
Automatic
summarization
vertices
summary
choice
information
heuristic
paraphrasing
classifier
similarity
extrinsic
keyphrases
words
binary
edge
multiple
syntactic
threshold
supervised
algorithmically
methods
random
final
probability
form
graph
semantic
importance
sentence
automatic
step
examples
summarization
phrase
adaptive
diversity
feature
abstraction
classification
original
document
automatically
extraction
number
coherence
news
qualitatively
metric
vertex
similar
keyphrase
unsupervised
trigram
centrality
corpus
undirected
training
important
stationary
technique
system

From Single to Multi-Document Summarization: A Protoype System and Its Evaluation

Lin & Hovy: ACL 2002

by Dan Vollmer

Why Summarize?

Prezi

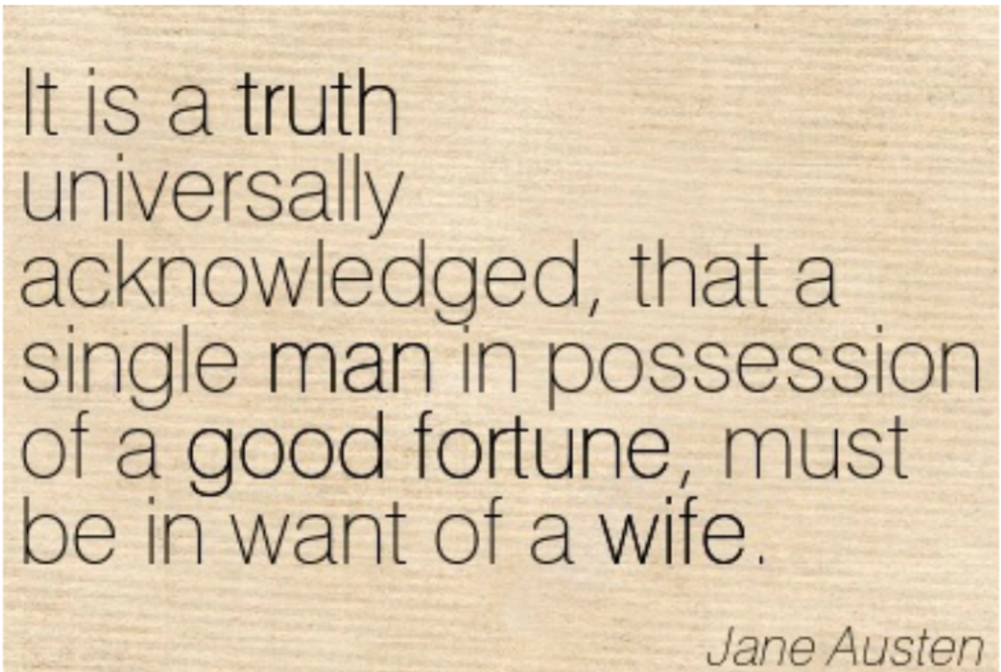
Abstractive

"I read War and Peace....
It involves Russia."
(Woody Allen)

- "Gisting"
- Text comprehended
- Reformulated in shorter words
- Quite difficult and very little work until recently

Prezi

Extractive



- Salient sentences drawn-out
- Relatively easy
- Method of most summarization systems

Elements of a Summary

Brevity

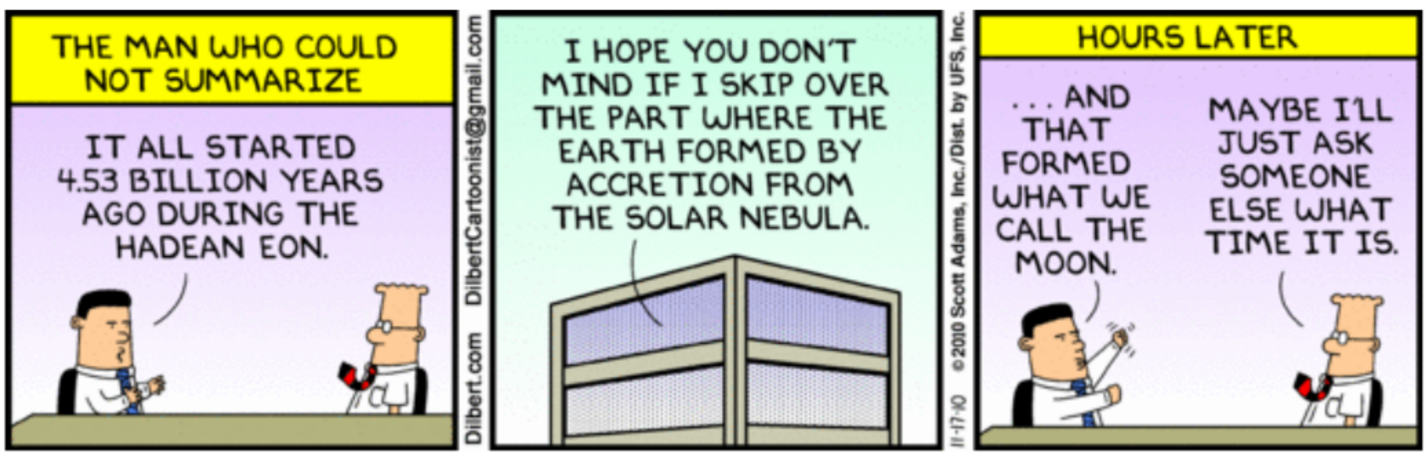
- No longer than half the original text
- But, we can go shorter as well
- DUC tasks: 50, 100, 200, 400 words

