

Statistical NLP Spring 2010



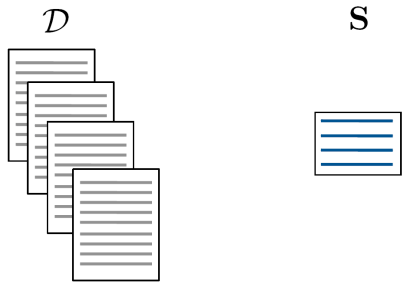
Lecture 22: Summarization

Dan Klein – UC Berkeley
Includes slides from Aria Haghighi, Dan Gillick

Summarization



Sentence Extraction



Selection

mid-'90s

- Maximum Marginal Relevance [Carbonell and Goldstein, 1998]

Greedy search over sentences

present

Maximize similarity to the query

Minimize redundancy

$$MMR = \underset{S_1, \dots, S_n}{\operatorname{argmax}} \left(\lambda \operatorname{Sim}_1(D, Q) + (1 - \lambda) \operatorname{Sim}_2(D_1, D_2) \right)$$

Selection

mid-'90s

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

present

Selection

mid-'90s

- Maximum Marginal Relevance
- Graph algorithms

present

Nodes are sentences

Selection

mid-'90s

- Maximum Marginal Relevance
- Graph algorithms

present

Nodes are sentences

Edges are similarities

Selection

mid-'90s

- Maximum Marginal Relevance
- Graph algorithms

present

Stationary distribution represents node centrality

Nodes are sentences

Edges are similarities

Selection

mid-'90s

- Maximum Marginal Relevance
- Graph algorithms
- Word distribution models

present

w	$P_0(w)$
Obama	0.017
speech	0.024
health	0.009
Montana	0.002

Input document distribution

~

w	$P_A(w)$
Obama	?
speech	?
health	?
Montana	?

Summary distribution

Selection

mid-'90s

- Maximum Marginal Relevance
- Graph algorithms
- Word distribution models

present

SumBasic [Nenkova and Vanderwende, 2005]

Value(w_i) = $P_0(w_i)$

Value(s_i) = sum of its word values

Choose s_i with largest value

Adjust $P_0(w)$

Repeat until length constraint

Selection

mid-'90s

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

present

word values	position	length	
S1	12	1	24
S2	4	2	14
S3	6	3	18

frequency is just one of many features

Selection

mid-'90s

- Maximum Marginal Relevance
- Graph algorithms
- Word distribution models
- Regression models
- Topic model-based [Haghighi and Vanderwende, 2009]

present

University of California Berkeley

Summarization Criterion

$P_C(\cdot)$
 Barack Obama: 0.15
 Serve America Act: 0.13
 signed: 0.12

$P_S(\cdot)$
 Barack Obama: 0.18
 Serve America Act: 0.16
 signed: 0.10

University of California Berkeley

Summarization Criterion

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

$P_C(\cdot)$
 Barack Obama: 0.15
 Serve America Act: 0.13
 signed: 0.12

$P_S(\cdot)$
 Barack Obama: 0.18
 Serve America Act: 0.16
 signed: 0.10

[Haghighi & Vanderwende, NAACL '09]

University of California Berkeley

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.

University of California Berkeley

Document Structure

General
 Ted Kennedy
 Ameri Corps

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...

University of California Berkeley

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

Ameri-Corps
 Ameri-Corps: 0.11
 Bill Clinton: 0.16
 expand: 0.08
 ...

Cost
 cost: 0.11
 republicans: 0.09
 congress: 0.07
 budget: 0.05
 ...

