Automatic Summarization

COMP-550

Nov 14, 2017

Outline

Types of automatic summarization

Summarization evaluation

Single-document summarization

- Supervised machine learning
- TF*IDF
- Topic signatures

Multi-document summarization

- SumBasic

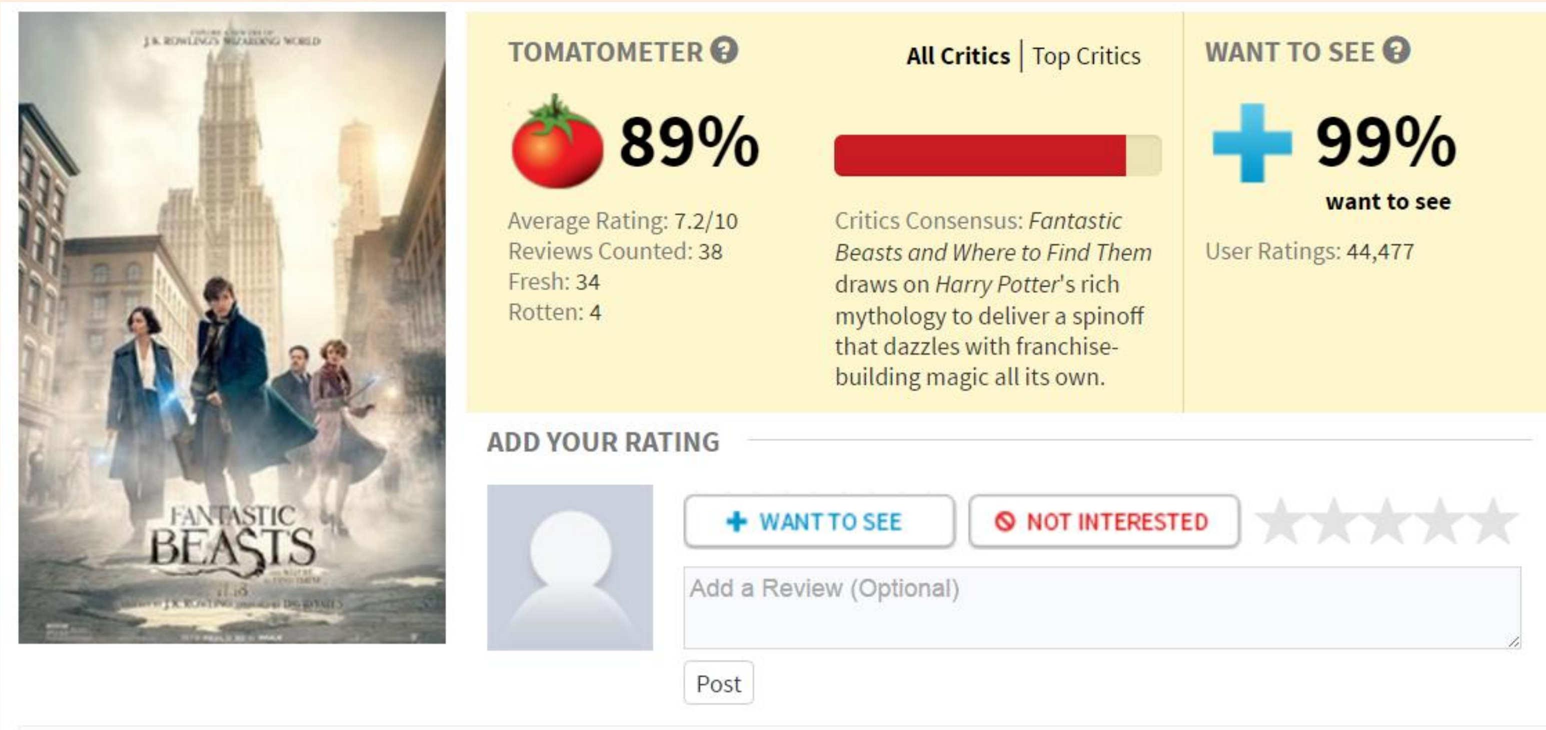
Automatic Summarization

Shortening some **source text** into a **summary**.

