Entity Summarization of News Articles

Roi Blanco, Gianluca Demartini, Malik Muhammad Saad Missen, Hugo Zaragoza

Internship at Yahoo!Lab Barcelona
September – November 2009
Motivation

• Going beyond document retrieval
• Finding entities relevant to a query in a document collection (e.g., Wikipedia)

• In collections of documents over time
  – Decide about relevance at document level (Entity Summarization)
  – Analyse and exploit relevance evolution
Scenario

• An event
  – Charles Schulz dies
• Get Relevant Docs
• Entities
  – Peanuts, his wife, media companies, hometown, other cartoonists, ...
• Timeline of relevant news:
  – 10/1999-09/2000:
    • 11/99 cancer diagnosed
    • 12/99 he retires
    • 02/00 he dies
    • 03/00 peanuts future discussed
    • ... Honors, museums, statues, airports, ...
Tasks

• Entity Ranking (ER)
  – Find the set of entities $e_i$ that best describe the relevant documents $d_i$
  – Yahoo! Correlator

• Entity Summarization (ES)
  – Find the set of entities $e_i$ that best describe document $d$ wrt a query $q$
  – Subtask: Find $e_i$ for $d$ wrt a query $q$ given history $d_i < d$
Tasks

- Entity Profiling (EP)
  - Construct temporal development of entity relevance
Outline

- Dataset
- Data analysis
- Entity Summarization
- Entity Profiles
- Conclusions
Dataset

• TREC Novelty Track 2004
  – Sentence retrieval
  – 25 event topics
  – 779 relevant news

• Entity annotations (7481 entities)
  – Persons (26%), Locations (10%), Organizations (57%), Products (7%)

• Relevance judgements
  – Of each entity wrt to topic in this current news
  – 21,213 judgements on 3 levels
  – Cohen’s Kappa 0.59
Data Analysis

• How useful is to find relevant sentences?
  – $P(e \text{ is Rel})$ 0.411 [0.404-0.417]
  – $P(e \text{ is NotRel})$ 0.168 [0.163-0.173]
  – $P(e \text{ is Rel } | s \text{ is Rel})$ 0.547 [0.534-0.559]
  – Sentences:
    • 21727 total 1.46 entity occurrences
    • 5122 relevant 1.88 entity occurrences
    • 2122 novel 1.92 entity occurrences

• How useful is to find novel sentences?
  – $P(e \text{ is Rel } | s \text{ is Nov})$ 0.510 [0.491-0.531]
Data Analysis

• How useful is looking at the past?
  – \( P(e|d_1) \) 0.893 [0.881-0.905]
  – \( P(e|d_{-1}) \) 0.701 [0.677-0.726]

• Is useful to consider sentence co-occurrence?

<table>
<thead>
<tr>
<th></th>
<th>Relevant</th>
<th>Related</th>
<th>NotRelevant</th>
<th>NotAnEntity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Relevant</td>
<td>0.24</td>
<td>0.08</td>
<td>0.03</td>
<td>0.07</td>
</tr>
<tr>
<td>Related</td>
<td>0.07</td>
<td>0.03</td>
<td>0.03</td>
<td>0.03</td>
</tr>
<tr>
<td>NotRelevant</td>
<td>0.07</td>
<td>0.07</td>
<td>0.05</td>
<td></td>
</tr>
<tr>
<td>NotAnEntity</td>
<td>0.07</td>
<td></td>
<td></td>
<td>0.04</td>
</tr>
</tbody>
</table>
Outline

• Dataset
• Data analysis
• **Entity Summarization**
  – Local Features
  – History Features
• Entity Profiles
• Conclusions
Entity Summarization

- Evaluation
  - P3, P5, AvgPrec
  - Ties aware measures [McSherry and Najork, ECIR08]

- Paired t-test
  - ** p<<0.01
  - * p<0.05

- Related considered NonRelevant
Local Features

- Looking at the document
  - Occurrences of $e$: $F(e,d)$
  - Occurrences of $e$ as subject: $F_{subj}(e,d)$

![Graphs showing occurrences and subject occurrences of $e$ with different functions and probabilities.](image-url)
Local Features

• Look at the position of e in the document
  – Length of the first sentence where e appears
  – Position of the first sentence where e appears

![Graph showing P(e) as a function of First Sentence Position and First Sentence Length]
## Local Features

