Lecture 22: Summarization

Sentence Extraction

\[ D \rightarrow S \]

Selection

- Maximum Marginal Relevance
- Graph algorithms [Mihalcea 05++]

Selection

- Maximum Marginal Relevance
- Graph algorithms
Selection

- Maximum Marginal Relevance
- Graph algorithms

Edges are similarities

Nodes are sentences

Selection

- Maximum Marginal Relevance
- Graph algorithms

Stationary distribution represents node centrality

Nodes are sentences

Edges are similarities

Selection

- Maximum Marginal Relevance
- Graph algorithms
- Word distribution models

<table>
<thead>
<tr>
<th>w</th>
<th>Pr(w)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Obama</td>
<td>0.017</td>
</tr>
<tr>
<td>speech</td>
<td>0.024</td>
</tr>
<tr>
<td>health</td>
<td>0.009</td>
</tr>
<tr>
<td>Montana</td>
<td>0.002</td>
</tr>
</tbody>
</table>

Input document distribution

Summary distribution

Selection

- Maximum Marginal Relevance
- Graph algorithms
- Word distribution models

SumBasic [Nenkova and Vanderwende, 2005]

value(w) = Pr(w)

value(s) = sum of its word values

Choose s, with largest value

Adjust Pr(w)

Repeat until length constraint

Selection

- Maximum Marginal Relevance
- Graph algorithms
- Word distribution models
- Regression models

<table>
<thead>
<tr>
<th>word values</th>
<th>position</th>
<th>length</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>1</td>
<td>24</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>14</td>
</tr>
<tr>
<td>6</td>
<td>3</td>
<td>18</td>
</tr>
</tbody>
</table>

F(x)

frequency is just one of many features
**Summarization Criterion**

\[ S^* = \min_{S: \text{words}(S) \leq L} KL(P_C \parallel P_S) \]

**Raw Count Content Model**

\[ P_C(\cdot) \]

- Barack Obama: 0.15
- Serve America Act: 0.13
- signed: 0.12

President Barack Obama received the Serve America Act after congress’ vote. The ailing senator was instrumental in its passage.

**Structured Content Models**

**General**

- Barack Obama: 0.15
- Serve America Act: 0.13
- signed: 0.12

**Ted Kennedy**

- Ted Kennedy: 0.18
- introduced: 0.12
- ailing senator: 0.11

**AmeriCorps**

- AmeriCorps: 0.11
- Bill Clinton: 0.16
- expanded: 0.08

**Cost**

- cost: 0.11
- republicans: 0.09
- congress: 0.07
- budget: 0.05

**Document Structure**

- **General**
  - Barack Obama: 0.15
  - Serve America Act: 0.13
  - signed: 0.12

- **Specific**
Example General Summary

Former House Speaker Newt Gingrich is asking a judge to force his estranged wife to turn over money he says she is hoarding.

On Thursday, accusers of wrongdoing and the mining of oil in the former U.S. House speaker’s divorce case gave way to a secret settlement between Gingrich and his wife of 18 years, Marianne Gingrich.

Gingrich filed for divorce July 29 amid allegations he is having an affair with 33-year-old congressional aide Callista Bisek.

Example Topical Summary

Gingrich Bio / Post-Speaker Life

Gingrich is best known leading the Republican Party’s takeover of the House in 1994. During that so-called Republican Revolution, Gingrich emphasized that “family values” should be a core pillar in American society.

Since resigning as speaker and from the congressional seat he held for 20 years, Gingrich has been making a living giving speeches, sitting on corporate boards, consulting and appearing as a political analyst on Fox News.

U.S. Rep. J.D. Hayworth (R-Ariz.) argued that Gingrich’s new job as a political commentator for Fox News makes it inappropriate to include him in political gatherings. “Time marches on. He’s gone on to other pursuits,” Hayworth said.

Selection

Optimal search using MMR

- Maximum Marginal Relevance
- Graph algorithms
- Word distribution models
- Regression models
- Topic models
- Globally optimal search

[McDonald, 2007]

Selection

[Intelligent Linear Program]

Maximize \( \sum \alpha_i \lambda_i - \sum \beta_i d_i \)

Subject to \( \sum \beta_i \leq D \)

\( \alpha_i \leq \lambda_i \leq \lambda_i \) \( \forall i,j \)

\( \lambda_i + \lambda_j - \lambda_k \leq 1 \) \( \forall i,j,k \)

\( \alpha_i \in \{0,1\} \) \( \forall i \)

\( \beta_i \in \{0,1\} \) \( \forall i,j \)

[Gillick and Favre, 2008]
Universal health care is a divisive issue.

The health care bill is a major test for the Obama administration.

President Obama remained calm.

Obama addressed the House on Tuesday.
How to include sentence position?

First sentences are unique

Selection

Only allow first sentences in the summary
Up-weight concepts appearing in first sentences
Identify more sentences that look like first sentences

How to include sentence position?

Selection

Some interesting work on sentence ordering
[Barzilay et. al., 1997; 2002]

But choosing independent sentences is easier
- First sentences usually stand alone well
- Sentences without unresolved pronouns
- Classifier trained on OntoNotes: <10% error rate

Baseline ordering module (chronological) is not obviously worse than anything fancier

Selection

Results [G & F, 2009]

52 submissions
27 teams
44 topics
- 10 input docs
- 100 word summaries
Gillick & Favre

- Rating scale: 1-10
- Humans in [8.3, 9.3]
- Rating scale: 0-1
- Humans in [0.62, 0.77]

Error Breakdown?

Sentence extraction is limiting
... and boring!

But abstractive summaries are much harder to generate...

Beyond Extraction?

in 25 words?
Huckleberry Finn
By Mark Twain

Don Quixote
By Cervantes

http://www.rinkworks.com/bookaminute/