Exploring Content Models for Multi-Document Summarization

Aria Haghighi
UC Berkeley

Lucy Vanderwende
MSR Redmond
Summarization
Summarization

$D$
Summarization
Sentence Extraction

\( D \)

\( S \)
Sentence Extraction

$D$
Summarization
Summarization

Representation

$P_c(\cdot)$

Sotomayor: 0.16
court: 0.13
supreme: 0.13
nominee: 0.11
....
Summarization

Representation   Extraction

\[ P_c(\cdot) \]

Sotomayor: 0.16
court: 0.13
supreme: 0.13
nominee: 0.11

....
SumBasic: Representation

[Nenkova & Vanderwende] 2006
SumBasic: Representation

Simple Unigram MLE

[Nenkova & Vanderwende] 2006
SumBasic: Representation

Simple Unigram MLE

\[ P_C(w) = \hat{P}_D(w) \]

- Sotomayor: 0.15
- Washington: 0.13
- supreme: 0.12
- Obama: 0.10

[Renkova & Vanderwende] 2006
SumBasic: Extraction

\[ \text{Score}(S') = \frac{1}{|S'|} \sum_{w \in S} P_c(w) \]
SumBasic: Extraction

Sentence Score

\[ \text{Score}(S') = \frac{1}{|S'|} \sum_{w \in S} P_c(w) \]
SumBasic: Extraction

Sentence Score

\[ \text{Score}(S') = \frac{1}{|S'|} \sum_{w \in S} P_c(w) \]

“Obama announced the nomination of Sonia Sotomayor”
SumBasic: Extraction

Sentence Score

\[ \text{Score}(S') = \frac{1}{|S'|} \sum_{w \in S} P_c(w) \]

“Obama announced the nomination of Sonia Sotomayor”

0.12  0.01  0.05  0.04  0.15
SumBasic: Extraction

Sentence Score

\[ \text{Score}(S') = \frac{1}{|S'|} \sum_{w \in S} P_c(w) \]

Score: 0.074

"Obama announced the nomination of Sonia Sotomayor"

0.12 0.01 0.05 0.04 0.15
Score($S'$) = $\frac{1}{|S'|} \sum_{w \in S} P_c(w)$

$S = \{\}$
\begin{align*}
\text{Score}(S) &= \frac{1}{|S|} \sum_{w \in S} P_c(w) \\
S &= \{\} \\
\text{while } words(S) < L:
\end{align*}
**SumBasic: Extraction**

$$\text{Score}(S) = \frac{1}{|S|} \sum_{w \in S} P_c(w)$$

$S = \{\}$

while $\text{words}(S) < L$:

$S^* = \max_{S \notin S} \text{Score}(S)$
**SumBasic: Extraction**

\[
\text{Score}(S) = \frac{1}{|S|} \sum_{w \in S} P_c(w)
\]

\[
S = \{
\}
\]

while \( \text{words}(S) < L \):

\[
S^* = \max_{S \notin S} \text{Score}(S)
\]

\[
S = S \cup S^*
\]
SumBasic: Extraction

\[ \text{Score}(S) = \frac{1}{|S|} \sum_{w \in S} P_c(w) \]

\[ S = \{\} \]

while \( \text{words}(S) < L \):

\[ S^* = \max_{S \not\subseteq S} \text{Score}(S) \]

\[ S = S \cup S^* \]

\[ P_c(w) = P_c(w)^2, \text{ for } w \in S^* \]
Experimental Results
Experimental Results

- DUC 2006
Experimental Results

- **DUC 2006**
  - 50 document sets, 25 docs each
Experimental Results

- **DUC 2006**
  - 50 document sets, 25 docs each
  - Max 250 tokens
Experimental Results

- **DUC 2006**
  - 50 document sets, 25 docs each
  - Max 250 tokens

- **ROUGE-2**
Experimental Results

- **DUC 2006**
  - 50 document sets, 25 docs each
  - Max 250 tokens

- **ROUGE-2**
  - Recall over bigrams w/o stop words against human summaries
Experimental Results

