BioSnowball: Automated Population of Wikis

Xiaojiang Liu, Zaiqing Nie, Nenghai Yu, Ji-Rong Wen

Web Search and Mining Group
Microsoft Research Asia
Outline

• **EntityCube: Web-Scale Entity Summarization**
  – [http://entitycube.research.microsoft.com](http://entitycube.research.microsoft.com)

• **Related Work**
  – Wikipedia: Collaborative Editing
  – Decoupled: Fact Extraction and Biography Summarization

• **BioSnowball**
  – Bio-Fact Duality
  – Bootstrapping framework
  – Joint summarization

• **Experiments**

• **Conclusion**
Wikipedia: Collaborative Editing

- Wikipedia
  - Collaborative Editing
  - First stop for celebrities and notable entities
- Biography
  - Introduction & Infobox
- NPOV Policy
  - Hard for everyday individuals
EntityCube: Web-Scale Entity Summarization

http://entitycube.research.microsoft.com

- Knowledge Mining over 3 Billion Web pages
- Automatically Generates
  - A summarization page for a person
  - A social-network graph for a person
  - A shortest-relationship path between two people
  - All titles of a person that are found on the Web
Automated Fact and Biography Extraction

• Separate Attempts
  – Fact Extraction
    • StatSnowball: Jun Zhu, WWW’09
    • KnowItAll: Oren Etzioni, WWW’04
  – Biography Summarization
    • Sentence Extraction, L. Zhou, EMNLP, 2004
    • Information Extraction, M. Collins, NAACL-ANLP, 2000

• De-coupled Attempts
  – Multi-document summarization via information extraction, M. White, HLT’2001
Bio-Fact Duality

• Good biography contains key facts
• If a text segment is a biography, fact extraction is easier

William Henry “Bill” Gates III (born October 28, 1955) is an American business magnate, philanthropist, author, and chairman of Microsoft, the software company he founded with Paul Allen. He is ranked consistently one of the world’s wealthiest people and the wealthiest overall as of 2009. During his career at Microsoft, Gates held the positions of CEO and chief software architect, and remains the largest individual shareholder with more than 8 percent of the common stock. He has also authored or co-authored several books.
BioSnowball

• Jointly summarize facts and biographies
  – Bio-Fact Duality
  – Markov Logic Networks as the underlying model
    • probabilistic extension of first-order logic
      \[ p(q|x) = \frac{1}{Z(w,x)} \exp \left( \sum_{i \in P_Q} \sum_{j \in G_i} w_i g_j(q,x) \right), \]
      • discriminative model, easy to specify complex relations

• Bootstrapping Framework
  – Easy to get seeds
  – No need to get training data set
Architecture of BioSnowball

P1 (Input):
initial fact and biography seeds
initial model can be provided

P2 (Bootstrapping Summarization Model):
Four Bootstrapping steps
1. Select good patterns
2. Train a joint summarization model
3. Joint Summarize
4. Generate new patterns

P3 (Post-processing and Output):
1. Use facts to do name disambiguation
2. Biography Re-ranking: remove duplicates
Joint Summarization Model

- **Fact Extraction**
  - $\text{Pattern}(e, np, b, +r) \Rightarrow \text{Fact}(e, np, b, +r)$
  - Can be general patterns or keywords patterns

- **Biography Ranking**
  - $\text{BioPattern}(e, b) \Rightarrow \text{BioBlock}(e, b)$,
  - Such as “is born”, “graduates from”
Joint Facts and Biographies

• Fact can be used to Rank Biography
  - $\text{Fact}(e, np, b, +r) \Rightarrow \text{BioBlock}(e, b)$

• Co-reference in Biography
  - $\text{BioBlock}(e, b) \land \text{CoRef}(e, pr) \land \text{Pattern}(pr, np, b, +r)$
    $\Rightarrow \text{Fact}(e, np, b, +r)$

• Facts in Biography are in certain order
  - $\text{BioBlock}(e, b) \land \text{Fact}(e, np, b, +r) \land \text{Next}(np, np)$
    $\Rightarrow \text{Fact}(e, np, b, +r)$
Experiments

