Broadcast News Story Segmentation Using Probabilistic Latent Semantic Analysis and Laplacian Eigenmaps

Mimi Lu*†, Lilei Zheng*†, Cheung-Chi Leung†, Lei Xie*, Bin Ma† and Haizhou Li†
* Shaanxi Provincial Key Laboratory of Speech and Image Information Processing,
School of Computer Science, Northwestern Polytechnical University, Xi'an
E-mail: {mlu,lzheng}@nwpu-aslp.org, lxie@nwpu.edu.cn
† Institute for Infocomm Research, A*STAR, Singapore
E-mail: {ccleung,mabin,hli}@i2r.a-star.edu.sg

***Abstract*—This paper proposes to integrate probabilistic latent semantic analysis (PLSA) and Laplacian Eigenmaps (LE) for broadcast news story segmentation. PLSA can address synonymy and polysemy problems by exploring underlying semantic relations beneath the actual occurrences of words. LE can provide a data transformation with the advantage of preserving the original temporal structure of sentence cohesive relations. We adopt PLSA statistics to replace term frequency as the representation of sentences and measure their connective strength. LE analysis is then performed on the connective strength matrix so that the sentence relations becomes geometrically evident for discriminating different stories. A dynamic programming (DP) algorithm is used for story boundary identification. Experiments show that the proposed method achieves superior story segmentation performances with the highest F1-measure of 0.7536 on TDT2 Mandarin BN corpus.**

I. Introduction

Story segmentation is the task of dividing a multimedia stream into homogenous segments each addressing a main topic. With the ever increasing of Internet bandwidth and rapid decline of storage cost, multimedia contents such as broadcast news, lecture clips, and meeting records are explosively available on the web. Meanwhile, Internet users require efficient retrieval systems, which can provide the access to their desired components rather than a whole document. Specifically for a broadcast news retrieval task, it is useful to segment self-contained stories from the complete program. Thus automatic story segmentation is highly demanded to avoid the tedious and labor-intensive manual annotation work.

Since a story usually consists of semantically analogous words, the indicator of lexical cohesion is more intuitive than audio or video cues and has been successfully adopted in many classic segmentation methods [1], [2], [3], [4]. *Lexical cohesion* refers to the phenomenon that words in a story hang together by semantic relations and different stories tend to employ different set of words. Therefore, inter-sentence connective strength is measured in text and story boundary detection is performed through local comparison [1], [2] or global optimization [3], [4].

In the lexical cohesion based methods mentioned above, term frequency within a sentence is used to calculate cohesive strength. It is based on the assumption that word occurrences can reflect meaning of text and sentences belonging to the same story tend to deploy the same set of words. This rigid word count comparison takes only word repetition into consideration while word choices may be of randomness in real-world due to the polysemy and synonymy phenomena. A word in different contexts may convey irrelevant meanings related to different stories and the main topic may be expressed by different words throughout the story. Strictly matching sentences depending on the actual appearances of words provides unreliable cues for lexical cohesion in both cases. Hence, the strategies which provide conceptual matching should be considered. Probabilistic Latent Semantic Analysis (PLSA) aims to explore the underlying semantic relations in text and has been proven to provide better performance than standard Latent Semantic Analysis (LSA) [5]. Recently we introduced PLSA to story segmentation task and achieved substantial improvement compared to LSA [6].

Our previous work proposed an effective lexical cohesion based approach using Laplacian Eigenmaps (LE) for story segmentation on broadcast news (BN) LVCSR transcripts [7]. LE is a geometrically motivated algorithm recently proposed for data representation [8]. We carry LE analysis on the sentence connective strength matrix and construct a Euclidean space in which each sentence is mapped to a vector. As a result, the cohesive relations between sentences become geometrically evident in the Euclidean space for discriminating different stories. The LE based approaches significantly outperform several state-of-the-art methods.

In this paper, we adopt PLSA statistics to replace term frequencies as the representation of sentences, and use LE technique to reinforce story boundaries. Further analysis of the LE mapping leads to a straightforward criterion for dynamic programming (DP) to seek the optimal segmentation. Experiment results show that the proposed approach can achieve good performances for BN story segmentation.

