*[2]*, Haizhou Li*[2]

[1] Shaanxi Provincial Key Laboratory of Speech and Image Information Processing, School of Computer Science, Northwestern Polytechnical University, Xi'an, China
[2] Institute for Infocomm Research, A*STAR, Singapore

## ABSTRACT

This paper proposes to use Laplacian Probabilistic Latent Semantic Analysis (LapPLSA) for broadcast news story segmentation. The latent topic distributions estimated by LapPLSA are used to replace term frequency vector as the representation of sentences and measure the cohesive strength between the sentences. Subword *n*-gram is used as the basic term unit in the computation. Dynamic Programming is used for story boundary detection. LapPLSA projects the data into a low-dimensional semantic topic representation while preserving the intrinsic local geometric structure of the data. The locality preserving property attempts to make the estimated latent topic distributions more robust to the noise from automatic speech recognition errors. Experiments are conducted on the ASR transcripts of TDT2 Mandarin broadcast news corpus. Our proposed approach is compared with other approaches which use dimensionality reduction technique with the locality preserving property, and two different topic modeling techniques. Experiment results show that our proposed approach provides the highest F1-measure of 0.8228, which significantly outperforms the best previous approaches.

***Index Terms***— story segmentation, dimensionality reduction, topic modeling, laplacian probabilistic latent semantic analysis

## 1. INTRODUCTION

Story segmentation is the task of partitioning a multimedia stream into a number of units each addressing a main topic or a coherent story [1]. Manual segmentation is accurate but labor-intensive, costly and infeasible due to the exponential growth of multimedia data. Therefore, automatic story segmentation approaches are highly in demand.

Taking advantage of the lexical-cohesion based approach which originated from text segmentation [2-4] has been widely studied. In this way, the audio portion of the multimedia stream is firstly passed to an automatic speech recognition (ASR) system and lexical cues are extracted from the ASR transcripts. Lexical cohesion [5] refers to the phenomenon that terms in a coherent story tend to hang together by semantic relations and different stories tend to use different sets of terms. Term repetition is the most common appearance of the lexical cohesion phenomenon. Therefore, the cohesive strength between sentences is usually measured using the cosine similarity between the term frequency vectors of different sentences. Cohesive strength scores are then used to detect story boundaries based on local [6, 7] or global [8] optimization.

The method mentioned above only relies on rigid term repetition, while term association in lexical cohesion is not considered. Moreover, it suffers from the problems of polysemy and synonymy in text. To deal with these problems, a topic technique known as Probabilistic Latent Semantic Analysis (PLSA) [9] has been introduced to the story segmentation task, in which conceptual matching through latent topics is considered in measuring inter-sentence cohesive strength. This method also shows significant improvement compared to using Latent Semantic Analysis (LSA) [10] in conceptual matching [11]. To overcome the overfitting problem of PLSA, Latent Dirichlet Allocation (LDA) [12] has been proposed to assume each latent topic have a Dirichlet prior.

Using a geometrically motivated dimensionality reduction method known as Laplacian Eigenmaps (LE) [13, 22] has recently been proposed to improve the performance of broadcast news story segmentation [14]. LE projects the data (term frequencies in sentences used in [14]) into a low-dimensional representation while preserving the intrinsic local geometric structure of the sentences. The locality preserving property attempts to make the algorithm more robust to the noise from ASR errors. To further improve the segmentation performance, the latent topic distributions estimated using PLSA are used to replace the term frequency vectors in estimating the LE projection [15]. Despite the promising result of this PLSA-LE approach, the parameters in PLSA and LE are estimated independently. We believe that certain information is lost in constructing a Laplacian matrix for estimating the data projection. To better utilize the benefits of topic modeling techniques and LE in data representation, LapPLSA [16] has recently been proposed and it shows better text clustering performance than PLSA and LDA. LapPLSA considers the information from data manifold by constructing a Laplacian matrix, so the model estimation maximizes the joint probability over a set of training data and simultaneously respects the data manifold.

In this paper, we propose to use LapPLSA to estimate the latent topic distribution of each text block for the story segmentation of broadcast news. Note that in the document clustering task in [16], the weight matrix which represents the similarity between document pairs is constructed in an unsupervised manner. The nearest neighbors in terms of word occurrences are used in the matrix construction. However, in our segmentation task, story boundary information is available in the training data, so the natural choice is to make use of this information in the matrix construction. Moreover, as in [14], we incorporate the temporal distances between text block pairs as a penalty factor in the weight matrix.