Relevance

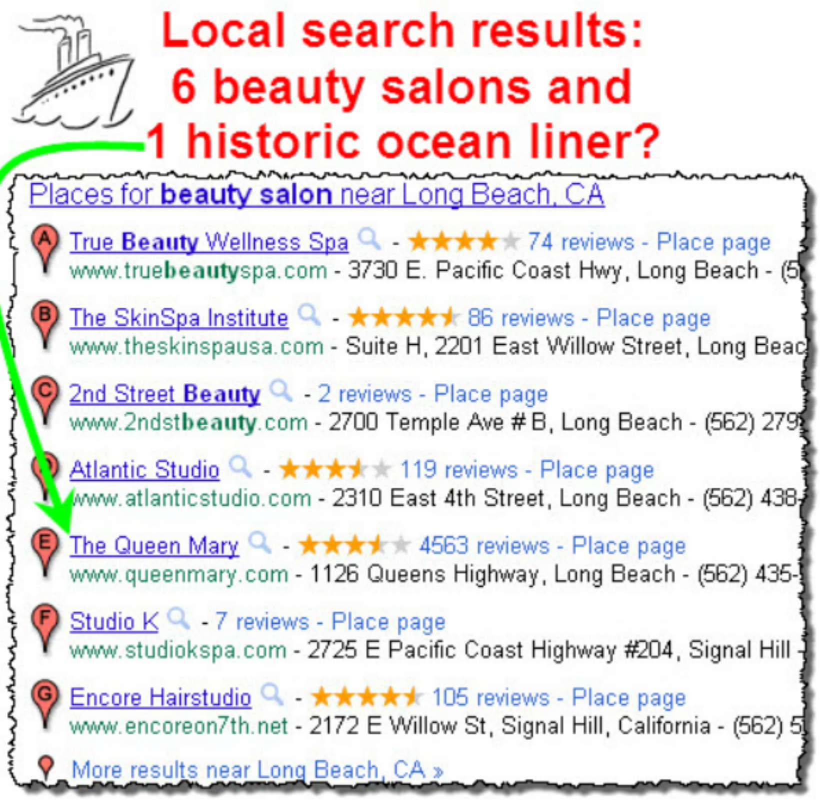
- We need to distill the document to central concepts
- Exclude irrelevant and redundant information

Brevity

- No longer than half the original text
- But, we can go shorter as well
- DUC tasks: 50, 100, 200, 400 words

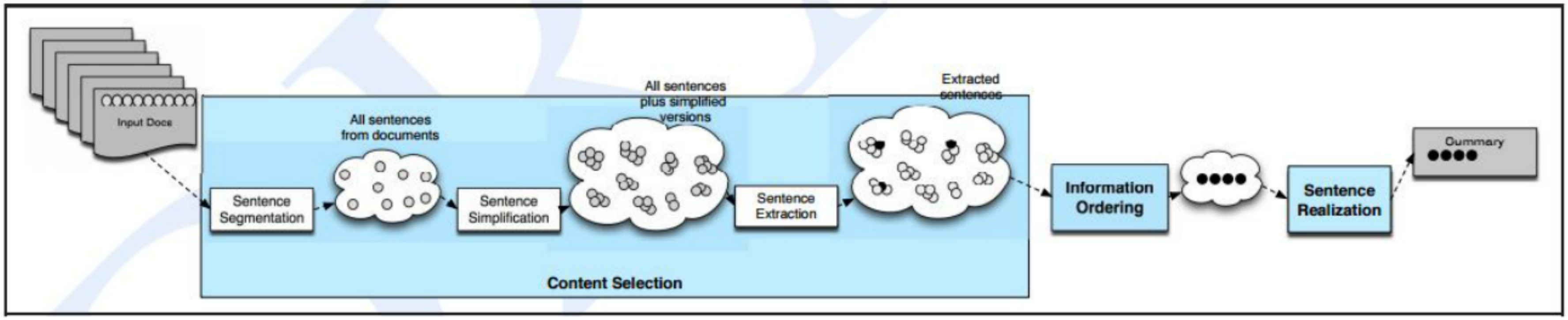


Relevance



- We need to distill the document to central concepts
- Exclude irrelevant and redundant information