[Haghighi & Vanderwende, NAACL '09]

University of California Berkeley

Document Structure

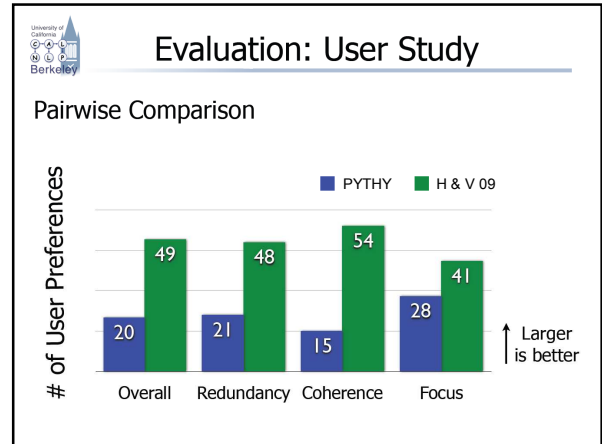
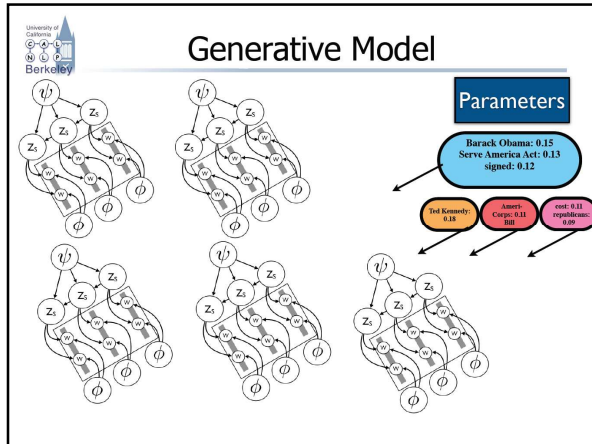
General
 ↓
 Specific

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

Ted Kennedy: 0.18
 introduced: 0.12
 ailing senator: 0.11

Ameri-Corps: 0.11
 Bill Clinton: 0.16
 expand: 0.08
 ...

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



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 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.

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

mid-'90s

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

present

[McDonald, 2007]

Optimal search using MMR

Integer Linear Program

Maximize: $\sum_i Rel_i s_i - \sum_{ij} Red_{ij} s_i s_j$

Subject to: $\sum_j I_j s_j \leq L$

$s_{ij} \leq s_i \quad s_{ij} \leq s_j \quad \forall i, j$

$s_i + s_j - s_{ij} \leq 1 \quad \forall i, j$

$s_i \in \{0, 1\} \quad \forall i$

$s_{ij} \in \{0, 1\} \quad \forall i, j$

Selection

[Gillick and Favre, 2008]

- S1 The health care bill is a major test for the Obama administration.
- S2 Universal health care is a divisive issue.
- S3 President Obama remained calm.
- S4 Obama addressed the House on Tuesday.

concept	value

Selection

[Gillick and Favre, 2008]

- S1 The health care bill is a major test for the Obama administration.
- S2 Universal health care is a divisive issue.
- S3 President Obama remained calm.
- S4 Obama addressed the House on Tuesday.

concept	value
obama	3

Selection

[Gillick and Favre, 2008]

- S1 The health care bill is a major test for the Obama administration.
- S2 Universal health care is a divisive issue.
- S3 President Obama remained calm.
- S4 Obama addressed the House on Tuesday.

concept	value
obama	3
health	2

Selection

[Gillick and Favre, 2008]

- S1 The health care bill is a major test for the Obama administration.
- S2 Universal health care is a divisive issue.
- S3 President Obama remained calm.
- S4 Obama addressed the House on Tuesday.

concept	value
obama	3
health	2
house	1

Selection

[Gillick and Favre, 2008]

- S1 The health care bill is a major test for the Obama administration.
- S2 Universal health care is a divisive issue.
- S3 President Obama remained calm.
- S4 Obama addressed the House on Tuesday.

concept	value
obama	3
health	2
house	1

Length limit: 18 words

summary	length	value
{S1, S3}	17	5
{S2, S3, S4}	17	6

← greedy (points to {S1, S3})
← optimal (points to {S2, S3, S4})