Our task is performed on ASR transcripts, which are inevitably error-prone and suffer from out-of-vocabulary issues. This induces noises on words and breaks certain lexical cohesion. To deal these problems, as in [14, 15, 17, 20], we attempt to use not only word but also subword bigram [21] as the basic unit in LapPLSA and the measurement of the cohesive strength between text blocks. Dynamic Programming (DP) [8] is used in story boundary detection. Our proposed algorithm is evaluated on TDT2 Mandarin broadcast news corpus. Different approaches involving the use of PLSA, LDA and LE are compared in this work.

## 2. LAPLACIAN EIGENMAPS

Laplacian Eigenmaps (LE) is a geometrically motivated algorithm to project data into a low-dimensional representation while preserving the local neighborhood information of the data. LE is aimed to make the low-dimensional representation of sentences robust to the noise from ASR errors [14, 15].

### 2.1. Construction of weight matrix

Given the ASR transcripts in $N$ units of text blocks, we denote the corresponding latent topic distributions $\mathbf{X}=[\mathbf{x}_1,\mathbf{x}_2,\cdots,\mathbf{x}_N]$ in $\mathbb{R}^K$, where $K$ is the size of latent topics. The details of the construction of the text blocks can be found in section 4.1. Note that in our task, a single news program is divided into a number of text blocks, each of whose starting time and ending time represent story boundary candidates. Moreover, the text block construction for the test data differs from that for the training/development data. However, in [16], each text block in which a latent topic distribution is estimated refers to an independent and well-defined text document.

Let $G$ denote a graph with $N$ nodes which represents the relationship between text block pairs. We put an edge between nodes $i$ and $j$ if $\mathbf{x}_i$ and $\mathbf{x}_j$ come from the same story. We define a weight matrix $\boldsymbol{S}=(s_{ij})_{(i,j=1,2,\cdots,N)}$ of the graph $G$ to model inter-sentence cohesive strength as:

$$s_{ij}=\cos(\mathrm{x}_i,\mathrm{x}_j)\cdot\alpha^{|i-j|}=\frac{\sum_t x_{i,t}x_{j,t}}{\sqrt{\sum_t x_{i,t}^2\sum_t x_{j,t}^2}}\cdot\alpha^{|i-j|} \quad (1)$$

where $\cos(\mathrm{x}_i,\mathrm{x}_j)$ is the cosine similarity between $\mathbf{x}_i$ and $\mathbf{x}_j$. $\alpha^{|i-j|}$ is the penalty of factor of the distance of $|i-j|$. $\alpha$ is a constant, which is set to 1.0 and 0.9 in training and test stages respectively. In the test stage, if the distance between two sentences is much larger than the ordinary length of a story, the cohesive strength will dramatically decrease because of the term $\alpha^{|i-j|}$. $t$ ranges over the latent topics in the dimension, and $x_{i,t}$ is the $t^{\text{th}}$ element of distribution $\mathbf{x}_i$.

### 2.2. Projection of data

Given the weight matrix $\boldsymbol{S}$, we define $\boldsymbol{C}$ as a diagonal matrix whose entries are column (or row, because $\boldsymbol{S}$ is symmetric) sums of $\boldsymbol{S}$. We also define $\boldsymbol{L}=\boldsymbol{C}-\boldsymbol{S}$, which is called Laplacian matrix in spectral graph theory.

$\mathbf{Y}=[\mathbf{y}_1,...,\mathbf{y}_n]$ is the low-dimensional representation of $\mathbf{X}$. This mapping can be represented by:

$$f:\mathbf{x}_i\Rightarrow\mathbf{y}_i \quad (2)$$

A reasonable criterion for obtaining an optimal mapping solution is to minimize the following objective function:

$$\sum_{ij}(\mathbf{y}_i-\mathbf{y}_j)^2 s_{ij} \quad (3)$$

Weight $s_{ij}$ would incur heavy costs if points $\mathbf{x}_i$ and $\mathbf{x}_j$ are mapped too far or too close to each other. Therefore, minimizing the objective function is to ensure that $\mathbf{y}_i$ and $\mathbf{y}_j$ have the same local geometrical relationship as between $\mathbf{x}_i$ and $\mathbf{x}_j$. It turns out that for each representation $\mathbf{y}_i$, the objective function can be transformed as:

$$\sum_{ij}(\mathbf{y}_i-\mathbf{y}_j)^2 s_{ij}=\mathrm{tr}(\boldsymbol{Y}^T\boldsymbol{L}\boldsymbol{Y}) \quad (4)$$

