Document Summarization

Multi-document Summarization

Extractive Summarization

Selection

Selection

- Maximum Marginal Relevance [Carbonell and Goldstein, 1998]

Greedy search over sentences

Maximize similarity to the query

Minimize redundancy

- Maximum Marginal Relevance

- Graph algorithms [Mihalcea 05++]

... 27,000+ more
Selection

mid-'90s - present

- Maximum Marginal Relevance
- Graph algorithms

Nodes are sentences

Edges are similarities

Selection

mid-'90s - present

- Maximum Marginal Relevance
- Graph algorithms

Nodes are sentences

Selection

mid-'90s - present

- Maximum Marginal Relevance
- Graph algorithms

Nodes are sentences

Stationary distribution represents node centrality

Selection

mid-'90s - present

- Maximum Marginal Relevance
- Graph algorithms
- Word distribution models

Nodes are sentences

Edges are similarities

Selection

mid-'90s - present

- Maximum Marginal Relevance
- Graph algorithms
- Word distribution models

Selection

mid-'90s - present

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

F(x)

word values | position | length
---|---|---
12 | 1 | 24
4 | 2 | 14
6 | 3 | 18

frequency is just one of many features

SumBasic [Nenkova and Vanderwende, 2005]

Value(w_i) = P_i(w_i)

Value(s_i) = sum of its word values

Choose s_i with largest value

Adjust P_i(w_i)

Repeat until length constraint
Selection

- Maximum Marginal Relevance
- Graph algorithms
- Word distribution models
- Regression models
- Topic model-based

[Haghighi and Vanderwende, 2009]

Summarization Criterion

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

\[ P_C(\cdot) \quad P_S(\cdot) \]

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

Barack Obama: 0.18
Serve America Act: 0.16
signed: 0.10

[Haghighi & Vanderwende, NAACL ’09]

Summarization Criterion

Raw Count Content Model

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

Document Structure

President Barack Obama received the Serve America Act after congress’ vote...

The bill is named after Massachusetts Senator Ted Kennedy who was present at its signing. The ailing senator was instrumental...

The legislation would greatly expand the ranks of Ameri-Corps, which was created by President Bill Clinton in 1993...

Structured Content Models

General

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

Ted Kennedy
Amiri-Corps
Cost

Ted Kennedy: 0.18
introduced: 0.12
ailing senator: 0.11
expanded: 0.08

Ameri-Corps: 0.11
Bill Clinton: 0.16

Cost: 0.01
republicans: 0.09
congress: 0.07
budget: 0.05

[Haghighi & Vanderwende, NAACL ’09]
Evaluation: User Study

Pairwise Comparison

<table>
<thead>
<tr>
<th># of User Preferences</th>
<th>PYTHY</th>
<th>H &amp; V 09</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall</td>
<td>20</td>
<td>41</td>
</tr>
<tr>
<td>Redundancy</td>
<td>49</td>
<td></td>
</tr>
<tr>
<td>Coherence</td>
<td>48</td>
<td></td>
</tr>
<tr>
<td>Focus</td>
<td>15</td>
<td></td>
</tr>
</tbody>
</table>

Larger is better

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, accusations of wrongdoing and the timing of a date 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.
Obama addressed the House on Tuesday.

President Obama remained calm.

Selection

Maximize Concept Coverage

Optimization problem: Set Coverage

\[
\max_{s \in S(D)} \sum_{c \in C(s)} v_c
\]

Set of extractive summaries of document set D

Set of concepts present in summary s

Value of concept c

Results

<table>
<thead>
<tr>
<th></th>
<th>Bigram Recall</th>
<th>Pyramid</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>4.00</td>
<td>23.5</td>
</tr>
<tr>
<td>2009</td>
<td>6.65</td>
<td>35.0</td>
</tr>
</tbody>
</table>

(Gillick and Favre, 2008)
Selection

Integer Linear Program for the maximum coverage model

[Gillick, Riedhammer, Favre, Hakkani-Tur, 2008]

\[
\begin{align*}
\text{Maximize:} & \quad \sum_i \alpha_i c_i \\
\text{Subject to:} & \quad \sum_j \sum_i \alpha_i \epsilon_{ij} \leq L \\
\end{align*}
\]

\[
\begin{align*}
\forall i, j, \epsilon_{ij} \\
\sum_j \sum_i \alpha_i \epsilon_{ij} & \leq \epsilon_i \\
\epsilon_i & \in [0, 1] \\
\epsilon_j & \in [0, 1]
\end{align*}
\]

maintain consistency between selected sentences and concepts

This ILP is tractable for reasonable problems

Results [G & F, 2009]

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

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

Error Breakdown?

- Rating scale: 0-1
- Humans in [0.11, 0.15]

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
What would a human do?

It is therefore unsurprising that Lindsay pleaded not guilty yesterday afternoon to the charges filed against her, according to her publicist.

Sentence Rewriting

She flew to France. In France she remained.

$C_{\text{she flew}} = 1$  $C_{\text{she's in France}} = 1$  $C_{\text{she remained}} = 1$

New Optimization problem: Safe Deletions

$$\max_{s \in S(D)} \left[ \sum_{c \in C(s)} v_c + \sum_{d \in D(s)} v_d \right]$$

Value of deletion $d$

How do we know how much a given deletion costs?

$$v_d = w^T f(d)$$
Learning

Features:

\[ f(a) - \left( \sum_{c \in C(a)} f(c) + \sum_{d \in D(a)} f(d) \right) \]

\[ \text{score}(s) = w^1 f(a) \]

Embed ILP in cutting plane algorithm.

Results

<table>
<thead>
<tr>
<th></th>
<th>Bigram Recall</th>
<th>Pyramid</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td></td>
<td>40.0</td>
</tr>
<tr>
<td>2009</td>
<td>6.65</td>
<td>35.0</td>
</tr>
<tr>
<td>Now</td>
<td>17.5</td>
<td>41.0</td>
</tr>
</tbody>
</table>

Beyond Extraction / Compression?

Sentence extraction is limiting

... and boring!

But abstractive summaries are much harder to generate...

in 25 words?

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