NeATS



- Authors' prototype system
- Takes an input set of newspaper articles
- summaries are created via three steps:
    - Selection
    - Filtering
    - Presentation

Prezi

NeATS Content Selection

$$Log\ Likelihood = -2log\lambda$$

$$\lambda = \frac{\max_{\omega\varepsilon\Omega_0} H(\omega;k)}{\max_{\omega\varepsilon\Omega} H(\omega;k)}$$

Where:

- Omegas are parameters
- K's are observations

Log Likelihood is then used to identify relevant n-grams

Rank	Unigram	(-2λ)	Bigram	(-2λ)	Trigram	(-2λ)
1	Slovenia	319.48	federal army	21.27	Slovenia central bank	5.80
2	Yugoslavia	159.55	Slovenia Croatia	19.33	minister foreign affairs	5.80
3	Slovene	87.27	Milan Kucan	17.40	unallocated federal debt	5.80
4	Croatia	79.48	European Community	13.53	Dmovsek prime minister	3.86
5	Slovenian	67.82	foreign exchange	13.53	European Community countries	3.86

Figure 2. Top 5 unigram, bigram, and trigram concepts for topic "Slovenia Secession from Yugoslavia".

- Compute the likelihood ratio
- Then identify key concepts in unigrams, bigrams, and trigrams
- On-topic & Off-topic document collections used to learn relevancy
- Concepts are clustered to find major subtopics
- Via strict lexical lookup
- Each sentence then ranked based on key concepts contained
- Not much time is devoted to the algorithm...

$$Log\ Likelihood = -2log\lambda$$

$$\lambda = \frac{\max_{\omega \varepsilon \Omega_0} H(\omega; k)}{\max_{\omega \varepsilon \Omega} H(\omega; k)}$$

Where:

- Omegas are parameters
- K's are observations

Log Likelihood is then used to identify relevant n-grams

Rank	Unigram	(-2λ)	Bigram	(-2λ)	Trigram	(-2λ)
1	Slovenia	319.48	federal army	21.27	Slovenia central bank	5.80
2	Yugoslavia	159.55	Slovenia Croatia	19.33	minister foreign affairs	5.80
3	Slovene	87.27	Milan Kucan	17.40	unallocated federal debt	5.80
4	Croatia	79.48	European Community	13.53	Drnovsek prime minister	3.86
5	Slovenian	67.82	foreign exchange	13.53	European Community countries	3.86

Figure 2. Top 5 unigram, bigram, and trigram concepts for topic "Slovenia Secession from Yugoslavia".

Ordering

Given a selected set of sentences, choose the optimal order for presenting them in a summary.

- "Optimal" usually defined using some distance measure
- E.g. TF-IDF & cosine similarity
- Can anyone see the challenge here?

- NeATS ranking causes lots of tie-scores, so filtering is needed...

FILTERING

Position

- Use genre specific knowledge
- Identify important sections in documents
- Edmundson (1969)
- NeATS is simple - first 10 sentences only

Stigma Words

Some words are likely to cause incongruities

- conjunctions
- the verb "say"
- quotation marks
- pronouns

NeATS doesn't do any discourse level selection

So, we just penalize sentences containing stigma words to drop their overall scores

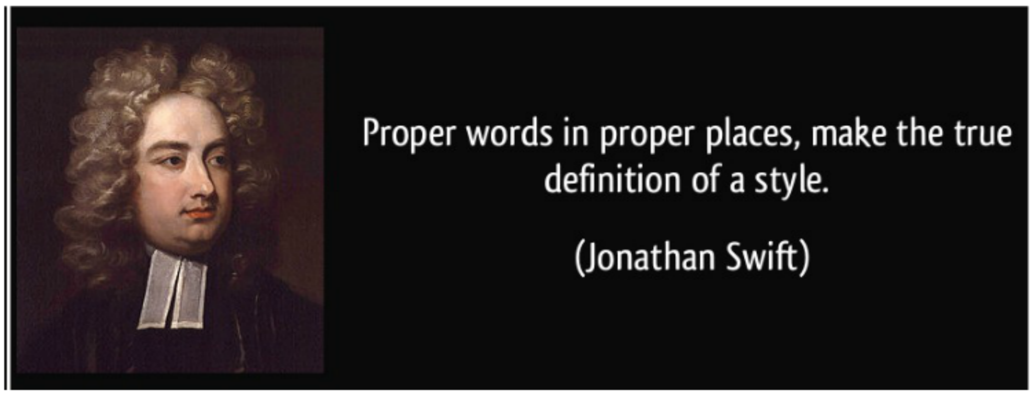
MMR

Maximal Marginal Relevance

$$MMR \stackrel{def}{=} \arg\max_{D_i \in R-S}\left[\lambda(Sim_1(D_i, Q) - (1-\lambda)\max_{D_j \in S} Sim_2(D_i, D_j))\right]$$