By Rayleigh-Ritz theorem [19], the solution of this function could be provided by the lowest $Q$ eigenvalues for the generalized eigenmaps problem:

$$\mathbf{X}\boldsymbol{L}\mathbf{X}^T\mathbf{y}=\lambda\mathbf{X}\boldsymbol{C}\mathbf{X}^T\mathbf{y} \quad (5)$$

$\mathbf{Y}$ is the solutions which are in the order of their eigenvalues $\lambda_1\le\lambda_2\le\cdots\le\lambda_Q$. $\mathbf{y}_i$ is a $Q$-dimensional ($Q<K$) representation of topic distribution vector $\mathbf{x}_i$.

## 3. TOPIC MODELING

### 3.1. Probabilistic latent semantic analysis

The essence of PLSA is a latent variable model in which each co-occurrence data, i.e., the occurrence of a term $w_m\in W=\{w_1,\cdots,w_M\}$ in a particular document $d_i\in D=\{d_1,\cdots,d_N\}$, is associated with an unobserved topic variable $z_k\in Z=\{z_1,\cdots,z_K\}$, which can be considered as a class label or topic. It is a generative model for word-document co-occurrences.

We obtain a co-occurrence pair $(d_i,w_m)$ on the state of associated latent variable $z_k$. Translating the document generation process into a probability model results in the expression:

$$P(d_i,w_m)=P(d_i)\sum_{k=1}^{K}P(w_m\mid z_k)P(z_k\mid d_i) \quad (6)$$

The conditional probability distributions $P(w_m \mid z_k)$ and $P(z_k \mid d_i)$ can be estimated by maximizing the log-likelihood:

$$\zeta_{PLSA} = \sum_{i=1}^{N}\sum_{m=1}^{M} n(d_i, w_m) \log P(d_i, w_m) \tag{7}$$

where $n(d_i, w_m)$ is the number of occurrences of word $w_m$ in document $d_i$.

The standard Expectation Maximization (EM) alternates two steps [18, 23]: i) an expectation (E) step where posterior probabilities are computed for the latent variables, based on the current estimates of the parameters as:

$$P(z_k \mid d_i, w_m) = \frac{P(w_m \mid z_k) P(z_k \mid d_i)}{\sum_l^K P(w_m \mid z_l) P(z_l \mid d_i)} \tag{8}$$

and ii) a maximization (M) step, where parameters in Eq.(8) are updated as:

$$P(w_m \mid z_k) = \frac{\sum_{i=1}^{N} n(d_i, w_m) P(z_k \mid d_i, w_m)}{\sum_{j=1}^{M}\sum_{i=1}^{N} n(d_i, w_j) P(z_k \mid d_i, w_j)} \tag{9}$$

$$P(z_k \mid d_i) = \frac{\sum_{j=1}^{M} n(d_i, w_j) P(z_k \mid d_i, w_j)}{n(d_i)} \tag{10}$$

with an initial random value. PLSA alternately applies the E-step and M-step until a convergence threshold is met.

### 3.2. Laplacian probabilistic latent semantic analysis

Notice that PLSA fails to discover the local geometrical structure in the sentence. There is no decipherable relation between $P_D = \{P(d_i)_{i=1,\cdots,N}\}$ and the conditional probability distribution $P(z_k \mid d_i)$. This makes the knowledge of $P_D$ unlikely to be useful. To address this issue, LapPLSA makes a specific assumption about the connection between $P_D$ and $P(z_k \mid d_i)$. In LapPLSA, if two documents $d_1, d_2 \in D$ are close in the intrinsic geometry of $P_D$, then the conditional probability distributions $P(z_k \mid d_1)$ and $P(z_k \mid d_2)$ are similar to each other. So we would obtain a document manifold which can be approximated through the graph $G$ (as the one defined in section 2.1) on a scatter of data points. We also define the weight matrix $\boldsymbol{S}$, diagonal matrix $\boldsymbol{C}$ and Laplacian matrix $\boldsymbol{L}$ as in section 2. To make the conditional probability distributions sufficiently smooth, we minimize the following function:

$$\Re_k = \frac{1}{2}\sum_{i,j=1}^{N} (P(z_k \mid d_i) - P(z_k \mid d_j))^2 s_{ij} \tag{11}$$

