Objective: Single-document Extractive Summarization for Wikipedia Articles

Pseudo-optimal Summary

- Finding optimal summary that is maximally similar to the given summary intractable.
- Approximated by finding top sentence from the article that was maximally similar to each sentence in the given summary.

Pipeline

Schematic diagram of the pipeline

Filtering Summarizable Articles

- Removed articles with very small original summary or very small total article length.
- Removed articles where the ratio of the original summary length and the total article length was not close to 0.1.
- Removed articles where pseudo-optimal extractive summary and random summary were comparable in performance.

Graph-based approach

- Nodes correspond to sentences.
- Edge weights are average distance (cosine / Euclidean) between top k pairs of words using word embedding representation.

Machine Learning based approach

A2. Sample summary: ML-based model

As a multi ethnic and multi religious nation India has had some sectarian violence and insurgencies in various parts of the country but has stayed together as a vibrant democracy. It has unresolved border disputes with China which escalated into the brief Sino Indian War in 1962 and with Pakistan which resulted in wars in 1947, 1965, and 1971 and in 1999 in Kargil at the northern fringe of Jammu and Kashmir State.

Results

<table>
<thead>
<tr>
<th></th>
<th>Random summary</th>
<th>Optimal summary</th>
<th>Model summary Graph</th>
<th>Model summary ML</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>0.0886</td>
<td>0.3254</td>
<td>0.1587</td>
<td>0.2972</td>
</tr>
<tr>
<td>Median</td>
<td>0.0879</td>
<td>0.3245</td>
<td>0.1507</td>
<td>0.2902</td>
</tr>
</tbody>
</table>

References