- *"Relevant Novelty"*
- Q ~ document centroid/user query
- D ~ document collection
- R ~ ranked list
- S ~ subset of documents in R already selected
- Sim ~ similarity metric (e.g. term frequency)
- Lambda = 1 produces most significant ranked list
- Lambda = 0 produces most diverse ranked list

Position



- Use genre specific knowledge
- Identify important sections in documents
- Edmundson (1969)
- NeATS is simple - first 10 sentences only

Stigma Words

Some words are likely to cause incongruities

- conjunctions
- the verb "say"
- quotation marks
- pronouns

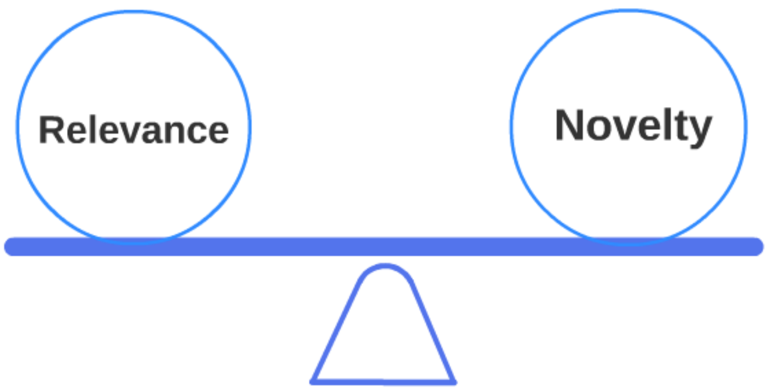
NeATS doesn't do any discourse level selection

So, we just penalize sentences containing stigma words to drop their overall scores

MMR

Maximal Marginal Relevance

$$MMR \stackrel{\text{def}}{=} \arg\max_{D_i \epsilon R-S} \left[\lambda(Sim_1(D_i, Q) - (1-\lambda) \max_{D_i \epsilon S} Sim_2(D_i, D_j)) \right]$$



- "*Relevant Novelty*"
- Q ~ document centroid/user query
- D ~ document collection
- R ~ ranked list
- S ~ subset of documents in R already selected
- Sim ~ similarity metric (e.g. term frequency)
- Lambda = 1 produces most significant ranked list
- Lambda = 0 produces most diverse ranked list

PRESENTATION

The Buddy System

How to handle definite noun phrases?

- E.g. "The...drought relief program of 1988" needs some context
- NeATS explicitly chooses an introductory sentence for context
- Assumed that lead sentences of documents contain introductory information

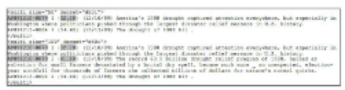


Figure 3. 50 and 100 word summaries for topic "US Drought of 1988".

Time Annotation and Sequencing

Examples

- weekdays (Sunday, Monday, etc.)
- (past | next | coming) + weekdays
- today, yesterday, last night

- A type of ordering - not NP-hard
- Sorting out temporal relationships
- Since the evaluation task uses news articles, publication dates allow for explicit computation of dates
- Ordering is relatively straightforward thereafter

Figure 4. [illegible]

The Buddy System

How to handle definite noun phrases?

- E.g. "The...drought relief program of 1988" needs some context
- NeATS explicitly chooses an introductory sentence for context
- Assumed that lead sentences of documents contain introductory information

```
<multi size="50" docset="d50i">
AP891210-0079 1 (32.20) (12/10/89) America's 1988 drought captured attention everywhere, but especially in
Washington where politicians pushed through the largest disaster relief measure in U.S. history.
AP891213-0004 1 (34.60) (12/13/89) The drought of 1988 hit …
</multi>
<multi size="100" docset="d50i">
AP891210-0079 1 (32.20) (12/10/89) America's 1988 drought captured attention everywhere, but especially in
Washington where politicians pushed through the largest disaster relief measure in U.S. history.
AP891210-0079 3 (41.18) (12/10/89) The record $3.9 billion drought relief program of 1988, hailed as
salvation for small farmers devastated by a brutal dry spell, became much more _ an unexpected, election-
year windfall for thousands of farmers who collected millions of dollars for nature's normal quirks.
AP891213-0004 1 (34.60) (12/13/89) The drought of 1988 hit …
</multi>
```

Figure 3. 50 and 100 word summaries for topic "US Drought of 1988".

