Linking Today's Wikipedia and News from the Past

Arunav Mishra
amishra@mpi-inf.mpg.de

Linking Problem

Link excerpts from Wikipedia articles, coined Wiki-excerpts, summarizing an entire event or a specific story of a multi-faceted event to past news articles providing contemporary accounts

Retrieval Tasks

Wiki2News “How did it happen?”
Given a Wiki-excerpt, retrieve past news articles that provide detailed accounts

Input: A Wiki-excerpt
Output: Ranked list of past news articles

News2Wiki “How is it remembered?”
Given a set of past news articles, retrieve Wiki-excerpts that could be linked to event

Input: Set of news articles
Output: Ranked list of Wiki-excerpts

Challenges

Wiki2News
• Leveraging of spatio-temporal expressions
• Leveraging named entity mentions
• Bridging vocabulary gap

News2Wiki
• Reduction of verbosity
• Estimation of event span
• Addressing language change in historic news articles

Simplified Wiki2News

Link Wikipedia events to past news articles providing contemporary accounts with fine-grained details.

2-Stage Cascade Model

Use pseudo-relevance feedback to improve temporal query model and rerank documents based on their divergence

Baselines

Publication Dates (LM+P): \( P(q \mid d) = P(q_{text} \mid d_{text}) \cdot P(q_{time} \mid t) \)
- First factor is estimated with query-likelihood approach
- Second factor is estimated as \( P(q_{time} \mid t) = \frac{1}{1 + e^{-r(q_{time} - t)}} \)

Temporal expressions (LM+T): \( P(q \mid d) = P(q_{text} \mid d_{text}) \cdot P(q_{time} \mid d_{time}) \)
- First factor is estimated with query-likelihood approach
- Second factor is estimated as \( P(q_{time} \mid d_{time}) = \frac{1}{|d_{time}|} \sum_{b \in |d_{time}|} \frac{1(q_{time} \in [b, c])}{c - b + 1} \)

Publication Dates + Temporal Expressions (LM+PT): \( P(q \mid d) = P(q_{text} \mid d_{text}) \cdot P(q_{time} \mid t) \cdot P(q_{time} \mid d_{time}) \)

Results

Benchmark: 50 random Wikipedia events from years 1987 to 2007
Dataset: The New York Times Corpus, with 2 million documents

<table>
<thead>
<tr>
<th>Benchmark</th>
<th>LMAP</th>
<th>LMAP+P</th>
<th>LMAP+T</th>
<th>LMAP+TP</th>
<th>CM</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAP</td>
<td>0.35</td>
<td>0.43</td>
<td>0.40</td>
<td>0.42</td>
<td>0.45</td>
</tr>
<tr>
<td>P@5</td>
<td>0.55</td>
<td>0.63</td>
<td>0.61</td>
<td>0.61</td>
<td>0.66</td>
</tr>
<tr>
<td>P@10</td>
<td>0.48</td>
<td>0.57</td>
<td>0.54</td>
<td>0.55</td>
<td>0.58</td>
</tr>
<tr>
<td>NDCG@5</td>
<td>0.53</td>
<td>0.59</td>
<td>0.58</td>
<td>0.60</td>
<td>0.60</td>
</tr>
<tr>
<td>NDCG@10</td>
<td>0.54</td>
<td>0.62</td>
<td>0.60</td>
<td>0.62</td>
<td>0.63</td>
</tr>
</tbody>
</table>

The Cascade Model (CM) consistently outperforms the baselines across all effectiveness measures, proving to be the most effective approach.

References