- DUC 2006
  - 50 document sets, 25 docs each
  - Max 250 tokens

- ROUGE-2
  - Recall over bigrams w/o stop words against human summaries
  - Bad at summary quality, decent at content
SumBasic: Performance

SumBasic  5.3
SumBasic: Extraction Issues
SumBasic: Extraction Issues

- What are we optimizing?
SumBasic: Extraction Issues

- What are we optimizing?
  - Without word length, when to stop?
SumBasic: Extraction Issues

- What are we optimizing?
  - Without word length, when to stop?
- Not Recall Oriented
SumBasic: Extraction Issues

- What are we optimizing?
  - Without word length, when to stop?
- Not Recall Oriented
  - No direct penalty for missing freq. words
KLSum: Extraction

$P_c(\cdot)$

Sotomayor: 0.15
Washington: 0.13
supreme: 0.12
KLSum: Extraction

\[ P_c(\cdot) \]

- Sotomayor: 0.15
- Washington: 0.13
- supreme: 0.12
KLSum: Extraction

$P_c(\cdot)$

Sotomayor: 0.15
Washington: 0.13
supreme: 0.12

$P_s(\cdot)$

Sotomayor: 0.20
Obama: 0.14
Washington: 0.11

Saturday, September 26, 2009
KLSum: Extraction

$P_c(\cdot)$

- Sotomayor: 0.15
- Washington: 0.13
- supreme: 0.12

$P_s(\cdot)$

- Sotomayor: 0.20
- Obama: 0.14
- Washington: 0.11
KLSum: Extraction

\[ P_c(\cdot) \]

Sotomayor: 0.15  
Washington: 0.13  
supreme: 0.12

\[ P_s(\cdot) \]

Sotomayor: 0.18  
Washington: 0.11  
supreme: 0.10
KLSum: Extraction

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

\( P_C(\cdot) \)
- Sotomayor: 0.15
- Washington: 0.13
- supreme: 0.12

\( P_S(\cdot) \)
- Sotomayor: 0.18
- Washington: 0.11
- supreme: 0.10

See Paper For Details
KLSum: Performance

<table>
<thead>
<tr>
<th></th>
<th>SumBasic</th>
<th>KLSum</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>5.3</td>
<td>6.0</td>
</tr>
</tbody>
</table>
Improving Representation
Improving Representation

- Flexible stop words
Improving Representation

- Flexible stop words
  - e.g. “stock” in financial document sets
Improving Representation

- **Flexible stop words**
  - e.g. “stock” in financial document sets
- **No pref. for multi-document usage**
Improving Representation

- **Flexible stop words**
  - e.g. “stock” in financial document sets
- **No pref. for multi-document usage**
  - Many docs indicate content importance
Improving Representation

- Flexible stop words
  - e.g. “stock” in financial document sets
- No pref. for multi-document usage
  - Many docs indicate content importance
- Prefer words in early sentences
Adding Topics

Background Distribution
Adding Topics

Background Distribution

$\phi_B$

- the: 0.12
- of: 0.09
- ....
- washington: 0.04
- policy: 0.03
Adding Topics

Content Distribution
Adding Topics

Content Distribution

\[ \phi C \]

Sotomayor: 0.16
supreme: 0.13
Obama: 12
court: 11
nominee: 10

....
Adding Topics

Document-Specific Distribution

similar to [Daume & Marcu, 2006]
Adding Topics

Document-Specific Distribution

\[ \phi_D \]

similar to [Daume & Marcu, 2006]
Adding Topics

Document-Specific Distribution

similar to [Daume & Marcu, 2006]
Adding Topics

Document-Specific Distribution

similar to [Daume & Marcu, 2006]
Adding Topics

Document-Specific Distribution

similar to [Daume & Marcu, 2006]
Adding Topics

Document Level

•

•

•

•
Adding Topics

Document Level

\[ \psi_t \]

\[ B \quad D \quad C \]
Adding Topics

Document Level

\[ B \quad D \quad C \]

\[ \psi_t \]

\[ B \quad D \quad C \]

\[ \psi_t \]
Adding Topics

Sentence

\[ B \quad D \quad C \]