Proceedings of the Third APSIPA Annual Summit and Conference, pages 356-360
Xi'an, China, 18-21 October, 2011.

II. PLSA based Sentence Connection

A. Sentence Construction

Since sentence delimiters are not available in BN LVCSR transcripts for implying sentence boundaries, a *sentence* here refers to a fixed number of consecutive terms in the input stream. The starting point of each sentence is a story boundary candidate. The word overlap between sentences is allowed in order to obtain adequate boundary candidates without severely restricting the length of sentence.

B. The PLSA Model

PLSA is a generative model which was first introduced in information retrieval [5] and developed for semantic matching between documents and queries. As for story segmentation, we measure sentence connective strength using PLSA statistics. In PLSA, each co-occurrence observation, i.e., the occurrence of a word $w \in \mathcal{W} = \{w_1, \ldots, w_M\}$ in a particular story $d \in \mathcal{D} = \{d_1, \ldots, d_N\}$, is associated with an unobserved variable $z \in \mathcal{Z} = \{z_1, \ldots, z_T\}$, which can be considered as a class label or topic. Given the assumption that d and w are independently conditioned on the state of the associated latent variable z, a joint probability model of story d and word w can be defined by:

$$P(d,w) = P(d) \sum_{z \in Z} P(w|z)P(z|d) \tag{1}$$

$P(w|z)$ and $P(z|d)$ are two parameters to be learnt in the PLSA model. An iterative Expectation Maximization (EM) algorithm is adopted for the maximum likelihood estimation by maximizing:

$$L = \sum_{d \in D} \sum_{w \in W} f(d,w) \log P(d,w) \tag{2}$$

where $f(d,w)$ denotes the frequency of word w in story d. Starting from random initial values, EM procedure alternates two steps: i) E-step where $P(z|d,w)$, the posterior probabilities of latent variables given the observations are computed based on the current estimates of model parameters and ii) M-step, where Eq.(2) is maximized by re-estimating parameters $P(w|z)$ and $P(z|d)$ with the new expected values $P(z|d,w)$.

After the parameters are learnt using the stories in a training corpus, the estimated $P(w|z)$ are used to compute $P(z|s)$ for a sentence s constructed in Section II-A through a *folding-in* process [5]. The process consists of maximizing the likelihood of s with a partial version of EM algorithm described above: the E-step is identical while in the M-step $P(w|z)$ are kept fixed and only $P(z|s)$ are updated. $P(z|s)$ is lated used to calculate sentence connective strength.

C. Sentence Connective Strength Matrix

For a lexical cohesion based story segmentation method, a lexical similarity indicator is used to represent semantic cohesiveness between sentences. We adopt cosine measure between pairwise sentences to depict their lexical similarity. Term frequency in each sentence is commonly used in the cosine similarity measure [1], [4]. In this paper, PLSA statistics are employed as a substitute for term frequencies to reveal the underlying semantic relations between sentences and the lexical similarity between sentences s_i and s_j is defined as:

$$cos(s_i, s_j) = \frac{\sum_z P(z|s_i)P(z|s_j)}{\sqrt{\sum_z P(z|s_i)^2 \sum_z P(z|s_j)^2}} \tag{3}$$

where $P(z|s_i)$ is the topic specific distribution and z ranges over the latent topic space.

Considering the fact that sentences are less likely to pertain to one story as the distance between sentences extends the regular length of a story, we integrate this distance into the cosine similarity and the sentence connective strength finally becomes:

$$Co(s_i, s_j) = cos(s_i, s_j) \cdot \alpha^{|i-j|} \tag{4}$$

The first part of Eq.(4) is the cosine similarity measure in Eq.(3) and the second part serves as a penalty factor where α is a constant parameter slightly lower than 1.0. If the distance between sentences s_i and s_j is much larger than the ordinary length of a story, $Co(s_i, s_j)$ will dramatically decrease by multiplying $\alpha^{|i-j|}$.