LapPLSA parameters are estimated by minimizing the following regularized log-likelihood:

$$\zeta_{LapPLSA} = \zeta_{PLSA} - \lambda \sum_{k=1}^{K} \Re_k \tag{12}$$

where $\lambda$ is the regularization parameter.

Similar to PLSA, LapPLSA paremeters are updated iteratively by an E-step and an M-step. The E-step in LapPLSA is the same as that in PLSA. In the M-step, we define:

$$\mathrm{Q}_{LapPLSA}(\psi) = \mathrm{Q}_{PLSA}(\psi) - \lambda \sum_{k=1}^{K} \Re_k \tag{13}$$

where $\mathrm{Q}_{LapPLSA}(\psi)$ and $\mathrm{Q}_{PLSA}(\psi)$ are the expected data log-likelihood for LapPLSA and PLSA respectively.

The computation in the M-step is summarized as follows:

1. compute $P(w_j \mid z_k)^{(t)}$ and $P(z_k \mid d_i)^{(t)}$ as in Eq.(9) and Eq.(10) respectively, where $(t)$ represents the $t$-th iteration in the current M-step;
2. update $P(z_k \mid d_i)^{(t+1)}$ with $P(z_k \mid d_i)^{(t)}$ as follows:

$$P(z_k \mid d_i)^{(t+1)} = (1-\gamma) P(z_k \mid d_i)^{(t)} + \gamma \frac{\sum_{j=1}^{N} s_{ij} P(z_k \mid d_i)^{(t)}}{\sum_{j=1}^{N} s_{ij}} \tag{14}$$

3. repeat step1 and step 2 until $\mathrm{Q}(\psi^{(t+1)}) \geq \mathrm{Q}(\psi^{(t)})$.

## 4. OUR PROPOSED APPROACH FOR STORY SEGMENTATION

### 4.1. Data preprocessing and latent topic estimation

For the training data, ASR transcripts with manually labeled story boundary tags are used. Text streams are broken into block units, each of which is a complete story. In the test data to be segmented, since there is no boundary information available, the text streams are divided into block units using the time labels of pauses in the ASR transcripts. If a pause duration is longer than 1.0 sec, it is considered as a story boundary candidate. This approach is compared with the formation of overlapping fixed-number-of-word pseudo-sentences [14]. No significant difference between the two approaches in story segmentation performance is found in a preliminary test. Note that if the ASR transcripts are at word level, and sub-words are used as the basic units in LapPLSA and the measurement of cohesive strength, word-to-subword conversion is needed.

Given the text blocks of the training data, term frequencies in each text block, a weight matrix $\boldsymbol{S}$, and a Laplacian matrix $\boldsymbol{L}$ are obtained. These are used for LapPLSA parameter estimation as described in section 3.2 with a preset topic number. This estimation process yields $P(w_j \mid z_k)$ as term distribution over a certain latent topic $z_k$. Then folding-in process is used to get the latent topic distribution of the other text blocks from the test data.

### 4.2. Story boundary detection

In story boundary detection, Dynamic Programming (DP) is used to obtain the global optimal solution. DP can more effectively capture smooth story shifts, compared with classical TextTiling method [6]. When using DP for story boundary detection, a target function can be defined as follows:

$$\Im = \sum_{t=1}^{Ns} \Big( \sum_{i,j \in Seg_t} \left\| \mathbf{z}_i - \mathbf{z}_j \right\|^2 \Big) \tag{15}$$

where $\mathrm{z}_i$ and $\mathrm{z}_i$ are the latent topic distributions of text blocks $i$ and $j$ respectively. $\| \mathbf{z}_i - \mathbf{z}_j \|^2$ is the Euclidean distance between the two distributions. $Seg_t$ defines a set of

text blocks that are assigned to a story. $N_S$ is the number of stories. The story boundaries which minimize the target function $\Im$ form the optimal results.

