CS60092: Information Retrieval

Text Summarization

Debaditya Roy

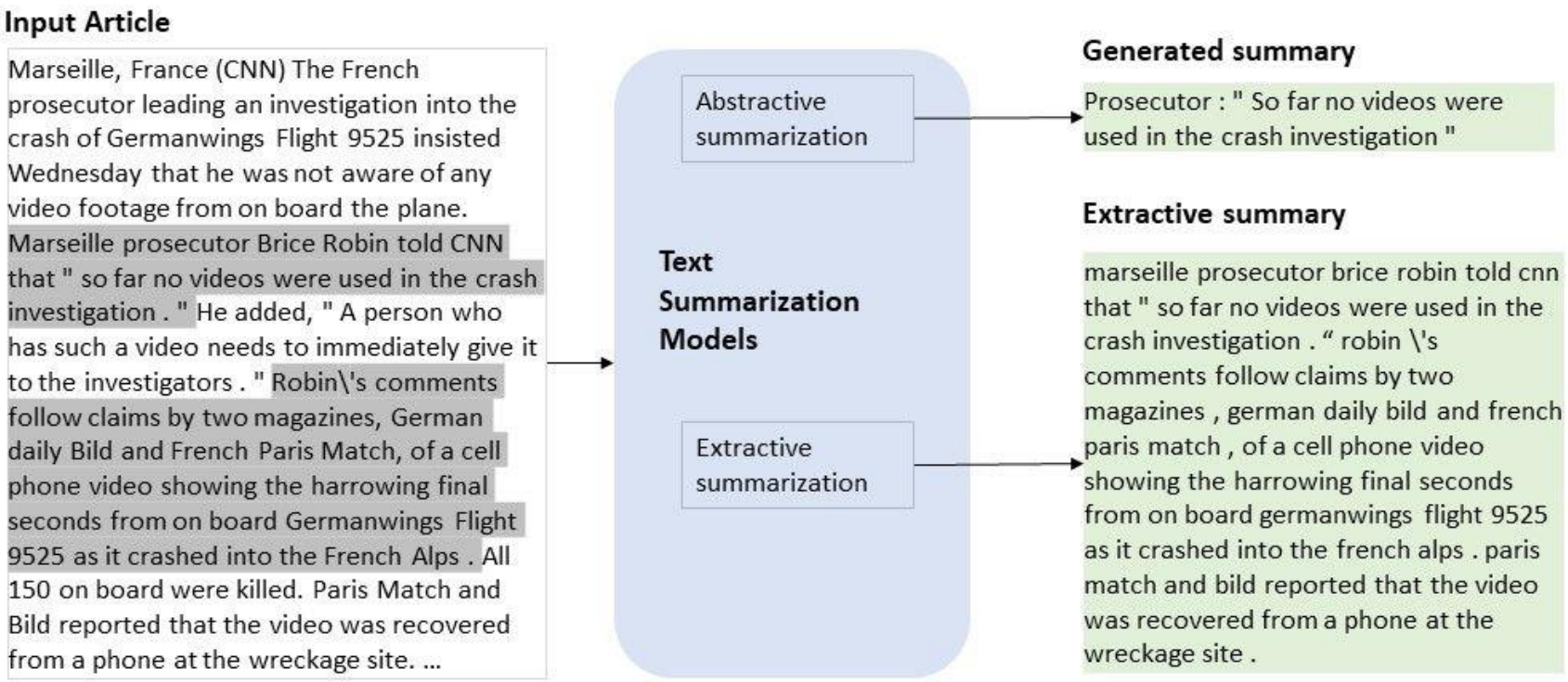
Text Summarization – The Task

- Automatic generation of a shorter text S from a source document D, such that

$$\max Salience(S, D) \;\; s.t. |S| \leq k$$

- $Salience$ ≈ information preservation
- k = length constraint

Text Summarization



Types of Summarization

Extractive

- Selects sentences directly from the source

$$S = \{s_i \in D \mid i \in I\}$$

Common methods:

- TF-IDF ranking
- TextRank (graph-based ranking)
- Supervised classifiers

Factually safe

Limited paraphrasing

Abstractive

- Generates new text using sequence-to-sequence models

$$P(S \mid D) = \prod_{t=1}^{|S|} P(w_t \mid w_{<t}, D)$$

Modern methods

- Transformer encoder–decoder models
- Large Language Models (LLMs)

