On the robustness of event detection evaluation: a case study

Matthias Feys, Thomas Demeester, Blaz Fortuna, Johannes Deleu, Chris Develder

06/12/2014
Overview

- Event Detection
- Evaluation
- Robustness Issues
- Annotation Coverage
- Conclusion
Event Detection
Document level event detection
Document level event detection

FIFA World Cup 2014

European Elections 2014

Disappearance Flight 370
Vector Space Model

Bag Of Words

‘this text contains the word text twice’

features ←

\[
\begin{array}{c|c|c|c|c|c|c|c}
\hline
\textbf{term} & \text{this} & \text{text} & \text{contains} & \text{the} & \text{twice} & \text{word} \\
\hline
\text{term frequencies (TF)} & 1 & 2 & 1 & 1 & 1 & 1 \\
\hline
\end{array}
\]

→ term frequencies (TF)
Evaluation
(Mean) Average Precision

- Golden Truth
- Area under precision/recall curve
(Mean) Graded Average Precision

- Generalized (M)AP for the case of multi-graded relevance

Datasets

- Chinese dataset:
  - Binary relevance judgements
  - Moderate (>100 and <300 articles) and large events (>300 articles)

- Flemish dataset:
  - Graded relevance judgements:
    - Strong (S)
    - Weak (W)
    - Distant (D)
  - ± Moderate+large events (>90 articles)
Issues
## Results

<table>
<thead>
<tr>
<th></th>
<th>MAP ± stdev</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Chinese</td>
<td>Flemish</td>
<td></td>
</tr>
<tr>
<td></td>
<td>moderate</td>
<td>large</td>
<td>all</td>
</tr>
<tr>
<td>tf-idf</td>
<td>0.50±0.25</td>
<td>0.35±0.24</td>
<td>0.24±0.22</td>
</tr>
<tr>
<td>BurstVSM</td>
<td>0.54±0.21</td>
<td>0.46±0.20</td>
<td>0.22±0.20</td>
</tr>
<tr>
<td>( B)-BurstVSM</td>
<td>0.59±0.18*</td>
<td>\textbf{0.53±0.21}*</td>
<td>0.24±0.20</td>
</tr>
<tr>
<td>( B)-tf-idf</td>
<td>\textbf{0.60±0.21}*</td>
<td>0.53±0.22*</td>
<td>\textbf{0.29±0.20}</td>
</tr>
</tbody>
</table>

### Issues:
- Large standard deviations
- Differences between datasets
Possible causes

- Ambiguity of event definition
- Subjectivity of annotations
Ambiguity of relevance definition

- Relevance definition can impact systems ranking!
Ambiguity of event type

- Homogeneous vs. heterogeneous:
  - E.g.: “Kadhafi’s Death” vs. “Occupy Wall Street protest”

<table>
<thead>
<tr>
<th></th>
<th>homogeneous</th>
<th>heterogeneous</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>MAP</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>tf-idf</td>
<td>0.30±0.22</td>
<td>0.19±0.24</td>
</tr>
<tr>
<td>B BurstVSM</td>
<td>0.15±0.13</td>
<td>0.34±0.23</td>
</tr>
<tr>
<td>BurstVSM</td>
<td>0.12±0.13</td>
<td>0.33±0.21</td>
</tr>
<tr>
<td>B tf-idf</td>
<td>0.26±0.21</td>
<td>0.33±0.22</td>
</tr>
</tbody>
</table>
Subjectivity of annotations

- Correlation between ranking based on annotations of U1 vs. U2 for each event
- Annotator rank correlation
  - AP: 0.53
  - Graded AP: 0.72

However!

- Ranking for both MAP and GAP is 1
Annotation coverage
Annotation coverage

- Do we need to annotate all articles for each event?
- Impact on mean rank correlation:
  - Correlation between:
    - mean ranking
    - ranking for specific event
Impact of annotation coverage
Conclusion
Conclusion

- Event detection evaluation is an ambiguous problem
- Graded relevance levels help:
  - Assess impact of relevance definition
  - Reduce impact of user disagreement
- User disagreement has little impact on system ranking
- Not needed to annotate all articles (sample!)
...questions...