## 5. EXPERIMENTAL SETUP

We conducted experiments on the ASR transcripts of 53-hour TDT2 VOA Mandarin broadcast news corpus as in [15]. The 177 news programs of the corpus were separated into three non-overlapping sets: a training set of 90 programs for parameter estimation in topic models and LE, a development set of 43 programs for empirical tuning and a test set of 44 programs for performance evaluation.

The following five approaches, in which DP is used in story boundary detection, were evaluated in the experiments:

- PLSA-DP: PLSA topic distributions were used to compute sentence cohesive strength.
- LDA-DP: LDA topic distributions were used to compute sentence cohesive strength.
- PLSA-LE-DP: PLSA topic distributions followed by LE projection were used to compute sentence cohesive strength.
- LDA-LE-DP: LDA topic distributions followed by LE projection were used to compute sentence cohesive strength.
- LapPLSA-DP: LapPLSA topic distributions were used to compute sentence cohesive strength.

We evaluate these story segmentation approaches using both word unigram and syllable bigram. The syllable sequences were obtained from the word transcripts using an in-house Mandarin word-to-syllable lexicon. F1-measure is used as the evaluate criterion. We follow the evaluation rule in TDT2: a detected boundary is considered correct if it lies within a 15-second tolerant window on each side of a reference boundary.

The convergence threshold in LapPLSA, PLSA and LDA was set to $1.0\times10e^{-4}$. The number of latent topics in PLSA-DP was set to 64 according to the empirical tuning on word unigram. For a fair comparison with PLSA, the number of latent topics in LDA-DP was set to 64. In PLSA-LE-DP approach, after the number of latent topics was fixed to 64, the dimensionality after LE mapping was set to 32 according to the empirical tuning on word unigram. For a comparison with PLSA-LE-DP, the number of latent topics in LapPLSA-DP was set to 32.

## 6. EXPERIMENTAL RESULTS AND ANALYSIS

Table I provides the story segmentation results on the test set in terms of F1-measure. These results reveal the following observations:

- LapPLSA-DP on syllable bigram performs the best (0.8228). Among all the approaches on word unigram, LapPLSA-DP also performs the best (0.8142). This demonstrates that considering the intrinsic manifold of the data in latent topic estimation is important. LapPLSA-DP achieves a 14% relative F1-measure improvement (from 0.7411 to 0.8228) over PLSA-LE-DP on syllable bigram.
- Applying LE on PLSA/LDA topic distributions performs better than the corresponding approaches without using LE. This implies that the intrinsic local geometrical structure of the data carries important information for obtaining better latent topic distributions, though LE operates on the topic distributions estimated by PLSA/LDA. Note that PLSA/LDA dose not consider the intrinsic local geometrical information of the data in the parameter estimation, so much information is lost in a certain extent.
- In all the five approaches, using syllable bigram performs better than using word unigram. This result is consistent with that in previous works [14,15,17,20]. But this performance improvement by using syllable bigram becomes diminished in LapPLSA-DP.
- LDA and PLSA perform similarly (i.e., PLSA-DP vs. LDA-DP and PLSA-LE-DP and LDA-LE-DP) though LDA is believed to successfully address the overfitting issue in PLSA. PLSA does slightly better in both sets of comparisons.

**Table I** story segmentation results (F1-measure)

| | Word Unigram | Syllable Bigram |
|---|---|---|
| PLSA-DP | 0.6398 | 0.6612 |
| LDA-DP | 0.6317 | 0.6594 |
| PLSA-LE-DP | 0.7102 | 0.7411 |
| LDA-LE-DP | 0.6954 | 0.7334 |
| LapPLSA-DP | 0.8142 | 0.8228 |

## 7. CONCLUSIONS

We purpose to use LapPLSA for broadcast news story segmentation. We compute the cohesive strength between sentences using the latent topic distributions estimated by LapPLSA. Our experiments on the ASR transcripts of TDT2 Mandarin broadcast news corpus show that the LapPLSA-DP approach brings a significant improvement over the previous best PLSA-LE-DP approach. This indicates the importance of the computation of latent topic distributions with respect to the intrinsic manifold of the data simultaneously. Although the use of LapPLSA for topic representation was first proposed for the document clustering task without any training labels associated to the documents, our experiments show that LapPLSA also works promisingly on the story segmentation task with labeled training data.