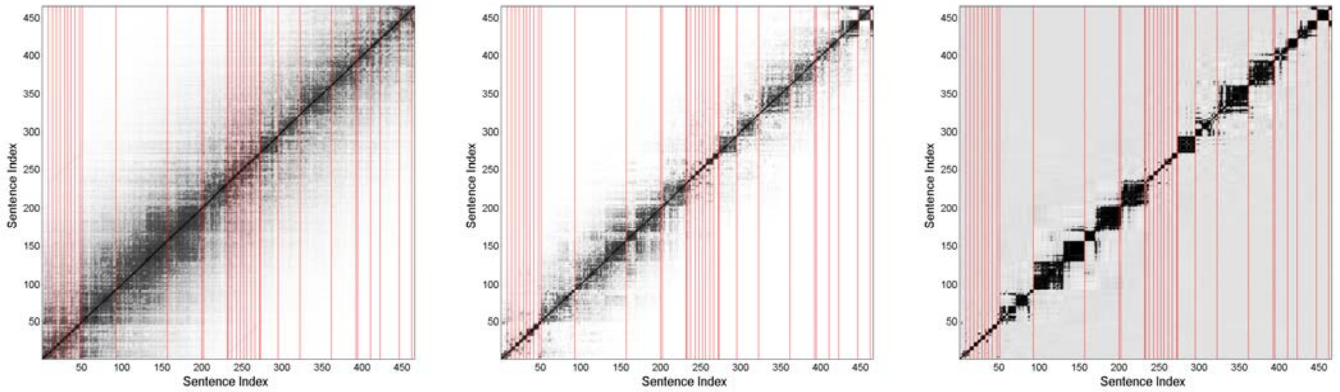
After measuring connective strength for all sentence pairs, the connective strength matrix $\boldsymbol{C}$ is defined as:

$$\boldsymbol{C} = \begin{bmatrix} Co(\mathbf{s}_1,\mathbf{s}_1) & Co(\mathbf{s}_1,\mathbf{s}_2) & \cdots & Co(\mathbf{s}_1,\mathbf{s}_n) \\ Co(\mathbf{s}_2,\mathbf{s}_1) & Co(\mathbf{s}_2,\mathbf{s}_2) & \cdots & Co(\mathbf{s}_2,\mathbf{s}_n) \\ \vdots & \vdots & \ddots & \vdots \\ Co(\mathbf{s}_n,\mathbf{s}_1) & Co(\mathbf{s}_n,\mathbf{s}_2) & \cdots & Co(\mathbf{s}_n,\mathbf{s}_n) \end{bmatrix}, \tag{5}$$

where n is the number of sentences. It is easy to prove that $\boldsymbol{C}$ is symmetric and non-negative. Figure 1 (a) and (b) compare the dotplots using different vector representations (term frequencies versus PLSA statistics) for a broadcast news program. The intensity of pixels corresponds to the value of the entry, i.e. a higher connective strength is represented by a darker pixel. Each dotplot figure contains dark square regions along the diagonal and these regions indicate cohesive story segments with high sentence connective strength. We can see that the introduction of PLSA in measuring connective strength, as shown in Figure (b), makes the segmentation easier in two aspects: (1) the intensity of intra-story area gets higher, which means sentences within a story are more closely connected and (2) noisy dark dots in inter-story areas are significantly reduced, and this makes the story boundaries clear to differentiate.

III. Laplacian Eigenmaps for Story Boundary Identification

In our previous work on lexical cohesion based story segmentation, we introduced a data transformation procedure known as Laplacian Eigenmaps (LE) to project data into a Euclidean space in which the natural clusters in the data are implicitly emphasized. Specifically, the locality preserving characteristic makes the LE algorithm relatively robust to noises in data. Take advantage of this characteristic, the LE approaches can reinforce the story boundary positions more



(a)

(b)

(c)

Fig. 1. Dotplots for a one-hour program in the TDT2 Mandarin corpus: (a) connective strength measured using term frequency; (b) connective strength measured using PLSA statistics; (c) cosine similarities between sentences (i.e., $\mathbf{y}_i$) after LE mapping.

effectively [7]. In this paper, we applied the LE procedure to the vectors of PLSA statistics.