<table>
<thead>
<tr>
<th>Feature</th>
<th>P3</th>
<th>P5</th>
<th>MAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>F(e,d)</td>
<td>.65</td>
<td>.56</td>
<td>.60</td>
</tr>
<tr>
<td>FirstSenLen</td>
<td>.37</td>
<td>.36</td>
<td>.45</td>
</tr>
<tr>
<td>FirstSenPos</td>
<td>.31</td>
<td>.31</td>
<td>.43</td>
</tr>
<tr>
<td>$F_{subj}$</td>
<td>.49</td>
<td>.44</td>
<td>.50</td>
</tr>
<tr>
<td>AvgBM25s</td>
<td>.27</td>
<td>.30</td>
<td>.41</td>
</tr>
<tr>
<td>SumBM25s</td>
<td>.50</td>
<td>.44</td>
<td>.52</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Feature</th>
<th>P3</th>
<th>P5</th>
<th>MAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>All Tied</td>
<td>.34</td>
<td>.34</td>
<td>.42</td>
</tr>
</tbody>
</table>
Entity Summarization

• Look at previous documents
  – Entity occurrences so far $F(e,H)$
  – Docs where the entity appeared so far $DF(e,H)$
  – Entity occurrences in the previous doc $F(e,d_{-1})$
  – Frequency of entity the first time? $F(e,d_1)$
  – Number of other entities with which the entity co-occurred so far $CoOcc(e,H)$
• We also tried
  – Weight history features with doc length
  – Weight history features with BM25
Using the History

• Conclusion
  – Evidence from past documents is very important
  – Effectiveness should improve over time (run $F(e,H)$)
Comparing Features

P(e) vs Feature Value

- P(e|F(e,d))
- P(e|F(e,d-1))
- P(e|F(e,d1))
- P(e|F(e,H))
- P(e|DF(e,H))
- P(e|CoOcc(e,H))
- P(e|Fsubj)
Feature Combinations

\[ \text{score}(e, \vec{f}) = \sum_{i=1}^{n} w_i g(f_i, \Theta_i) \]

\[ g(x, t) = \frac{x}{x + t} \]
Combining 2 Features

\[ \text{score}(e, F_1, F_2) = \left( \frac{F_1}{F_1 + t_1} \right) + w \left( \frac{F_2}{F_2 + t_2} \right) \]

Graphs showing the effect of \( w \) on \( \text{AvgPrec} \) for different scenarios:
- \( F(e,d)+(w*F(e,H)) \)
- \( F(e,d)+(w*F(e,d_{-1})) \)
- \( F(e,d)+(w*DF(e,H)) \)
- \( F(e,d)+(w*F(e,d_{-1})) \)
Combining 2 Features

\[ F(e,d) + (w \times F_{\text{Subj}}(e,d)) \]

\[ F(e,d) + (w \times \text{CoOcc}(e,H)) \]

\[ F(e,d) + (w \times \text{FirstSenLen}) \]

\[ F(e,d) + (w \times \text{FirstSenPos}) \]
## Combining 2 features

<table>
<thead>
<tr>
<th>Feature</th>
<th>Function g</th>
<th>W</th>
<th>((F(e,d)+wF)) AvgPrec</th>
</tr>
</thead>
<tbody>
<tr>
<td>(F(e,d))</td>
<td></td>
<td>-</td>
<td>.60</td>
</tr>
<tr>
<td>(F(e,H))</td>
<td></td>
<td></td>
<td>.66</td>
</tr>
<tr>
<td>(F(e,d_{1}))</td>
<td></td>
<td>0.2</td>
<td>.62**</td>
</tr>
<tr>
<td>(F(e,d_{-1}))</td>
<td></td>
<td>0.6</td>
<td>.65**</td>
</tr>
<tr>
<td>(F(e,H))</td>
<td></td>
<td>1.5</td>
<td>.68****</td>
</tr>
<tr>
<td>CoOcc(e,H)</td>
<td></td>
<td>2.2</td>
<td>.67**</td>
</tr>
<tr>
<td>DF(e,H)</td>
<td></td>
<td>0.5</td>
<td>.67****</td>
</tr>
</tbody>
</table>
Combining 3 features

\[
score(e, F_1, F_2, F_3) = \left( \frac{F_1}{F_1 + t_1} \right) + w_1 \left( \frac{F_2}{F_2 + t_2} \right) + w_2 \left( \frac{F_3}{F_3 + t_3} \right)
\]

• Optimizing \( w_1, w_2 \)

<table>
<thead>
<tr>
<th>( F_1 )</th>
<th>( F_2 )</th>
<th>( F_3 )</th>
<th>( w_1, w_2 )</th>
<th>AvgPrec</th>
</tr>
</thead>
<tbody>
<tr>
<td>F(e,d)</td>
<td>F(e,d_{-1})</td>
<td>F(e,H)</td>
<td>0.4, 1.0</td>
<td>0.68</td>
</tr>
<tr>
<td>F(e,d)</td>
<td>CoOcc(e,H)</td>
<td>F(e,H)</td>
<td>0.12, 1.84</td>
<td>0.68</td>
</tr>
<tr>
<td>F(e,d)</td>
<td>CoOcc(e,H)</td>
<td>FistSenLen</td>
<td>2.1, 0.01</td>
<td>0.67</td>
</tr>
<tr>
<td>F(e,d)</td>
<td>CoOcc(e,H)</td>
<td>F(e,d_{-1})</td>
<td>2.2, 0</td>
<td>0.67</td>
</tr>
</tbody>
</table>