Trudeau affirms 'true friendship' between Canada and Cuba at meeting with President Castro

Prime Minister Justin Trudeau met with Cuban President Raul Castro at his presidential palace in Havana on Tuesday and praised the "true friendship" between the two countries. 9:36 PM ET

Another Example



Modality:

Text, graphical, speech, multi-model

We will focus on **text** summarization.

Summarization Systems: Purposes

Informative – tries to be a substitute for the source text, expressing as much of the important points as possible

Indicative – provides a link to the source text, to help users decide whether to read it or not

Critical – provides an opinion of the source text (positive or negative)

Summarization Systems: Method

Extraction – copy and extract parts of the source text

Abstraction – synthesize and produce novel text

Requires more advanced semantic analysis and NLG

Summarization Systems: Focus

Generic – no particular point of view taken; source text author's views are preserved

User-tailored or **query-focused** – summary reflects upon a specific goal or priority specified by the user

Summarization Systems: Source

Single-document

Multi-document

Additional issues to handle:

- Conflicting or contradictory information
- Redundancy between documents
- Combining information from multiple documents

Summarization Systems: Background

Level of background to assume in readers

An **update summary** is a summary written to provide an update on a situation, assuming that the reader already knows about previous related events.

Steps in Summarization

1. **Analysis / Content selection**
   - Determining what to say. What is important? Novel? Interesting? Relevant?
2. **Transformation / Refinement**
   - Aggregating common or contradictory points
   - Drawing new inferences from source text
3. **Synthesis / Surface realization**
   - Determining the final form of the summary

Steps in Extractive Summarization

Let's look at these three steps for single-document extractive summarization.

1. Analysis / Content selection
   - Determine which sentences or other text spans to select
2. Transformation / Refinement
   - Minimal amount of work needed.
3. Synthesis / Surface realization
   - Minimal amount of work needed: arranging different snippets

How does this change for multi-document summarization?

Summarization Evaluation

How do you tell if you've got a good summary?

Aspects to be rated:

Summary content

- Does it accurately reflect the original content?
- Does it include the most important content?
- Does it include non-redundant content?

Linguistic quality

- Grammaticality of the individual sentences
- Coherence of the output

Human Judgments

Ask people to rate the summary

- *From a scale of 1 to 5, how would you rate the quality of this summary?*

Advantages

- Can focus in on the different aspects of the summary
- Does not require gold standard summaries

Disadvantages

- Expensive – need to conduct for each system
- Different people have different interpretations of the scale
- Results do not generalize across different evaluation runs

ROUGE (Lin, 2004)

Compare automatic summary against human gold standard summaries

$$ROUGE_n = \frac{\sum_{S\in\{Refs\}} \sum_{ngram\in S} Count_{match}(ngram)}{\sum_{S\{Refs\}} \sum_{ngram\in S} Count(ngram)}$$

Sum over reference summaries

For each ngram in S

ngram count/match

Common choices for n: 1, 2

ROUGE Example

Let's compute ROUGE-1:

System: *We learned about evaluating summarization with ROUGE.*

Ref 1: *Extractive summarization can be evaluated using automatic methods.*

Ref 2: *ROUGE was devised to evaluate automatically generated summaries.*

Ref 3: *This class covers language generation, including summarization.*

Other Evaluation Methods

Pyramid Method (Nenkova and Passonneau, 2004)

- A structured kind of content evaluation which focuses on selecting important summary content units (SCUs).
- Requires human annotation effort.

Extrinsic evaluation

- Test if providing summaries can improve learning (e.g., by taking a quiz on the material) (McCallum et al., 2012)
- Test if summaries can improve speed of identifying relevant documents (Dorr et al., 2005)

Single-Document Summarization

View this as a supervised machine learning method

Not all factors can be easily learned in this approach

Which of the following do you think are best for supervision?

