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

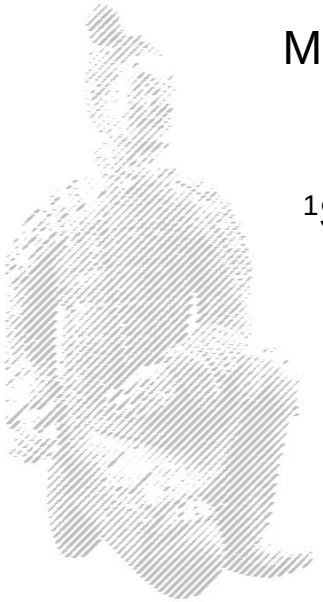
Mimi LU[1,2], Lilei Zheng[1,2], Cheung-Chi LEUNG[2], Lei XIE[1],

Bin MA[2] and Haizhou LI[2]

[1]Shaanxi Provincial Key Lab of Speech and Image Information Processing, Northwestern Polytechnical University, China

[2]Institute for Infocomm Research, A*STAR, Singapore

APSIPA ASC2011, Xi'an, China



Outline

- Introduction
- Motivation
- Methodology
  - Probabilistic Latent Semantic Analysis
  - Laplacian Eigenmap
- Experiments and analysis
- Conclusion



Broadcast news story segmentation

- The task of dividing broadcast news (BN) programs into homogeneous units each addressing a main topic

- A key precursor to various tasks, such as spoken document retrieval and summarization



Motivation

- Lexical cohesion based methods
  - Words in a story hang together by semantic relations
  - Different stories deploy different set of words
- Usually measured by rigid word counts: TF vector
  - Large vocabulary size leads to high dimension and sparse representation
  - Cohesive relations between sentences cannot be reflected clearly due to the sparseness



Motivation

- Dimension reduction (data transformation) required
- From vocabulary size to latent topic number
  - Literal matching is unreliable: polysemy, synonymy
  - Conceptual matching is introduced
- From latent topic number to approximate story number
  - Clustering on topics



Objectives

- PLSA statistics is adopted as the representation of sentences to replace term frequence vector and measure lexical cohesion
- LE analysis is performed to explore geometric relations between sentences and reinforce story boundaries



PLSA model

- Probabilistic latent semantic analysis

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

d: document, w: word, z: topic

- Maximum Likelihood Estimation
  - Maximize log-likelihood of co-occurrence pairs $L = \sum_d \sum_w n(d,w) \log P(d,w)$
  - E-step $P(z|d,w) = \frac{P(w|z)P(z|d)}{\sum_z P(w|z)P(z|d)}$
  - M-step

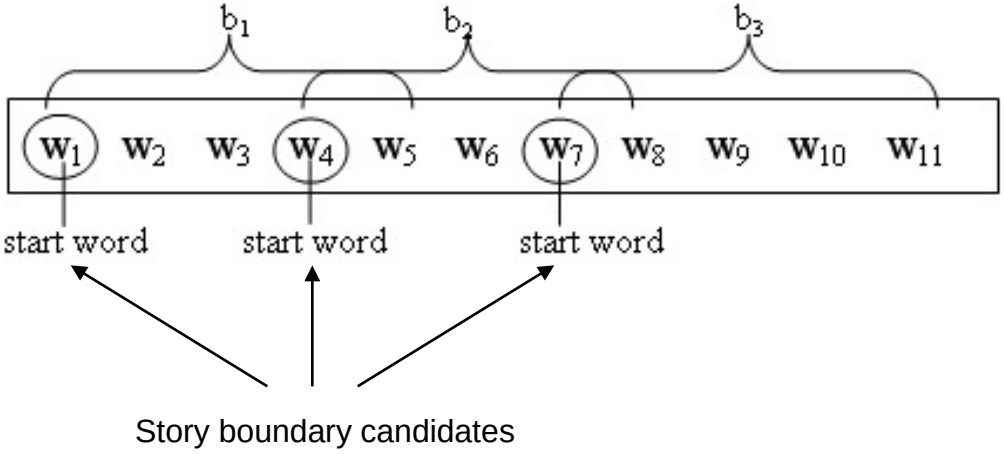
$$P(w|z) = \frac{\sum_d n(d,w)P(z|d,w)}{\sum_w \sum_d n(d,w)P(z|d,w)} \qquad P(z|d) = \frac{\sum_w n(d,w)P(z|d,w)}{\sum_z \sum_w n(d,w)P(z|d,w)}$$

- Folding-in process for unseen test data: keep P(w|z) fixed and only P(z|s) is updated



Sentence Construction

- Sentence delimiters are not available in LVCSR transcripts
- Pseudo-sentence: each text block with a fixed number of consecutive words is formed



PLSA based sentence connection

- Cosine measure is used to depict lexical similarity

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

- Sentence connective strength

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

penalty factor

- Connective strength matrix

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



Laplacian Eigenmaps analysis

- We propose to find a mapping from a sentence s_i to a vector y_i with lower dimension k
- Criterion for choosing the optimal mapping: minimize the objective function