## 8. ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China (61175018), the Natural Science Basic Research Plan of Shaanxi Province (2011JM8009) and the Fok Ying Tung Education Foundation (131059).

## 9. REFERENCES


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

[2] D. Beeferman, A. Berger, and J. Lafferty, "Statistical model for text segmentation," *Machine Learning*, vol. 34, no. 1-3, pp. 42-60, 1999.

[3] J. Yamron, I. Carp, L. Gillick, and P. Mulbregt, "A hidden Markov model approach to text segmentation and event tracking," in *Proc. of ICASSP*, pp. 333-336, 1999.

[4] F. Y. Y. Choi, "Advances in domain independent linear text segmentation," in *Proc. of NAACL*, pp. 26-33, 2000.

[5] M. Halliday, and R. Hasan, *Cohesion in English*, Longman Group, New York, 1976.

[6] M.A. Hearst, "TextTiling: segmenting text into multi-paragraph subtopic passages," *Computational Linguistics*, vol. 23, no. 1, pp. 33-64, 1997.

[7] N. Stokes, J. Carthy, and A. F. Smeaton, "Select: a lexical cohesion based news story segmentation system," *AI Communication*, vol. 17, pp. 3-12, 2004.

[8] P. Fragkou, V. Petridis, and A. Kehagias, "A dynamic programming algorithm for linear text segmentation," *Journal of Intelligent Information System*, vol. 23, pp. 179-197, 1997.

[9] T. Hofmann, "Probabilistic latent semantic indexing," in *Proc. of SIGIR'99*, pp. 50-57, 1999.

[10] F. Choi, P. W. Hastings, and J. Moore, "Latent semantic analysis for text segmentation," in *Proc. of EMNLP*, pp. 109-117, 2001.

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

[12] D. Blei, A. Ng, and M. Jordan, "Latent Dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993-1022, 2003.

[13] M. Belkin, and P. Niyogi, "Laplacian eigenmaps for dimensionality reduction and data representation," *Neural Computation*, vol. 15, pp. 1383-1396, 2002.

[14] L. Xie, L. Zheng, Z. Liu, and Y. Zhang, "Laplacian eigenmaps for automatic story segmentation of broadcast news," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 20, pp. 264-277, 2012.

[15] M. Lu, L. Zheng, C.-C. Leung, L. Xie, B. Ma, and H. Li, "Broadcast news story segmentation using probabilistic latent semantic analysis and Laplacian eigenmaps," in *Proc. of APSIPA ASC 2011*, pp. 356-360, 2011.

[16] D. Cai, Q. Zhu, J. Han, and C. Zhai, "Modeling hidden topics on document manifold," in *Proc. of the 17th ACM conference on Information and knowledge management*, pp.911-920, 2008.

[17] Y. Yang, and L. Xie, "Subword latent semantic analysis for TextTiling-based automatic story segmentation of Chinese broadcast," in *Proc. of ISCSLP*, pp.358-361, 2008.

[18] W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling, *Numerical Recipes in C: The Art of Scientific computing*, Cambridge University Press, 1992.

[19] H. Lutkepohl, *Handbook of matrices*, Wiley, Chichester, UK, 1997.

[20] L. Xie, Y. Yang, and Z. Liu, "On the effectiveness of subwords for lexical cohesion based story segmentation of Chinese broadcast news," *Information Sciences*, vol. 181, no. 13, pp. 2873-2891, 2011.

[21] L. Xie, and Y. Yang, "Subword lexical chaining for automatic story segmentation in Chinese broadcast news," in *Proc. of PCM*, pp. 248-258, 2008.

[22] M. Belkin, P. Niyogi, and V. Sindhwani, "Manifold regularization: A geometric framework for learning from examples," *Journal of Machine Learning Research*, vol. 7, pp. 2399-2434, 2006.

[23] R. Neal, and G. Hinton, "A view of the EM algorithm that justifies incremental, sparse, and other variants", *Learning in Graphical Models*, Kluwer, 1998.