- Lexical features
  - Content words
  - Function words
- Discourse features
  - Position within document
  - Discourse cues such as *because* or *therefore*
  - Discourse structure

A Machine Learning Method

Early methods rely on position and discourse cues (Luhn, 1959; Edmundson, 1968)

Lin and Hovy (1998) trained a supervised method:

- Input: source text + human abstracts
- For each sentence in human abstract, find position in source article that has highest similarity to it.
- On computer products newspaper corpus:
  - T1 (title)
  - P2S1 (first sentence of second paragraph)
  - P3S1 (first sentence of third paragraph)
- On WSJ:
  - T1, P1S1, P1S2, …

Leading Baseline

In fact, in some genres, such as news text, the beginning of the source text acts like a summary.

Baseline method: select the first sentences of the article, up until the word length limit is reached.

Let's check with actual news articles: http://www.bbc.com/news

Term Weighting

Not all words are equally important.

What do you know about an article if it contains the word

the?

penguin?

TF*IDF (Salton, 1988)

Term Frequency Times Inverse Document Frequency

A term is important/indicative of a document if it:

1. Appears many times in the document
2. Is a relative rare word overall

TF is usually just the count of the word

IDF is a little more complicated:

$$IDF(t, Corpus) = \log \frac{\#(\text{Docs in } Corpus)}{\#(\text{Docs with term } t) + 1}$$

- Need a separate large training corpus for this

Originally designed for document retrieval

TF*IDF Example

the appears in 8000 of 8500 documents

penguin appears in 50 of 8500 documents

the appears 35 times in current article

penguin appears twice in current article

TF*IDF of *the* is

TF*IDF of *penguin* is

Topic Signatures

A method designed by Lin and Hovy (2000)

First, determine two sets of related and unrelated articles.

e.g., Summarizing about vaccinations

Related (R) : articles in health domain

Unrelated ($\neg R$): articles in the finance, education domains

For each term t_i, compute following matrix:

	R	$\neg R$
t_i	O_{11}	O_{12}
$\neg t_i$	O_{21}	O_{22}

Binomial Distributions

We will consider each *row* of the contingency table

	R	$\neg R$
t_i	O_{11}	O_{12}
$\neg t_i$	O_{21}	O_{22}

e.g., from first row, we ask: what is the probability that occurrences of t_i are distributed between R and $\neg R$ in this way? This is a **binomial distribution**.

$$b(k; n, \theta) = \binom{n}{k} \theta^k (1-\theta)^{(n-k)}$$

Competing Hypotheses

Compare the following two hypotheses:

Hypothesis 1: the term t_i is not characteristic of the domain; the distribution of occurrences of t_i between R and $\neg R$ is the same as for all other terms, $\neg t_i$

Likelihood of data given this hypothesis:

$$L(H_1) = b(O_{11}; O_{11} + O_{12}, p)b(O_{21}; O_{21} + O_{22}, p)$$

Hypothesis 2: the term t_i is important to the domain; the distribution of occurrences of t_i between R and $\neg R$ is different from the distribution for all other terms, $\neg t_i$

$$L(H_2) = b(O_{11}; O_{11} + O_{12}, p_1)b(O_{21}; O_{21} + O_{22}, p_2)$$

Likelihood Ratio

We'll compute the following likelihood ratio:

$$-2\log\lambda = -2\log\frac{L(H_1)}{L(H_2)}$$

A high value of $-2\log\lambda$ for a term indicates that the term is indicative of the domain; good to include in summary.

Rank sentences by $-2\log\lambda$ and select sentences with words that score highly on this.

