
Mimi LU¹,², Lilei Zheng¹,², Cheung-Chi LEUNG², Lei XIE¹, Bin MA² and Haizhou LI²

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

²Institute 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 frequency 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**

\[ P(d, w) = P(d)P(w \mid d) \]

\[ P(w \mid d) = \sum_{z \in Z} P(w \mid z)P(z \mid d) \]

- **Maximum Likelihood Estimation**
  - Maximize log-likelihood of co-occurrence pairs

\[ L = \sum_{d} \sum_{w} n(d, w) \log P(d, w) \]

- **E-step**

\[ P(z \mid d, w) = \frac{P(w \mid z)P(z \mid d)}{\sum_{z} P(w \mid z)P(z \mid d)} \]

- **M-step**

\[ P(w \mid z) = \frac{\sum_{d} n(d, w)P(z \mid d, w)}{\sum_{w} \sum_{d} n(d, w)P(z \mid d, w)} \]

\[ P(z \mid d) = \frac{\sum_{w} n(d, w)P(z \mid d, w)}{\sum_{z} \sum_{w} n(d, w)P(z \mid d, w)} \]

- **Folding-in process for unseen test data:** keep \( P(w \mid z) \) fixed and only \( P(z \mid s) \) 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

![Diagram showing start words and story boundary candidates]
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

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

- Connective strength matrix

\[
C = \begin{bmatrix}
Co(s_1, s_1) & Co(s_1, s_2) & \cdots & Co(s_1, s_n) \\
Co(s_2, s_1) & Co(s_2, s_2) & \cdots & Co(s_2, s_n) \\
\vdots & \vdots & \ddots & \vdots \\
Co(s_n, s_1) & Co(s_n, s_2) & \cdots & Co(s_n, s_n)
\end{bmatrix}
\]
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

$$\sum_{i,j} \|y_i - y_j\|^2 c_{ij}$$

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

$$L = D - C$$

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

$$
\sum_{i,j} ||y_i - y_j||^2 c_{ij} = \text{tr}(Y^TLY)
$$

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

$$
Lv = \lambda Dv
$$

- 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:

\[ \sum_{t=1}^{N_s} \left( \sum_{i,j \in \text{Seg}_t} ||y_i - y_j||^2 \right) \]

- 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**

\[ F1\text{-measure} = \frac{2 \times \text{recall} \times \text{precision}}{\text{recall} + \text{precision}} \]
Experimental results

### TABLE I

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

<table>
<thead>
<tr>
<th>Approach</th>
<th>Word</th>
<th>Subword</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Unigram</td>
<td>Unigram</td>
</tr>
<tr>
<td>TF-LE-DP</td>
<td>0.6200</td>
<td>0.6232</td>
</tr>
<tr>
<td>PLSA-LE-DP</td>
<td>0.7138</td>
<td>0.7440</td>
</tr>
<tr>
<td>PLSA-DP</td>
<td>0.6407</td>
<td>0.6502</td>
</tr>
</tbody>
</table>

### TABLE II

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

<table>
<thead>
<tr>
<th></th>
<th>Word</th>
<th>Subword</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Unigram</td>
<td>Unigram</td>
</tr>
<tr>
<td>No. of OOV terms</td>
<td>4128</td>
<td>4159</td>
</tr>
<tr>
<td>No. of tokens</td>
<td>209919</td>
<td>209919</td>
</tr>
<tr>
<td>ratio</td>
<td>1.97%</td>
<td>1.98%</td>
</tr>
</tbody>
</table>
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 explained by the OOV problem
Thanks for your attention!