Prezi

- NeATS explicitly chooses an introductory sentence for context
- Assumed that lead sentences of documents contain introductory information

```
multi size="50" docset="d50i">
P891210-0079 1 (32.20) (12/10/89) America's 1988 drought captured attention everywhere, but especially in
ashington where politicians pushed through the largest disaster relief measure in U.S. history.
P891213-0004 1 (34.60) (12/13/89) The drought of 1988 hit …
/multi>
multi size="100" docset="d50i">
P891210-0079 1 (32.20) (12/10/89) America's 1988 drought captured attention everywhere, but especially in
ashington where politicians pushed through the largest disaster relief measure in U.S. history.
P891210-0079 3 (41.18) (12/10/89) The record $3.9 billion drought relief program of 1988, hailed as
alvation for small farmers devastated by a brutal dry spell, became much more _ an unexpected, election-
ear windfall for thousands of farmers who collected millions of dollars for nature's normal quirks.
P891213-0004 1 (34.60) (12/13/89) The drought of 1988 hit …
/multi>
```

Figure 3. 50 and 100 word summaries for topic "US Drought of 1988".

Time Annotation and Sequencing

Examples

- weekdays (Sunday, Monday, etc.)
- (past | next | coming) + weekdays
- today, yesterday, last night

```
<multi size="100" docset="d45h">
AP900625-0160  1 (26.60) (06/25/90) The republic of Slovenia plans to begin work on a constitution
that will give it full sovereignty within a new Yugoslav confederation, the state Tanjug news agency
reported Monday (06/25/90).
WSJ910628-0109 3 (9.48)  (06/28/91) On Wednesday (06/26/91), the Slovene soldiers manning this border
post raised a new flag to mark Slovenia's independence from Yugoslavia.
WSJ910628-0109 5 (53.77) (06/28/91) Less than two days after Slovenia and Croatia, two of Yugoslavia's
six republics, unilaterally seceded from the nation, the federal government in Belgrade mobilized
troops to regain control.
FBIS3-30788    2 (49.14) (02/09/94) In the view of Yugoslav diplomats, the normalization of relations
between Slovenia and the Federal Republic of Yugoslavia will certainly be a strenuous and long-term
project.
</multi>
```

Figure 4. 100 word summary with explicit time annotation.

- A type of ordering - not NP-hard
- Sorting out temporal relationships
- Since the evaluation task uses news articles, publication dates allow for explicit computation of dates
- Ordering is relatively straightforward thereafter

- today, yesterday, last night

```
:multi size="100" docset="d45h">
.P900625-0160  1 (26.60)  (06/25/90) The republic of Slovenia plans to begin work on a constitution
:hat will give it full sovereignty within a new Yugoslav confederation, the state Tanjug news agency
:eported Monday (06/25/90).
ISJ910628-0109 3 (9.48)   (06/28/91) On Wednesday (06/26/91), the Slovene soldiers manning this border
post raised a new flag to mark Slovenia's independence from Yugoslavia.
ISJ910628-0109 5 (53.77)  (06/28/91) Less than two days after Slovenia and Croatia, two of Yugoslavia's
six republics, unilaterally seceded from the nation, the federal government in Belgrade mobilized
:roops to regain control.
'BIS3-30788    2 (49.14)  (02/09/94) In the view of Yugoslav diplomats, the normalization of relations
between Slovenia and the Federal Republic of Yugoslavia will certainly be a strenuous and long-term
project.
:/multi>
```

Figure 4. 100 word summary with explicit time annotation.

EVALUATION

- 50, 100, 200, 400 word summaries generated on one set of documents
- Human-written reference summaries are created
- 2 Baselines: Lead & Coverage
- Sentence is the smallest unit evaluated
- Judged on grammaticality, cohesion, & coherence
- Content inclusion grades: all, most, some, hardly any, & none

Proposed Evaluation Metrics

Usually in Single Document Summarization We Use Recall & Precision...

E.g. $Precision = \frac{N_e}{N_s}$

$$Precision = \frac{\#\ Shared\ Sentences}{\#\ Sentences\ in\ Summary}$$

...but these methods are not appropriate

- Multiple system units contribute to multiple model units
- System-Summary and Model-Summary do not exactly overlap
- Overlap judgement is non-binary

We need new metrics!