• Data set
  – WikiSeed: 17850 with both infoboxes and bio-blocks from Wikipedia
  – Web1M: 1 million web blocks from EntityCube
Bio-Fact Duality

- **Most Frequent Fact Types**
  
<table>
<thead>
<tr>
<th>Property</th>
<th>Occurrence</th>
<th>Hit in Bio</th>
<th>Hit Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>birthdate</td>
<td>16959</td>
<td>16529</td>
<td>0.975</td>
</tr>
<tr>
<td>name</td>
<td>15539</td>
<td>12557</td>
<td>0.808</td>
</tr>
<tr>
<td>birthplace</td>
<td>13762</td>
<td>13142</td>
<td>0.955</td>
</tr>
<tr>
<td>spouse</td>
<td>10888</td>
<td>5471</td>
<td>0.502</td>
</tr>
<tr>
<td>occupation</td>
<td>8022</td>
<td>6553</td>
<td>0.817</td>
</tr>
<tr>
<td>birthname</td>
<td>6897</td>
<td>5524</td>
<td>0.801</td>
</tr>
<tr>
<td>deathdate</td>
<td>6272</td>
<td>6128</td>
<td>0.977</td>
</tr>
<tr>
<td>deathplace</td>
<td>5319</td>
<td>4918</td>
<td>0.925</td>
</tr>
<tr>
<td>location</td>
<td>5013</td>
<td>3624</td>
<td>0.723</td>
</tr>
<tr>
<td>alma mater</td>
<td>3822</td>
<td>3145</td>
<td>0.823</td>
</tr>
</tbody>
</table>

- **Biography Fact Position**
  - First Block: 5.98 facts on average
  - Second: 2.74 facts
  - 3rd to 8th: 1 fact(s)
Joint Summarization & Bootstrapping Model

Table 4: Evaluation of fact extraction results of different joint summarization models

<table>
<thead>
<tr>
<th>Types</th>
<th>Bio</th>
<th>Birth</th>
<th>Death</th>
<th>Overall Facts</th>
</tr>
</thead>
<tbody>
<tr>
<td>bnBioSnowball</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Precision</td>
<td>0.758</td>
<td>0.951</td>
<td>0.619</td>
<td>0.714</td>
</tr>
<tr>
<td>Recall</td>
<td>0.675</td>
<td>0.791</td>
<td>0.325</td>
<td>0.754</td>
</tr>
<tr>
<td>F1</td>
<td>0.714</td>
<td>0.864</td>
<td>0.426</td>
<td>0.734</td>
</tr>
<tr>
<td>jnBioSnowball</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Precision</td>
<td>0.402</td>
<td>0.615</td>
<td>0.742</td>
<td>0.71</td>
</tr>
<tr>
<td>Recall</td>
<td>0.380</td>
<td>0.064</td>
<td>0.442</td>
<td>0.08</td>
</tr>
<tr>
<td>F1</td>
<td>0.390</td>
<td>0.116</td>
<td>0.554</td>
<td>0.144</td>
</tr>
<tr>
<td>jpBioSnowball</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Precision</td>
<td>0.933</td>
<td>0.954</td>
<td>0.758</td>
<td>0.794</td>
</tr>
<tr>
<td>Recall</td>
<td>0.770</td>
<td>0.932</td>
<td>0.543</td>
<td>0.908</td>
</tr>
<tr>
<td>F1</td>
<td>0.844</td>
<td>0.943</td>
<td>0.633</td>
<td>0.847</td>
</tr>
</tbody>
</table>
Conclusions

• EntityCube: Web-Scale Entity Summarization

• BioSnowBall
  – Bio-Fact duality
  – Jointly perform biography ranking and fact extraction in an integrated statistical model.
  – A bootstrapping architecture
  • BioSnowball significantly reduces the number of human-tagged examples and iteratively mines facts and biography blocks.
Questions

Thank you!