ExB Text Summarizer

Stefan Thomas, Christian Beutenmüller, Xose de la Puente, Robert Remus & Stefan Bordag
ExB Research & Development GmbH
Outline

• Introduction
• ExB Summarizer
  – Preprocessing
  – (Fair)TextRank
• Official results
  – Single document summarization
  – Multi document summarization
Approaches to summarization

• Sentence extractive methods
  - State of the art
  - Limited results

• Abstractive summarization
  - The “human way“ of summarization
  - Additional difficulty: Producing text
Overview of ExB summarizer

• Goals:
  – Scalability
  – Language independence

• Key ideas:
  – Main parts unsupervised
  – TextRank on a similarity graph of sentences
Preprocessing steps

• Rule-based Tokenization
• Stop-word removal
• Stemming
• Sentence boundary detection
• Temporal expression detection for multi document summarization
TextRank

- Invented by Rada Mihalcea
- Origin: PageRank algorithm [Page & Brin]
- Graph-based ranking algorithm
  - Text is represented as nodes and edges
  - Nodes are ranked according to their importance
FairTextRank

• Sentence similarity graph
  - Sentences = nodes
  - Similarity between sentences = weighted edges (between 0 and 1)
  - Bag-of-words model with Jaccard index

• Iterative application of TextRank
  - Helps covering different topics in the produced summary

• Postprocessing
<table>
<thead>
<tr>
<th>System</th>
<th>#langs</th>
<th>Rank R-1</th>
<th>Rank R-2</th>
<th>Rank R-3</th>
<th>Rank R-4</th>
<th>Rank R-4SU</th>
</tr>
</thead>
<tbody>
<tr>
<td>BGU-SCE-M</td>
<td>3</td>
<td>2.0</td>
<td>3.3</td>
<td>3.7</td>
<td>4.3</td>
<td>3.0</td>
</tr>
<tr>
<td>BGU-SCE-P</td>
<td>3</td>
<td>5.0</td>
<td>4.7</td>
<td>5.0</td>
<td>4.3</td>
<td>4.3</td>
</tr>
<tr>
<td>CCS</td>
<td>38</td>
<td>2.1</td>
<td>2.1</td>
<td>2.2</td>
<td>2.3</td>
<td>2.5</td>
</tr>
<tr>
<td>ExB</td>
<td>38</td>
<td>3.2</td>
<td>3.3</td>
<td>3.7</td>
<td>3.8</td>
<td>2.8</td>
</tr>
<tr>
<td>LCS-IESI</td>
<td>38</td>
<td>4.1</td>
<td>4.1</td>
<td>4.0</td>
<td>4.0</td>
<td>4.1</td>
</tr>
<tr>
<td>NTNU</td>
<td>2</td>
<td>5.5</td>
<td>6.0</td>
<td>6.0</td>
<td>7.0</td>
<td>5.0</td>
</tr>
<tr>
<td>UA-DLSI</td>
<td>3</td>
<td>6.0</td>
<td>5.0</td>
<td>4.7</td>
<td>5.0</td>
<td>6.0</td>
</tr>
<tr>
<td>Lead</td>
<td>38</td>
<td>5.1</td>
<td>5.0</td>
<td>4.6</td>
<td>4.3</td>
<td>5.0</td>
</tr>
<tr>
<td>Oracles</td>
<td>38</td>
<td>1.1</td>
<td>1.2</td>
<td>1.2</td>
<td>1.2</td>
<td>1.2</td>
</tr>
</tbody>
</table>
# Results (MMS)

<table>
<thead>
<tr>
<th>Language</th>
<th>AutoSummENG</th>
<th>MeMoG</th>
<th>NPowER</th>
<th>Rank/Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arabic</td>
<td>0.135</td>
<td>0.164</td>
<td>1.717</td>
<td>7/9</td>
</tr>
<tr>
<td>Chinese</td>
<td>0.118</td>
<td>0.141</td>
<td>1.654</td>
<td>1/5</td>
</tr>
<tr>
<td>Czech</td>
<td>0.188</td>
<td>0.200</td>
<td>1.874</td>
<td>4/7</td>
</tr>
<tr>
<td>English</td>
<td>0.167</td>
<td>0.191</td>
<td>1.817</td>
<td>6/10</td>
</tr>
<tr>
<td>French</td>
<td>0.200</td>
<td>0.195</td>
<td>1.892</td>
<td>5/8</td>
</tr>
<tr>
<td>Greek</td>
<td>0.147</td>
<td>0.170</td>
<td>1.750</td>
<td>5/8</td>
</tr>
<tr>
<td>Hebrew</td>
<td>0.115</td>
<td>0.147</td>
<td>1.655</td>
<td>8/9</td>
</tr>
<tr>
<td>Hindi</td>
<td>0.123</td>
<td>0.139</td>
<td>1.662</td>
<td>3/7</td>
</tr>
<tr>
<td>Romanian</td>
<td>0.168</td>
<td>0.183</td>
<td>1.809</td>
<td>4/6</td>
</tr>
<tr>
<td>Spanish</td>
<td>0.193</td>
<td>0.202</td>
<td>1.886</td>
<td>3/6</td>
</tr>
</tbody>
</table>
Summary

- TextRank based approach
- Multitude of preprocessing steps
- Participated in all possible languages
- Competitive results
Negative findings

• No increase in performance via:
  – Semantic Text Similarity instead of bag-of-words/Jaccard index
  – Word2vec word embeddings
  – Named entities

• ROUGE measure is inappropriate
Example of FairTextRank
Example of FairTextRank
Example of FairTextRank