Weighted Recall (if C = 1 it is just Recall [R1])

$$Retention_w = \frac{(\#\ MUs\ Marked) \cdot C}{Total\ \#\ MUs\ in\ Model\ Summary}$$

Pseudo-Precision

$$Precision_p = \frac{\#\ SUs\ Marked}{Total\ \#\ SUs\ in\ System\ Summary}$$

- Participants in DUC were given raw data from the tests
- NIST asked for proposal metrics to "help progress the field"
- Authors propose several new methods:

Usually in Single Document Summarization We Use Recall & Precision...

E.g.

$$Precision = \frac{N_a}{N_s}$$

$$Precision = \frac{\#\ Shared\ Sentences}{\#\ Sentences\ in\ Summary}$$

...but these methods are not appropriate

- Multiple system units contribute to multiple model units
- System-Summary and Model-Summary do not exactly overlap
- Overlap judgement is non-binary

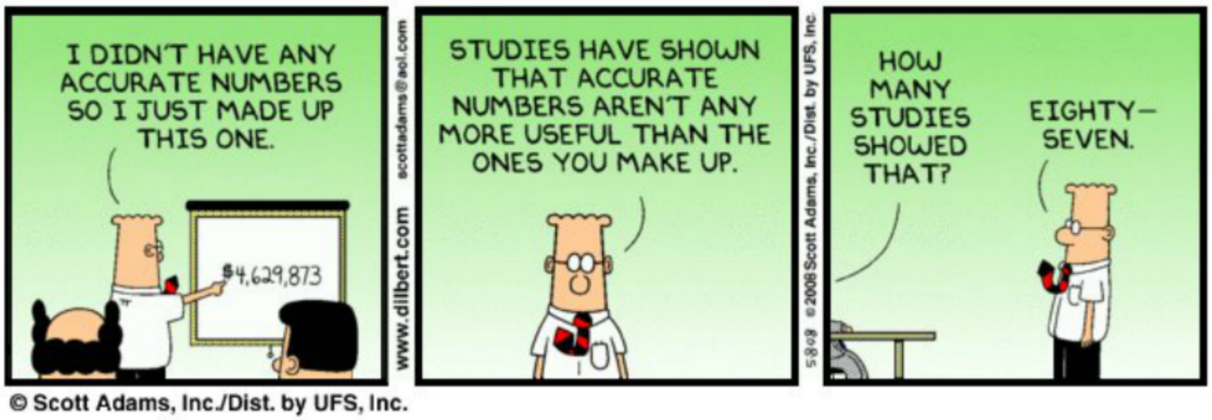
We need new metrics!

Weighted Recall (if C = 1 it is just Recall [R1])

$$Retention_w = \frac{(\#\ MUs\ Marked) \cdot C}{Total\ \#\ MUs\ in\ Model\ Summary}$$

Pseudo-Precision

$$Precision_p = \frac{\#\ SUs\ Marked}{Total\ \#\ SUs\ in\ System\ Summary}$$



- Unfortunately, these metrics are no longer widely used
- ROUGE is now standard

Results

SYS	Pp All	R1 All	Rw All	Pp 400	R1 400	Rw 400	Pp 200	R1 200	Rw 200	Pp 100	R1 100	Rw 100	Pp 50	R1 50	Rw 50
HM	58.71%	53.00%	28.81%	59.33%	52.95%	33.23%	59.91%	57.23%	33.82%	58.73%	54.67%	27.54%	56.87%	47.16%	21.62%
T	48.96%	35.53%(3)	18.48%(1)	56.51%(3)	38.50%(3)	25.12%(1)	53.85%(3)	35.62%	21.37%(1)	43.53%	32.82%(3)	14.28%(3)	41.95%	35.17%(2)	13.89%(2)
N*	58.72%(1)	37.52%(2)	17.92%(2)	61.01%(1)	41.21%(1)	23.90%(2)	63.34%(1)	38.21%(3)	21.30%(2)	58.79%(1)	36.34%(2)	16.44%(2)	51.72%(1)	34.31%(3)	10.98%(3)
Y	41.51%	41.58%(1)	17.78%(3)	49.78%	38.72%(2)	20.04%	43.63%	39.90%(1)	16.86%	34.75%	43.27%(1)	18.39%(1)	37.88%	44.43%(1)	15.55%(1)
P	49.56%	33.94%	15.78%	57.21%(2)	37.76%	22.18%(3)	51.45%	37.49%	19.40%	46.47%	31.64%	13.92%	43.10%	28.85%	9.09%
L	51.47%(3)	33.67%	15.49%	52.62%	36.34%	21.80%	53.51%	36.87%	18.34%	48.62%(3)	29.00%	12.54%	51.15%(2)	32.47%	9.90%
B2	47.27%	30.98%	14.56%	60.99%	33.51%	18.35%	49.89%	33.27%	17.72%	47.18%	29.48%	14.96%	31.03%	27.64%	8.02%
S	52.53%(2)	30.52%	12.89%	55.55%	36.83%	20.35%	58.12%(2)	38.70%(2)	19.93%(3)	49.70%(2)	26.81%	10.72%	46.43%(3)	19.23%	4.04%
M	43.39%	27.27%	11.32%	54.78%	33.81%	19.86%	45.59%	27.80%	13.27%	41.89%	23.40%	9.13%	31.30%	24.07%	5.05%
R	41.86%	27.63%	11.19%	48.63%	24.80%	12.15%	43.96%	31.28%	15.17%	38.35%	27.61%	11.46%	36.49%	26.84%	6.17%
O	43.76%	25.87%	11.19%	50.73%	27.53%	15.76%	42.94%	26.80%	13.07%	40.55%	25.13%	9.36%	40.80%	24.02%	7.03%
Z	37.98%	23.21%	8.99%	47.51%	31.17%	17.38%	46.76%	25.65%	12.83%	28.91%	17.29%	5.45%	28.74%	18.74%	3.23%
B1	32.92%	18.86%	7.45%	33.48%	17.58%	9.98%	43.13%	18.60%	8.65%	30.23%	17.42%	6.05%	24.83%	21.84%	4.20%
W	30.08%	20.38%	6.78%	38.14%	25.89%	12.10%	26.86%	21.01%	7.93%	28.31%	19.15%	5.36%	27.01%	15.46%	3.21%
U	23.88%	21.38%	6.57%	31.49%	29.76%	13.17%	24.20%	22.64%	8.49%	19.13%	17.54%	3.77%	20.69%	15.57%	3.04%