Given the connective strength matrix $\boldsymbol{C} := (c_{ij})_{(i,j=1,\cdots,n)}$, we define the unnormalized graph Laplacian matrix as:

$$\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{C}. \tag{6}$$

where $\boldsymbol{D}$ is the diagonal matrix with $d_i = \sum_{j=1}^{n} c_{ij}$.

Consider the problem of mapping the sentence $\mathbf{s}_i$ to a lower dimensional vector $\mathbf{y}_i$ so that the sentences in the same story stay as close together as possible. Let

$$f : \mathbf{s}_i \mapsto \mathbf{y}_i \tag{7}$$

be such a mapping to the target space. A reasonable criterion for choosing an optimal mapping is to minimize the objective function:

$$\sum_{i,j} \|\mathbf{y}_i - \mathbf{y}_j\|^2 c_{ij} \tag{8}$$

under appropriate constraints. We can see that if two sentences $\mathbf{y}_i$ and $\mathbf{y}_j$ are connected closely, a large value of c_{ij} between the two sentences will prevent them from being mapped far away from each other.

Assume the result of the mapping is an $n \times k$ matrix $\boldsymbol{Y}$, where k is the dimension of the target space and the i-th row of $\boldsymbol{Y}$ is the vector $\mathbf{y}_i$ that $\mathbf{s}_i$ is mapped to. In our work, k is the actual number of stories in an LVCSR transcript and is preset. Using Laplacian matrix $\boldsymbol{L}$, the objective function (8) can be rewritten as:

$$\sum_{i,j} \|\boldsymbol{y}_i - \boldsymbol{y}_j\|^2 c_{ij} = \mathrm{tr}(\boldsymbol{Y}^T \boldsymbol{L} \boldsymbol{Y}). \tag{9}$$

To prevent $\boldsymbol{Y}$ from degenerating to a zero matrix or other matrices with its rank less than k, the constraint below is attached:

$$\boldsymbol{Y}^T \boldsymbol{D} \boldsymbol{Y} = \boldsymbol{I}, \tag{10}$$

where $\boldsymbol{I}$ is an identity matrix.

Altogether, the problem of finding the optimal mapping can be written as below:

$$\begin{aligned} &\underset{\boldsymbol{Y}}{\mathrm{argmin}} \quad \mathrm{tr}(\boldsymbol{Y}^T \boldsymbol{L} \boldsymbol{Y}) \\ &\text{subject to} \quad \boldsymbol{Y}^T \boldsymbol{D} \boldsymbol{Y} = \boldsymbol{I}. \end{aligned} \tag{11}$$

By the Rayleigh-Ritz theorem [9], the solution of this problem can be provided by the eigenvectors corresponding to the smallest k eigenvalues of the generalized eigenvalue problem:

$$\boldsymbol{L}\boldsymbol{v} = \lambda \boldsymbol{D}\boldsymbol{v} \quad \text{or} \qquad \boldsymbol{D}^{-1}\boldsymbol{L} = \lambda \boldsymbol{v}. \tag{12}$$

The $n \times k$ matrix $\boldsymbol{Y}$, which is supposed to contain $\boldsymbol{y}_1, \cdots, \boldsymbol{y}_n$ as its rows, can be formed with the first k eigenvectors $\boldsymbol{v}_1, \cdots, \boldsymbol{v}_k$ as its columns. After the mapping, the relation between the sentences and stories is well revealed.

Dotplot in Figure 1 (c) shows the cosine similarities between sentences after LE mapping. Compared to Figure 1 (b), it is much easier to differentiate the intra-story area and inter-story area after mapping and story boundaries are clearly revealed.

We adopt a dynamic programming (DP) solution for story boundary identification. Specifically, we formalize the process as minimizing:

$$\sum_{t=1}^{N_s} \left(\sum_{i,j \in Seg_t} \|\mathbf{y}_i - \mathbf{y}_j\|^2 \right), \tag{13}$$

where $||\mathbf{y}_i - \mathbf{y}_j\|^2$ is the inter-sentence Euclidean distance in a story segment Seg_t and N_s is the number of stories. Due to the linear constraint of the story segmentation task [4], we can obtain the global minimization of Eq. (13) using DP algorithm in polynomial time [7].

