Eliminating Redundancy by Spectral Relaxation for Multi-Document Summarization

F. Fukumoto, A. Sakai and Y. Suzuki
Univ. of Yamanashi
Graph-based multi-document summarization

- LexPagerank (Erkan and Radev, 2004)
- PageRank and HITS (Mihalcea and Tarau, 2005)

  - Constructing graph consisting nodes and links
  - Applying graph-based ranking algorithm
  - Chose the sentences with large rank score into the summary

All the sentences are ranked based on a sentence as unit of information.
Semantically related two sentences with “high recommendation” are ranked with high score, and thus are regarded as a summary sentence.
The resulting summary still contains overlapping info.
Cluster-based Multi-document Summarization

- ClusterCMRW model (X.Wan et al)
- Classifying documents into theme clusters by using k-means
- Constructing a graph to reflect the relationships between sentences and clusters by using MRW model
- Spectral Clustering (Weiss et al)
- A transformation of the original sentences into a set of orthogonal eigenvectors.
Sentence extraction by ClusterCMRW

1. Weight between two sentences, conditioned on the two clusters containing the two sentences.

\[
f(i \rightarrow j \mid clus(s_i), clus(s_j)) = 
\begin{align*}
&f(i \rightarrow j) \cdot \{ \lambda \cdot \pi(clus(s_i)) \cdot \omega(s_i, clus(s_i)) \\
&+ (1 - \lambda) \cdot \pi(clus(s_j)) \cdot \omega(s_j, clus(s_j)) \}
\end{align*}
\]

\[
\pi(clus(s_i)) = \text{sim}_{\text{cosine}}(clus(s_i), D)
\]

\[
\omega(s_i, clus(s_i)) = \text{sim}_{\text{cosine}}(s_i, clus(s_i))
\]

\(\pi(clus(s_i))\): The importance of the cluster \(clus(s_i)\) in the document set \(D\)

\(\omega(s_i, clus(s_i))\): The correlation between the sentence \(s_i\) and its cluster \(clus(s_i)\)
Sentence extraction by ClusterCMRW

2. The transition probability from $s_i$ to $s_j$

$$p(i \rightarrow j \mid \text{clus}(s_i), \text{clus}(s_j)) = \frac{\sum_{k=1}^{V} f(i \rightarrow k \mid \text{clus}(s_i), \text{clus}(s_j))}{f(i \rightarrow j \mid \text{clus}(s_i), \text{clus}(s_j))}$$

3. The final transition matrix $\tilde{M}_{ij}$

$$\tilde{M}_{ij} = p(i \rightarrow j \mid \text{clus}(s_i), \text{clus}(s_j))$$

$$\tilde{\lambda}^* = \mu \tilde{M}^* + \frac{1 - \mu}{|V|} \bar{e} \bar{e}^T$$
ClusterCMRW model

Spectral Clustering

$clus(s_j)$  $clus(s_i)$
Sentence Classification by Spectral Clustering

1. Form a distance matrix $D$

Multi-doc.

\[
\begin{pmatrix}
1 & 0 & 1 & 1 \\
1 & 0 & 1 & 0 \\
1 & 1 & 1 & 0 \\
\end{pmatrix}
\]

\[D = \begin{bmatrix}
0 & 1 & 1.4 \\
1 & 0 & 1 \\
1.4 & 1 & 0 \\
\end{bmatrix}
\]

- Euclid distance
- Word freq.

2. Feature space and sentence classification

- $D$ is transformed to an affinity matrix $A_{ij}$

\[
A_{ij} = \begin{cases}
\exp\left(-\frac{D_{ij}^2}{2\sigma^2}\right) & \text{if } i \neq j \\
0 & \text{if } i = j
\end{cases}
\]

- Create a diagonal matrix $B$

\[
B_{ii} = \sum_{j=1}^{n} A_{ij}
\]

- Create $L$

\[
L = B^{-1/2} A B^{-1/2}
\]

- Each item has a vector of $l$ coordinates in the transformed space.
- These vectors are normalized to unit length, and K-means is applied to $S$ in $l$-dimensional space.

Multi-document summarization by ClusterCMRW
# Experiments

## 1. Data

<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td># of topics</td>
<td>30</td>
</tr>
<tr>
<td># of sentences / doc</td>
<td>30 to 350 sentences</td>
</tr>
<tr>
<td>Ext. of sentences</td>
<td>NTCIR-3 SUMM FBFREE Long and short according to the character length</td>
</tr>
<tr>
<td># of clusters</td>
<td>The square root of the number of sentences</td>
</tr>
</tbody>
</table>
Experiments

2. Two evaluation measures:

- Cosine similarity between the generated summary by the system and the human generated summary
- ROUGE score used in DUC

\[
\text{ROUGE} = \frac{\sum_{S \in \{R, f\text{Sum}\}} \sum_{\text{word} \in S} \text{Count}_{\text{match}}(\text{word})}{\sum_{S \in \{R, f\text{Sum}\}} \sum_{\text{word} \in S} \text{Count}(\text{word})}
\]
Parameter estimation used in the spectral clustering

- 10 topics to estimate two parameters $\sigma$ and $l$ in the $l$-dimensional space
  - $\sigma$ is searched in steps of 0.01 from 1.0 to 5.0
  - $l$ is searched in steps 10% from 0 to 80% against the total number of words in the training data

- The size that optimized the average F-score of 10 topics was chosen
  - $\sigma$ is set to 4.5
  - $l$ is set to 80%
### Summarization Results

<table>
<thead>
<tr>
<th></th>
<th># of doc</th>
<th># of sent</th>
<th># of sum</th>
<th>cos</th>
<th>ROUGE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>MRW</td>
<td>K-means</td>
</tr>
<tr>
<td>short</td>
<td>7.5</td>
<td>83.0</td>
<td>11.9</td>
<td>0.431</td>
<td>0.575</td>
</tr>
<tr>
<td>long</td>
<td></td>
<td></td>
<td>20.4</td>
<td>0.371</td>
<td>0.408</td>
</tr>
</tbody>
</table>

- Sp outperformed the baselines, MRW and k-means, regardless of the types of summary, and evaluation measures.
- Short was better than long. The rank score of correct sentences within the candidate sentences obtained by the MRW model works well.
## Sentence Similarities within a summary

<table>
<thead>
<tr>
<th></th>
<th># of doc</th>
<th># of sent</th>
<th># of sum</th>
<th>Similarity within a summary sentences</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Human</td>
</tr>
<tr>
<td>short</td>
<td>7.5</td>
<td>83.0</td>
<td>11.9</td>
<td>0.129</td>
</tr>
<tr>
<td>long</td>
<td>20.4</td>
<td></td>
<td></td>
<td>0.173</td>
</tr>
</tbody>
</table>
SP is more robust than k-means and simple MRW model even for a large number of input sentences
Sp outperformed the results obtained by directly applying MRW.

The results by k-means was worse than the results of MRW when the ratio of the # of cluster $k$ against the # of sentences as an input was larger than 80%. For a large number of topics, k-means is not effective.
Conclusion

- A method to detect salient sentences from documents that discuss the same event
- 10.6% improvement over a baseline MRW (cosine), and 2.9% (ROUGE score)

- Applying the method to the DUC evaluation data
- Extending the method to classify sentences into more than one clusters by using soft-clustering techniques