Selection

Integer Linear Program for the maximum coverage model
[Gillick, Riedhammer, Favre, Hakkani-Tur, 2008]

Maximize: $\sum_i w_i c_i$ ← total concept value

Subject to: $\sum_j l_j s_j \leq L$ ← summary length limit

$s_j Occ_{ij} \leq c_i, \forall i, j$ ← maintain consistency between selected sentences and concepts

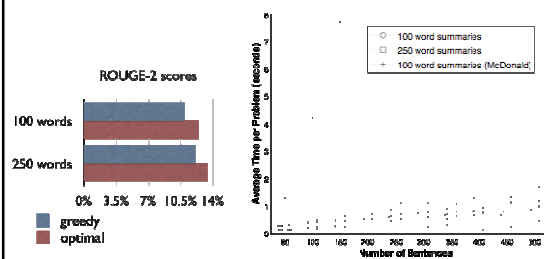
$\sum_j s_j Occ_{ij} \geq c_i \quad \forall i$

$c_i \in \{0, 1\} \quad \forall i$

$s_j \in \{0, 1\} \quad \forall j$

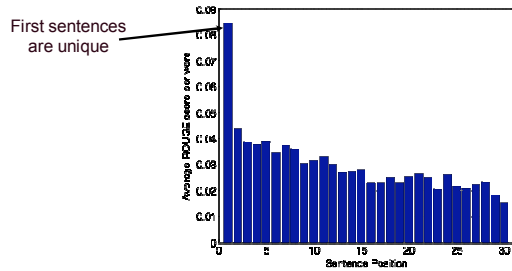
Selection

[Gillick and Favre, 2009]



This ILP is tractable for reasonable problems

Selection



How to include sentence position?

Selection

Only allow first sentences in the summary

surprisingly strong baseline

Up-weight concepts appearing in first sentences

included in TAC 2009 system

Identify more sentences that look like first sentences

first sentence classifier is not reliable enough yet

How to include sentence position?

Selection

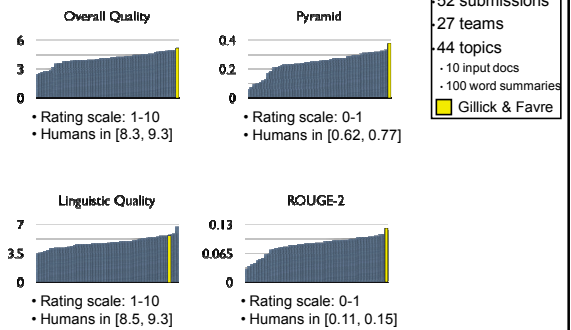
Some interesting work on sentence ordering [Barzilay et. al., 1997; 2002]

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

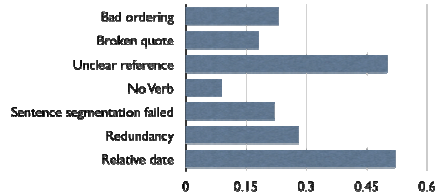
Results [G & F, 2009]



Error Breakdown?

[Gillick and Favre, 2008]

culprits in low-scoring summaries

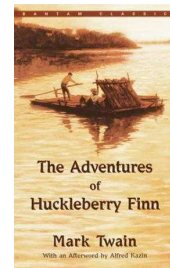


Beyond Extraction?

Sentence extraction is limiting

... and boring!

But abstractive summaries are much harder to generate...



in 25 words?

BOOK-A-MINUTE *Classics*

Huckleberry Finn

By Mark Twain

Ultra-Condensed by David J. Parker

Huckleberry Finn

(Goes rafting. Goes home.)

THE END

Don Quixote

By Cervantes

Ultra-Condensed by Scott Kivides

Don Quixote

Chivalry demands I destroy that evil thing.

Sancho Panza

No, master. It is something ordinary and harmless.

Don Quixote

(falls down)

THE END

<http://www.rinkworks.com/bookamminute/>