• Optimizing \( t_1, t_2, t_3 \)

<table>
<thead>
<tr>
<th>( F_1 )</th>
<th>( F_2 )</th>
<th>( F_3 )</th>
<th>( t_1, t_2, t_3 )</th>
<th>AvgPrec</th>
</tr>
</thead>
<tbody>
<tr>
<td>F(e,d)</td>
<td>F(e,d_{-1})</td>
<td>F(e,H)</td>
<td>5.9, 6.9, 13.8</td>
<td>0.69**</td>
</tr>
</tbody>
</table>
Combining Features with ML

- Logistic Regression for ranking entities
- 5-folds cross validation on 25 topics
- Similar results for combinations of 2 features

<table>
<thead>
<tr>
<th>Local Doc Features</th>
<th>History Features</th>
<th>Features</th>
<th>P3</th>
<th>P5</th>
<th>AvgPrec</th>
</tr>
</thead>
<tbody>
<tr>
<td>F(e,d)</td>
<td>F(e,d₁)</td>
<td>F(e,d)</td>
<td>.65</td>
<td>.56</td>
<td>.60</td>
</tr>
<tr>
<td>FirstSenLen</td>
<td>F(e,d₁)</td>
<td>Local</td>
<td>.65</td>
<td>.56</td>
<td>.62</td>
</tr>
<tr>
<td>FirstSenPos</td>
<td>F(e,H)</td>
<td>History</td>
<td>.66</td>
<td>.60</td>
<td>.67</td>
</tr>
<tr>
<td>F_subj</td>
<td>CoOcc(e,H)</td>
<td>All</td>
<td>.69</td>
<td>.62</td>
<td>.68</td>
</tr>
<tr>
<td>AvgBM25s</td>
<td>DF(e,H)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SumBM25s</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Outline

• Dataset
• Data analysis
• Entity Summarization
• **Entity Profiles**
• Conclusions
There are lots of entities in a story
- 31 judged docs per topic
- 27 judged entities per doc

For which entities we should build profiles?
• 67% of entities appear only in 1 document
• Relevant entities stay relevant all the time

• We pick 708 entities
  – That are relevant at least one day
  – That do not have always the same judgement
Future prediction

• Query: Will entity e be relevant in future?
• Predict appearence (and relevance) of an entity e in future documents given that
  – e has appeared in the past (as Relevant)
  – e does not appear today
  – 7% of entities appear
    • as at least twice as Relevant
    • with a gap in their profiles
Future prediction

• Goal: Extend summary with entities not present in the current doc

• Possible approach:
  – rank e from past docs
  – filter e in current d
  – evaluate with ground truth on future
Conclusions

- Defined new tasks: ES, EP
- Constructed evaluation benchmark
- Entity summarization
  - Investigated some features and combinations
  - Information from the past helps most
  - Obtain 15% improvement over F(e,d)
- Entity profiling
  - How to select interesting entities
Entity Profiles

• Evaluation setting
  – 708 entities
  – Average over all ON/OFF decisions (some entities may have more decisions)
  – Skip non judged (entity, date) pairs
  – Related considered ON-OFF
  – Skip entities which have never been relevant!
  – tp=ON, Rel tn=OFF, NonRel fp=ON, NonRel fn=OFF, Rel
## Entity Profiles

<table>
<thead>
<tr>
<th>Algo</th>
<th>$P=tp/(tp+fp)$</th>
<th>$R=tp/(tp+fn)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Always ON</td>
<td>.84 - .56</td>
<td>1 - 1</td>
</tr>
<tr>
<td>Always OFF</td>
<td>0 - 0</td>
<td>0 - 0</td>
</tr>
<tr>
<td>Freq-based</td>
<td>.84 - .57</td>
<td>.77 - .77</td>
</tr>
<tr>
<td>withPast (t/5)</td>
<td>.85 - .60</td>
<td>.51 - .54</td>
</tr>
<tr>
<td>withPast (t/3)</td>
<td>.85 - .61</td>
<td>.45 - .48</td>
</tr>
<tr>
<td>withPast (t/2)</td>
<td>.85 - .62</td>
<td>.39 - .42</td>
</tr>
<tr>
<td>with Past (t)</td>
<td>.89 - .68</td>
<td>.29 - .33</td>
</tr>
</tbody>
</table>