IV. Experiments

A. Experimental Setup

To evaluate the proposed approach, story segmentation is performed on TDT2 Mandarin broadcast news corpus [1]

[1] http://www.ldc.upenn.edu/Projects/TDT2

TABLE I

Story segmentation results (F1-measure) of experimented methods on the TDT2 Mandarin BN corpus

Approach	Word		Subword							
	Unigram		Unigram		Bigram		Trigram		Quadgram	
	Char.	Syl.	Char.	Syl.	Char.	Syl.	Char.	Syl.	Char.	Syl.
TF-LE-DP	0.6200	0.6232	0.6820	0.7011	0.7409	0.7281	0.6963	0.6932	0.6645	0.6693
PLSA-LE-DP	0.7138	0.7440	0.7536	0.7202	0.7202	0.7472	0.6518	0.6693	0.5469	0.5866
PLSA-DP	0.6407	0.6502	0.6836	0.6550	0.6550	0.6661	0.5866	0.6121	0.5405	0.5485

TABLE II

Statistics of the OOV terms, i.e., terms appearing in the development and test sets but not the training set.

	Word		Subword							
	Unigram		Unigram		Bigram		Trigram		Quadgram	
	Char.	Syl.	Char.	Syl.	Char.	Syl.	Char.	Syl.	Char.	Syl.
No. of OOV terms	4128	4159	380	74	50766	40702	127952	125048	270661	264607
No. of tokens	209919	209919	364801	364801	364714	364714	364627	364627	364540	364540
ratio	1.97%	1.98%	0.10%	0.02%	13.92%	11.16%	35.09%	34.29%	74.25%	72.59%

which contains about 53 hours of VOA Mandarin broadcast news audio. Manually annotated story boundaries and LVCSR transcripts are provided. The 177 news programs of the corpus are separated to three non-overlapping sets: a training set of 90 files for PLSA model estimation, a development set of 43 files for empirical parameters tuning and a test set of 44 files for performance evaluation.

We carried out story segmentation experiments with three methods, namely:

- TF-LE-DP, which uses term frequencies to compute sentence connective strength and applies DP after LE mapping;
- PLSA-LE-DP, which uses PLSA statistics to compute sentence connective strength and applies DP after LE mapping;
- PLSA-DP, which uses PLSA statistics to compute sentence connective strength and directly applies DP without LE mapping.

Due to the robustness of subword to deal with certain speech recognition errors and out-of-vocabulary words by partial matching [10], we experimented the segmentation approaches using both word unigram and character/syllable n-gram. The syllable sequences were obtained from the word transcripts using an in-house Mandarin word-to-syllable lexicon. Both character and syllable were chosen as the subword unit in the experiments since they do not always correspond to each other in pairs. For example, the two different words “负荷” (burden) and “附和” (chime in with) have the same syllable sequence “fu4 he4” while the same character “会” pronounces differently in “会议” (meeting, “hui4”) and “会计” (accountant, “kuai4”). The evaluation criterion used is *F1-measure*, i.e., the harmonic mean of *recall* and *precision*. According to the TDT2 standard, a detected boundary is considered correct if it lies within a 15-seconds tolerant window on each side of a reference boundary.

For each method under evaluation, empirical tuning was first performed on the development set to pick up the optimal parameter settings that achieve the highest F1-measure. Then we applied the best-tuned parameters on the test set for segmentation experiments. The number of topics in the PLSA model was preset to 64 according to empirical results on character word. Other parameters include sentence length, sentence overlap shift and α in Eq.(4).