Fluent, human-like

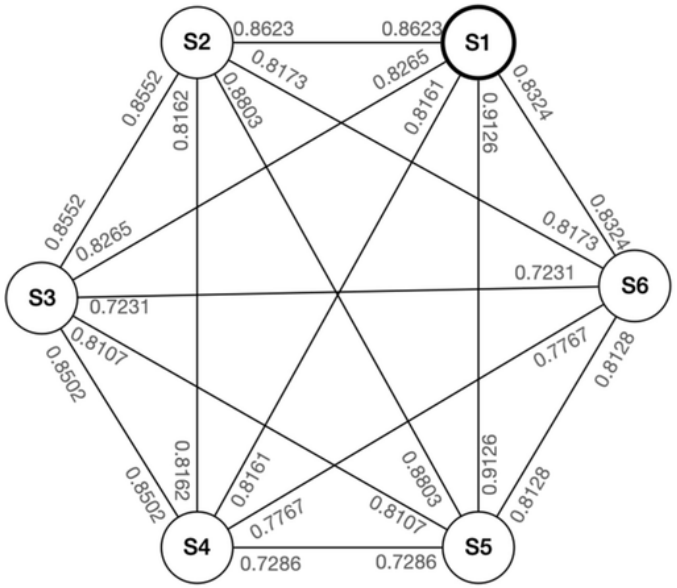
Risk of hallucination

TextRank

- Replace web pages with sentences.
- Build weighted graph

$$W_{ij} = \mathrm{sim}(s_i, s_j)$$

- Normalize rows → transition matrix.



Case IMM-2683-96, paragraph 4:

S1: The applicant alleges that the Board used standard form "boiler-plate" reasons and therefore denied the applicant a fair hearing.

S2: Key passages in the Board's reasons are identical or virtually identical to two other decisions, Jafari v. Minister of Citizenship & Immigration , Board

S3: In all three cases involving Iranian refugee claimants, Mr. Jack Davis was the presiding Board member.

S4: Jafari was heard three days after the case at bar.

S5: The respondent in turn argues that the Board clearly made an independent decision.

S6: The identical passages are nothing more than digests of the law on such legal questions as credibility and the documentary evidence

https://www.researchgate.net/figure/An-example-of-TextRank-graph-each-sentence-in-the-paragraph-is-represented-as-a-vertice_fig4_362616496

TextRank

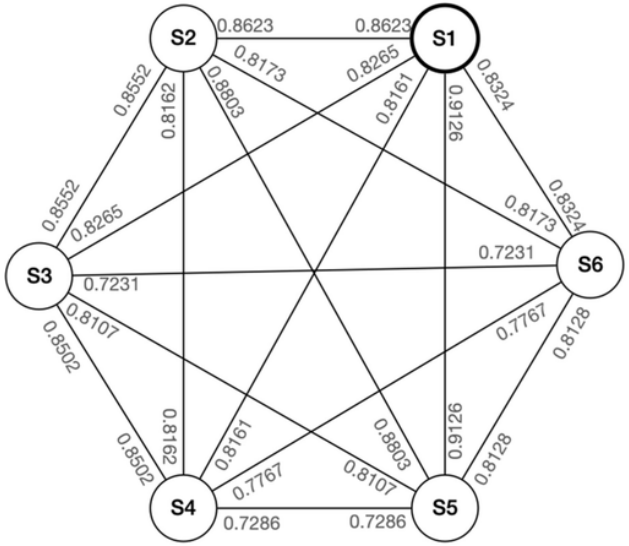
- Apply PageRank equation:

$$r = \alpha W^{\mathsf{T}} r + (1 - \alpha) v$$

- Result

$$r_i = \text{importance score of sentence } s_i$$

- $\text{Top} - k$ sentences form summary.



Case IMM-2683-96, paragraph 4:

S1: The applicant alleges that the Board used standard form "boiler-plate" reasons and therefore denied the applicant a fair hearing.

S2: Key passages in the Board's reasons are identical or virtually identical to two other decisions, Jafari v. Minister of Citizenship & Immigration , Board

S3: In all three cases involving Iranian refugee claimants, Mr. Jack Davis was the presiding Board member.

S4: Jafari was heard three days after the case at bar.

S5: The respondent in turn argues that the Board clearly made an independent decision.

S6: The identical passages are nothing more than digests of the law on such legal questions as credibility and the documentary evidence

https://www.researchgate.net/figure/An-example-of-TextRank-graph-each-sentence-in-the-paragraph-is-represented-as-a-vertice_fig4_362616496

How to evaluate summarization?

- Human evaluation is expensive and subjective
- **Need automatic metrics** to compare

DUC2001 -> d60k -> SJMN91-06106024

Reference summary

[Line-1] Rodney King spends his time seeing doctors and thinking about his injuries he fears may become permanent. ***[Line-2]*** He is staying with relatives and fears retribution by the police. ***[Line-3]*** His ex-wife says he's depressed and frightened; his attorney has hired guards to protect him. ***[Line-4]*** King suffers headaches and numbness of the face after five hours of plastic surgery to repair fractures of his cheek and eye bones, and has instituted an $83 million law suit against the city for excessive force. ***[Line-5]*** In another development, he's now a suspect in a February 21 robbery and shooting, a result of the wide publicity.