Sample Rankings

Topic 10 Signature Terms of Topic 151 — C			
Unigram	$-2log\lambda$	Bigram	$-2log\lambda$
jail	461.044	county jail	160.273
county	408.821	early release	85.361
overcrowding	342.349	state prison	74.372
inmate	234.765	state prisoner	67.666
sheriff	154.440	day fine	61.465
state	151.940	jail overcrowding	61.329
prisoner	148.178	court order	60.090
prison	145.306	local jail	56.440
city	133.477	prison overcrowding	55.373
overcrowded	128.008	central facility	52.909

Topic 10 Signature Terms of Topic 257 — Ci			
Unigram	$-2log\lambda$	Bigram	$-2log\lambda$
cigarette	476.038	tobacco industry	80.768
tobacco	313.017	bn cigarette	67.429
smoking	284.198	philip morris	54.073
smoke	159.134	cigarette year	48.045
rothmans	156.675	rothmans international	44.434
osha	148.372	tobacco smoke	44.269
seita	126.421	sir patrick	40.455
ban	113.849	cigarette company	39.399
smoker	104.110	cent market	36.223
bat	79.903	tax increase	36.223

Multi-Document Summarization

Additional issues to consider:

- Conflicting or contradictory information
- Redundancy between documents
- Combining information from multiple documents

But the second point can actually work to our advantage

- If everybody is talking about the same thing, that thing is likely to be important information.

SumBasic

(Nenkova and Vanderwende, 2005)

Uses unigram frequencies with a simple update for non-redundancy.

- Step 1: Compute $\mathrm{p}(w_i) = n_i / N$
- Repeat until summary length limit reached:
  - Step 2: Rank sentences by their average word probabilities
  - Step 3: Pick best scoring sentence S^{best}; add to summary.
  - Step 4: For each word w_j in S^{best}, update

$$p^{new}(w_j) = p^{old}(w_j)^2$$

This down-weights the words that were just selected

Later Developments

More sophisticated optimization procedures:

- Rather than a greedy selection and update step, select a globally optimum set of sentences, accounting for both informativeness and non-redundancy.

Account for similarities between bigrams

Other heuristics, such as avoiding sentences with pronouns

Removing words, such as discourse cues like *therefore*, that don't make sense out of context.

Modelling coherence or flow of summary sentences.

Conroy et al., 2006

This system combines the topic signature method, a sophisticated non-redundancy module, and the following eliminations:

- Gerund clauses

  Sally went to the store, skipping on one leg.

- Restricted relative-clause appositives

  Bob, who is the president of the club, disagreed.

- Intra-sentential attribution

  They would never do that, she said, without consulting us.

- Lead adverbs

  Hopefully, we will find a solution.

Performance

This simple method (with a few other details), achieves near-human performance on ROUGE-1:

Submission	Mean	95% CI Lower	95% CI Upper
F	0.36787	0.34442	0.39467
B	0.36126	0.33387	0.38754
O (ω)	0.35810	0.34263	0.37330
H	0.33871	0.31540	0.36423
A	0.33289	0.30591	0.35759
D	0.33212	0.30805	0.35628
E	0.33277	0.30959	0.35687
C	0.30237	0.27863	0.32496
G	0.30909	0.28847	0.32987
$\omega_{qs}^{(pr)}$	0.308	0.294	0.322
peer 65	0.308	0.293	0.323
SumBasic	0.302	0.285	0.319
peer 34	0.290	0.273	0.307
peer 124	0.286	0.268	0.303
peer 102	0.285	0.267	0.302

Table 4: Average ROUGE 1 Scores with stop words removed for DUC04, Task 2

Next Class

Abstractive summarization

- Text-to-text generation
- Semantics-to-text generation

Natural language generation

Please bring your A3 reading to class! I will look over your questions and we will discuss in class.

References

- Conroy et al. 2006. Topic-Focused Multi-document Summarization Using an Approximate Oracle Score. *COLING-ACL*.
- Edmunson. 1968. New Methods in Automatic Extraction. *Journal of the ACM.* 16(2), 264-285.
- Lin and Hovy. 1998. Automated Text Summarization and the SUMMARIST System. *ACL Workshop*.
- Lin and Hovy. 2000. The automated acquisition of topic signatures for text summarization. *COLING*.
- Luhn. 1959. The Automatic Creation of Literature Abstracts. *IBM Journal of Research and Development.* 159-165.
- Nenkova and Vanderwende. 2005. The impact of frequency on summarization. *MSR Tech Report*.