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

- Given the connective strength matrix C, the unnormalized graph Laplacian matrix is defined as:

$$\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{C}$$

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



Laplacian Eigenmaps analysis

- Using Laplacian matrix L, the objective function can be rewritten as

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

- By the Rayleigh-Ritz theorem, the solution of minimizing the above function is

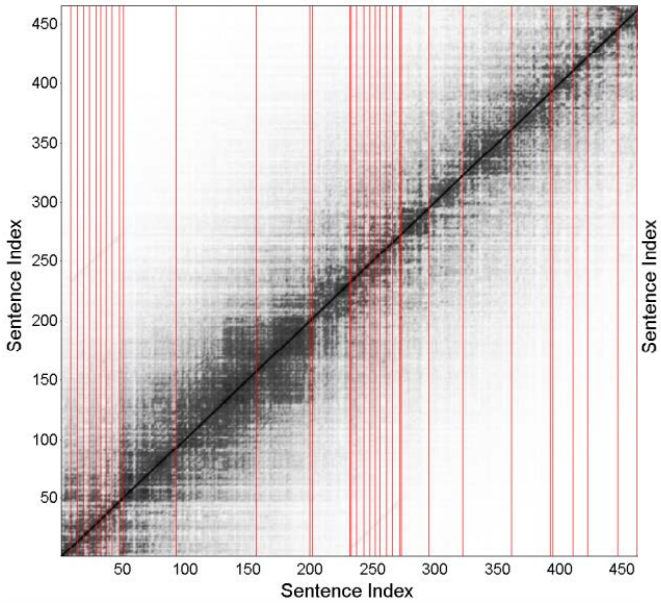
$$\boldsymbol{L}\boldsymbol{v} = \lambda \boldsymbol{D}\boldsymbol{v}$$

  - where v is the eigenvectors corresponding to the smallest k eigenvalues

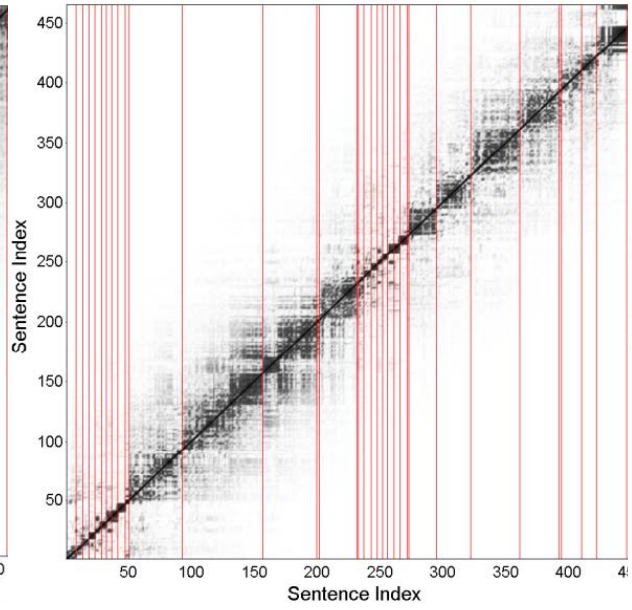


Laplacian Eigenmaps analysis

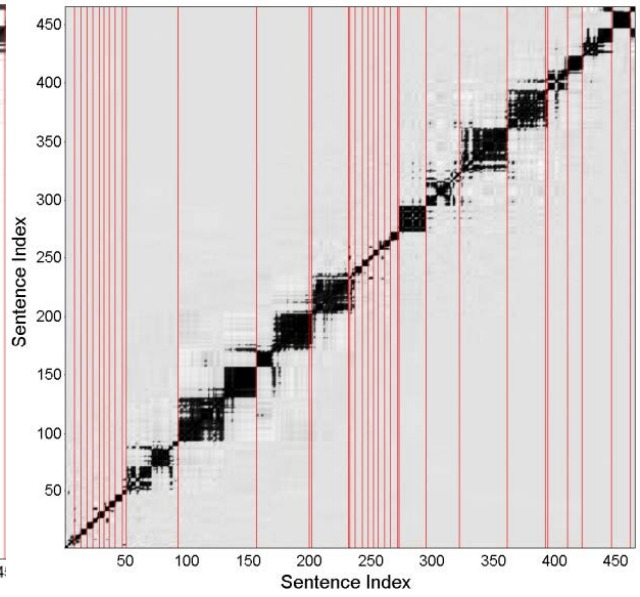
connective strength using TF vector



connective strength using PLSA vector



connective strength using LE mapping vector



Dotplots for a one-hour program in the TDT2 Mandarin corpus



Dynamic Programming solution

- A straightforward DP algorithm is adopted for story boundary identification
- The process is formalized as minimizing:

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

  - where N_s is the number of stories



Experimental setup

- Corpus:
  - LVCSR transcripts of TDT2 VOA Madarine broadcast news
  - Data used (Number of programs):

    training = 90, development = 43, test = 44
- Experiments conducted both on word unigram and character/syllable subwords
- Evaluation criterion: F1-measure

$$F1-measure = \frac{2 * recall * precision}{recall + precision}$$



Experimental results

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%	3.92%	11.16%	35.09%	34.29%	74.25%	72.59%



Conclusions

- We integrate PLSA and LE for BN story segmentation
  - PLSA statistics are employed as the representation of sentences and to measure sentence connective strength
  - LE analysis is conducted on the connective strength matrix to discriminate different stories
- Experimental results suggest:
  - The proposed combination of PLSA and LE can achieve good story segmentation performance
  - The approach performs considerably different on different word/subword level
  - Performance degradation could be also explaned by the OOV problem



Thanks for your attention!