B. Results and Analysis

The story segmentation results on the test set in term of F1-measure are summarized in Table I. We can observe that the introduction of PLSA significantly improves LE based segmentation performance when using word unigram and character/syllable unigram. The highest F1-measure of 0.7536 is obtained on the character level unigram and the highest relative improvement compared to TF-LE-DP, 19.38% (from 0.6232 to 0.7440) is achieved on syllable-word. These results can be explained by the superiority of PLSA to deal with the synonymy and polysemy problems: different words reflecting a similar concept are considered to be matched despite of their actual appearances and thus contribute to the intra-story connective strength; on the other hand, the meaning of a word may vary in different contexts, therefore the relativeness of two sentences may fall when taking the latent semantic into account although they employ several similar word usage.

We also notice the considerably different performances of PLSA when using different word/subword levels of n-gram. PLSA-LE-DP achieves superior results on word unigram and subword unigram/bigram while the performances of subword trigram/quadgram are inferior. This can be attributed by the fact that the most frequently used words in Chinese are one or two characters long and their semantic perspective can be exploited by PLSA. When the subword terms of higher order are deployed, the majority of the units are merely meaningless combinations of subwords rather than meaningful words, which makes PLSA fail to take advantage of latent semantic relations. Moreover, the OOV problem also draws our attention and could be another cause of the performance degradation. The vocabulary constructed by the training data for PLSA model estimation may not match with the data in the development and test sets. This produces incorrect statistics and affects the PLSA utilization. This OOV phenomenon becomes extremely serious when increasing the order of n-gram. Table II shows the statistics of OOV terms, i.e., terms appearing in the development and test sets but not the training set. OOV terms have taken up notable portion of the whole token set on bigram and higher order n-grams. Especially on character quadgram, 74.25% of the tokens are made up by OOV terms. According to this observation, we believe that the OOV problem hinders latent semantic analysis based on the estimated PLSA model and accounts for the drop of PLSA based segmentation performance.

The performance differences between the two PLSA based methods (PLSA-DP and PLSA-LE-DP) demonstrate the effectiveness of LE technique for story segmentation. We can see that with the LE mapping, the story segmentation performance improves on all word and subword levels of n-gram. For instance, PLSA-LE-DP notably outperforms PLSA-DP with a relative gain of 18.08% (from 0.6121 to 0.7472) on syllable bigram.

V. Conclusions

This paper integrates PLSA into our previous work on Laplacian Eigenmaps for broadcast news story segmentation. PLSA statistics are adopted as the representation of sentences and to measure sentence connective strength. Then we perform LE analysis on the connective strength matrix so that the sentence relations are evident in temporal structure for discriminating different stories. Taking advantages of PLSA and LE, the proposed method achieves the highest story segmentation performance of 0.7536 on character unigram. Additionally, we observe that the OOV problem seriously hinders the PLSA utilization, and this leads to an F1-measure degradation on higher order n-grams.

VI. Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (60802085, 61175018), the China MOE Program for New Century Excellent Talents in University (2008) and the Natural Science Basic Research Plan of Shaanxi Province (2011JM8009).

References

[1] M. A. Hearst, “TextTiling: segmenting text into multi-paragraph subtopic passages,” *Comput. Linguist.*, vol. 23, no. 1, pp. 33–64, 1997.

[2] N. Stokes, J. Carthy, and A. F. Smeaton, “Select: a lexical cohesion based news story segmentation system,” *AI Commun.*, vol. 17, pp. 3–12, January 2004.

[3] P. Fragkou, V. Petridis, and A. Kehagias, “A dynamic programming algorithm for linear text segmentation,” *J. Intell. Inf. Syst.*, vol. 23, pp. 179–197, September 2004.

[4] I. Malioutov and R. Barzilay, “Minimum cut model for spoken lecture segmentation,” in *Proc. of ACL*, 2006, pp. 25–32.

[5] T. Hofmann, “Probabilistic latent semantic indexing,” in *Proc. of SIGIR ’99*, 1999, pp. 50–57.

[6] 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.

[7] 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*, 2011.

[8] M. Belkin and P. Niyogi, “Laplacian eigenmaps for dimensionality reduction and data representation,” *Neural Computation*, vol. 15, pp. 1373–1396, 2002.

[9] H. Lutkepohl, *Handbook of Matrices*. Chichester:Wiley, 1997.

[10] 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.