Table 1. Pseudo precision, unweighted retention, and weighted retention for all summary lengths: overall average, 400, 200, 100, and 50 words.

SYS	Grammar	Cohesion	Coherence
Human	3.74	2.74	3.19
B1	3.18	2.63	2.8
B2	3.26	1.71	1.65
L	3.72	1.83	1.9
M	3.54	2.18	2.4
N*	3.65	2	2.22
O	3.78	2.15	2.33
P	3.67	1.93	2.17
R	3.6	2.16	2.45
S	3.67	1.93	2.04
T	3.51	2.34	2.61
U	3.28	1.31	1.11
W	3.13	1.48	1.28
Y	2.45	1.73	1.77
Z	3.28	1.8	1.94

Table 2. Averaged grammaticality, cohesion, and coherence over all summary sizes.

Results

YS	Pp All	R1 All	Rw All	Pp 400	R1 400	Rw 400	Pp 200	R1 200	Rw 200	Pp 100	R1 100	Rw 100	Pp 50	R1 50	Rw 50
M	58.71%	53.00%	28.81%	59.33%	52.95%	33.23%	59.91%	57.23%	33.82%	58.73%	54.67%	27.54%	56.87%	47.16%	21.62%
T	48.96%	35.53%(3)	18.48%(1)	56.51%(3)	38.50%(3)	25.12%(1)	53.85%(3)	35.62%	21.37%(1)	43.53%	32.82%(3)	14.28%(3)	41.95%	35.17%(2)	13.89%(2)
N*	58.72%(1)	37.52%(2)	17.92%(2)	61.01%(1)	41.21%(1)	23.90%(2)	63.34%(1)	38.21%(3)	21.30%(2)	58.79%(1)	36.34%(2)	16.44%(2)	51.72%(1)	34.31%(3)	10.98%(3)
Y	41.51%	41.58%(1)	17.78%(3)	49.78%	38.72%(2)	20.04%	43.63%	39.90%(1)	16.86%	34.75%	43.27%(1)	18.39%(1)	37.88%	44.43%(1)	15.55%(1)
P	49.56%	33.94%	15.78%	57.21%(2)	37.76%	22.18%(3)	51.45%	37.49%	19.40%	46.47%	31.64%	13.92%	43.10%	28.85%	9.09%
L	51.47%(3)	33.67%	15.49%	52.62%	36.34%	21.80%	53.51%	36.87%	18.34%	48.62%(3)	29.00%	12.54%	51.15%(2)	32.47%	9.90%
B2	47.27%	30.98%	14.56%	60.99%	33.51%	18.35%	49.89%	33.27%	17.72%	47.18%	29.48%	14.96%	31.03%	27.64%	8.02%
S	52.53%(2)	30.52%	12.89%	55.55%	36.83%	20.35%	58.12%(2)	38.70%(2)	19.93%(3)	49.70%(2)	26.81%	10.72%	46.43%(3)	19.23%	4.04%
M	43.39%	27.27%	11.32%	54.78%	33.81%	19.86%	45.59%	27.80%	13.27%	41.89%	23.40%	9.13%	31.30%	24.07%	5.05%
R	41.86%	27.63%	11.19%	48.63%	24.80%	12.15%	43.96%	31.28%	15.17%	38.35%	27.61%	11.46%	36.49%	26.84%	6.17%
O	43.76%	25.87%	11.19%	50.73%	27.53%	15.76%	42.94%	26.80%	13.07%	40.55%	25.13%	9.36%	40.80%	24.02%	7.03%
Z	37.98%	23.21%	8.99%	47.51%	31.17%	17.38%	46.76%	25.65%	12.83%	28.91%	17.29%	5.45%	28.74%	18.74%	3.23%
B1	32.92%	18.86%	7.45%	33.48%	17.58%	9.98%	43.13%	18.60%	8.65%	30.23%	17.42%	6.05%	24.83%	21.84%	4.20%
W	30.08%	20.38%	6.78%	38.14%	25.89%	12.10%	26.86%	21.01%	7.93%	28.31%	19.15%	5.36%	27.01%	15.46%	3.21%
U	23.88%	21.38%	6.57%	31.49%	29.76%	13.17%	24.20%	22.64%	8.49%	19.13%	17.54%	3.77%	20.69%	15.57%	3.04%