\[ \psi_t \]
Adding Topics

Sentence

Word

$B \quad D \quad C$

$\psi_t$
Adding Topics

Sentence

Word

\[ Z \rightarrow W \leftarrow \psi_t \]

\[ B \quad D \quad C \]
Adding Topics

Sentence

Word \{B,D,C\}

\[ \psi_t \]
Adding Topics

Sentence

Word

$Z$

$W$

$\psi_t$

$B$ $D$ $C$
Adding Topics

Sentence

Word

$\psi_t$

$B$ $D$ $C$

$\phi_B \phi_D \phi_C$

$W$

$Z$
Adding Topics

Representation  Extraction

$\phi C$

Sotomayor: 0.16
supreme: 0.13
Obama: 0.12
court: 0.11
nominee: 0.10
....

Saturday, September 26, 2009
Adding Topics

Representation

Extraction

$\phi C$

Sotomayor: 0.16
supreme: 0.13
Obama: 0.12
court: 0.11
nominee: 0.10
....
Adding Bigrams

Each sentence is a bag of bigrams
Adding Bigrams

Each sentence is a bag of bigrams

\[ \phi C \]

Obama announced: 0.09
Sonia Sotomayor: 0.08
Supreme Court: 0.05
.....
Performance

<table>
<thead>
<tr>
<th>Method</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>SumBasic</td>
<td>5.3</td>
</tr>
<tr>
<td>KLSum</td>
<td>6.0</td>
</tr>
<tr>
<td>+Topic</td>
<td>6.3</td>
</tr>
<tr>
<td>+Bigram</td>
<td>7.8</td>
</tr>
</tbody>
</table>
Structured Content Models
Structured Content Models

General Content

Sotomayor: 0.16
supreme: 0.13
court: 0.12
....
Structured Content Models

General Content

Specific Content “Sub-Stories”

Sotomayor: 0.16
supreme: 0.13
court: 0.12
....
Structured Content Models

General Content

Sotomayor: 0.16
supreme: 0.13
court: 0.12
....

Specific Content “Sub-Stories”

born: 0.15
puerto: 0.13
mother: 0.12
father: 0.10
...

confirmation: 0.16
Republican: 0.11
senators: 0.08
Limbaugh: 0.07
...

race: 0.11
identity: 0.09
firefighters: 0.07
discrimination: 0.05
...
Example Sub-Story

Biography

Sonia Sotomayor Born to Puerto Rican parents who moved to the Bronx in New York during World War Two.
Biography

Sonia Sotomayor Born to Puerto Rican parents who moved to the Bronx in New York during World War Two.
Example Sub-Story

Confirmation

Senate Republicans have combed over Sotomayor’s record on the federal bench ahead of confirmation hearings.

Saturday, September 26, 2009
Senate Republicans have combed over Sotomayor’s record on the federal bench ahead of confirmation hearings.
Structured Content Models

\[ \phi C_0 \]

Sotomayor: 0.16
supreme: 0.13
Obama: 0.12
......

\[ \phi C_1 \]

born: 0.15
puerto: 0.13
mother: 0.12
father: 0.10
...

\[ \phi C_2 \]

confirmation: 0.16
Republican: 0.11
senators: 0.08
Limbaugh: 0.07
...

\[ \phi C_3 \]

race: 0.11
identity: 0.09
firefighters: 0.07
discrimination: 0.05
...

Saturday, September 26, 2009
Structured Topics

General or Specific?
Structured Topics

General or Specific?

\[ B \quad D \quad C \]

\[ \psi_t \]
Structured Topics

General or Specific?

\[ B \quad D \quad C \]

\[ G \quad S \]

\[ \psi_t \]

\[ \psi_G \]
Structured Topics

General or Specific?

B  D  C
ψₜ

G  S
ψ₉
Structured Topics

General or Specific?

\[ B \quad D \quad C \quad G \quad S \]

\[ \psi_t \quad \psi_G \]
Structured Topics

What specific topic?
Structured Topics

What specific topic?
Structured Topics

What specific topic?

$\mathbb{Z}_S \{1,2,3\}$
Structured Topics

What specific topic?
Structured Topics

What specific topic?