Table 1. Pseudo precision, unweighted retention, and weighted retention for all summary lengths: overall average, 400, 200, 100, and 50 words.

SYS	Grammar	Cohesion	Coherence
Human	3.74	2.74	3.19
B1	3.18	2.63	2.8
B2	3.26	1.71	1.65
L	3.72	1.83	1.9
M	3.54	2.18	2.4
N*	3.65	2	2.22
O	3.78	2.15	2.33
P	3.67	1.93	2.17
R	3.6	2.16	2.45

SYS	Grammar	Cohesion	Coherence
Human	3.74	2.74	3.19
B1	3.18	2.63	2.8
B2	3.26	1.71	1.65
L	3.72	1.83	1.9
M	3.54	2.18	2.4
N*	3.65	2	2.22
O	3.78	2.15	2.33
P	3.67	1.93	2.17
R	3.6	2.16	2.45
S	3.67	1.93	2.04
T	3.51	2.34	2.61
U	3.28	1.31	1.11
W	3.13	1.48	1.28
Y	2.45	1.73	1.77
Z	3.28	1.8	1.94

Table 2. Averaged grammaticality, cohesion, and coherence over all summary sizes.

ROUGE

Recall-Oriented Understudy for Gisting Evaluation

ROUGE: How many reference n-grams are covered by the candidate

- Like BLEU in MT
- Uses N-Gram Overlap
- Actually a suite of metrics
- Recall measure rather than precision
- Proprietary :/

BLEU: How many candidate n-grams occurred in the reference

Where are we heading?

- RNNs for sentence ordering
- Abstractive summarization systems

Check out DEFT (Deep Exploration and Filtering of Text) for a look at some near cutting-edge proposals

tl;dr

- Extractive summarization dominates the field
- State-of-the-art systems are quite good: even the NeATS prototype was decent
- All extractive systems follow the same three steps:
  - selection
  - filtering
  - presentation
- Heuristics play a huge role in generating summaries (especially ordering)
- It's quite difficult to agree upon an evaluation metric (the ones used here are now out-of-use)
- ROUGE is now the default scoring metric
- True abstractive summaries still evade us

References

- Cao, Ziqiang, et al. "Ranking with Recursive Neural Networks and Its Application to Multi-document Summarization." (2015).
- Carbonell, Jaime, and Jade Goldstein. "The Use of MMR, Diversity-Based Reranking for Reordering Documents and Producing Summaries." (1998).
- Dunning, Ted. "Accurate methods for the statistics of surprise and coincidence." (1993).
- Edmundson, Harold P. "New methods in automatic extracting." (1969).
- Jurafsky, Daniel, and James H. Martin. "Speech and Language Processing." (2009).
- Khan, Atif, and Naomie Salim. "A Review on Abstractive Summarization Methods." (2014).
- Lin, Chin-Yew, and Eduard Hovy. "Identifying topics by position." (1997).
- Lin, Chin-Yew, and Eduard Hovy. "From Single to Multi-Document Summarization: A Prototype System and Its Evaluation." (2002).
- Lin, Chin-Yew. "Rouge: A Package for Automatic Evaluation of Summaries." (2004).
- Luhn, Hans Peter. "The automatic creation of literature abstracts." (1958).
- Zechner, Klaus, and Alex Waibel. "Minimizing word error rate in textual summaries of spoken language." (2000).

Questions?

Should you?