“Sticky” Specific Topics

\[
P(Z'_S | Z_S) = \begin{cases} 
\sigma & \text{if } Z'_S = Z_S \\
(1 - \sigma)\theta_{Z'_S | Z_S} & \text{o.w.}
\end{cases}
\]
Structured Topics

What specific topic?

“Sticky” Specific Topics

\[ P(Z'_S | Z_S) = \begin{cases} 
\sigma \\
(1 - \sigma)\theta_{Z'_S | Z_S} 
\end{cases} \]

if \( Z'_S = Z_S \)
o.w.
Structured Topics

Representation

Extraction

\( \phi C_0 \)

Sotomayor: 0.16
supreme: 0.13
Obama: 0.12
court: 0.11
nominee: 0.10
.....
Structured Topics

Representation

Extraction

\( \phi C_0 \)

Sotomayor: 0.16
supreme: 0.13
Obama: 0.12
court: 0.11
nominee: 0.10
....
Performance

- SumBasic: 5.3
- +Topic: 6.3
- +Struct: 6.4
- +Bigram: 8.3

Saturday, September 26, 2009
HierSum: Manual Evaluation

- **Pairwise Comparison**
  - 23 participants, 69 judgments
  - Each participant sees reference summary and two model summaries
  - Pythy: State-of-the-art discriminative system. Highest automatic score and high-performing on manual content evaluation
## Manual Evaluation

<table>
<thead>
<tr>
<th>Question</th>
<th>Pythy</th>
<th>Ours</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall</td>
<td>20</td>
<td>49</td>
</tr>
<tr>
<td>Redundancy</td>
<td>21</td>
<td>48</td>
</tr>
<tr>
<td>Coherence</td>
<td>15</td>
<td>54</td>
</tr>
<tr>
<td>Focus</td>
<td>28</td>
<td>41</td>
</tr>
</tbody>
</table>
DUC ’07 Results

- SumBasic: 5.9
- Pythy: 8.7
- Ours: 9.3

Saturday, September 26, 2009
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 mining of dirt 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.

Newt Gingrich's attorney, Randy Evans, says Marianne Gingrich has refused to even discuss a settlement until she questions Bisek.
Callista Bisek

Marianne Gingrich also wants to depose Callista Bisek, a congressional aide with whom the former U.S. House speaker has a relationship.

And in motions filed last week in Superior Court for the District of Columbia, Callista Bisek, a clerk for the House Agriculture Committee, asked a judge to overturn a Georgia court order requiring her to answer questions about her relationship with Gingrich.

Mayoue has said he intends to question Bisek about all aspects of her relationship with Newt Gingrich.
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.
Conclusion

- KL objective for sentence extraction summarization
- Topic models can yield state-of-the-art automatic and manual summary eval
- Also structured content models for topical summarization
Thanks!

Questions?

If not, I have more examples........
Slain chimp’s owner now says it wasn’t on Xanax
STAMFORD, Conn. - As authorities considered criminal charges, the woman whose 200-pound domesticated chimpanzee went berserk and mauled a friend backtracked Wednesday on whether she gave the animal the anti-anxiety drug Xanax.

Owner of chimp that went on Conn. rampage changes story, says she never gave animal Xanax
Sandra Herold told The Associated Press on Wednesday that she never gave the drug to her 14-year-old chimp, Travis, who was shot dead by Stamford police Monday after he grievously wounded Herold’s friend Charla Nash.

Owner of chimp that went on Conn. rampage changes story, says she never gave animal Xanax
In humans, Xanax can lead to aggression in people who are unstable to begin with, said Dr. Emil Coccaro, chief of psychiatry at the University of Chicago Medical Center.
Topical Summarization

Articles:

Kosovo comes back to bite the US
Ten days ago, a full-scale war broke out when Russian and Georgian forces clashed over the breakaway Georgian region of South Ossetia.

Russia will occupy buffer zone in Georgian territory
Anatoly Nogovitsyn, deputy chief of the Russian military's general staff, said a battalion of about 270 soldiers would occupy a swath of Georgian territory around the enclaves of Abkhazia and South Ossetia after the withdrawal of troops from central Georgia.