Predicted summary:

[Line-1] Six weeks after his beating by Los Angeles police and seemingly forgotten in the political turmoil that has followed -- Rodney G. King fears retribution, spends most of his time seeing doctors, and thinks a lot about the headaches, scars and facial numbness he worries might become permanent. ***[Line-2]*** Lerman has filed an $83 million claim against the city on King's behalf. King's neat, blue home in Altadena has the curtains drawn, its phone number and those of other family members long changed.

Two major metrics:

- **BLEU** (precision-oriented)
- **ROUGE** (recall-oriented)

Saini, N., Saha, S., Chakraborty, D., & Bhattacharyya, P. (2019). Extractive single document summarization using binary differential evolution: Optimization of different sentence quality measures. *PloS one*, *14*(11), e0223477.

BLEU (Bilingual Evaluation Understudy)

Originally introduced for **machine translation**, but often used for summarization.

Core Idea: N-gram Precision

Example:

- **Reference (R):** the cat sat on the mat**, Length** $r = 6$
- **Candidate (C):** the cat is on mat**, Length**: $c = 5$

Candidate Unigrams

Word	Count(C)
the	1
cat	1
is	1
on	1
mat	1

Reference Unigrams

Word	Count(C)
the	2
cat	1
sat	1
on	1
mat	1

BLEU

Unigram matches

- the → 1
- cat → 1
- on → 1
- mat → 1
- is → 0

Reference (R): the cat sat on the mat
Candidate (C): the cat is on mat

- Total matches = 4
- Total candidate unigrams = 5

$$P_1 = \frac{4}{5} = 0.8$$

Bigram matches

Candidate bigrams

- *the cat*
- *cat is*
- *is on*
- *on mat*

Total = 4

Reference bigrams

- *the cat*
- *cat sat*
- *sat on*
- *on the*
- *the mat*

Overlap

the cat

$$P_2 = \frac{1}{4} = 0.25$$

Final BLEU

- Geometric Mean of P_1 and P_2 with equal weights

$$GM = \exp\left(\frac{1}{2}\log 0.8 + \frac{1}{2}\log 0.25\right) \approx 0.447$$

- Brevity Penalty

$$BP = e^{1-\frac{r}{c}} = e^{1-\frac{6}{5}} \approx 0.819$$

- Final BLEU = $BP \cdot GM \approx 0.366$

Low BLEU because

- Bigram precision is poor
- Candidate is shorter than reference

BLEU – strengths and weaknesses

✔ Correlates with translation quality
✔ Penalizes overly short outputs
✔ Robust for long texts

✗ Precision-only (ignores recall)
✗ Bad for short summaries
✗ Cannot detect semantic similarity
✗ Sensitive to wording

ROUGE (Recall-Oriented Understudy for Gisting Evaluation)

- Designed **specifically for summarization**

Example

- **Reference (R):** *the cat sat on the mat*, **Length** $r = 6$
- **Candidate (C):** *the cat is on mat*, **Length**: $c = 5$

ROUGE-1 (Unigram Recall)

- Matches = 4
- Total reference unigrams = 6

$$ROUGE - 1 = \frac{4}{6} = 0.667$$

ROUGE-2

ROUGE-2 (Bigram Recall)

- Matches = 1
- Total reference unigrams = 5

$$ROUGE-2 = \frac{1}{5} = 0.2$$

- ROUGE measures: $\frac{\text{overlap}}{\text{reference size}}$
- BLEU measures: $\frac{\text{overlap}}{\text{candidate size}}$

That's precision vs recall

ROUGE-L (Longest Common Subsequence)

Reference (R): *the cat sat on the mat*
Candidate (C): *the cat is on mat*

- **LCS**: *the cat on mat*, **Length** = 4
- Compute Precision and Recall

$$R_{LCS} = \frac{LCS}{r} = \frac{4}{6} = 0.667$$

$$P_{LCS} = \frac{LCS}{c} = \frac{4}{5} = 0.8$$

F-score

- Standard formulation: $F_{LCS} = \frac{(1+\beta^2)RP}{R+\beta^2 P}$
- Usually $\beta = 1$

$$F = \frac{2 \cdot 0.667 \cdot 0.8}{0.667 + 0.8} \approx 0.727$$

ROUGE-S

Reference (R): the cat sat on the mat
Candidate (C): the cat is on mat

- Counts all ordered pairs
- A skip-bigram is $(w_i, w_j), i < j$ with gaps allowed

Reference Skip-Bigrams $(\mathbf{15})$

(the, cat), (the, sat), (cat, on),(sat, mat), …

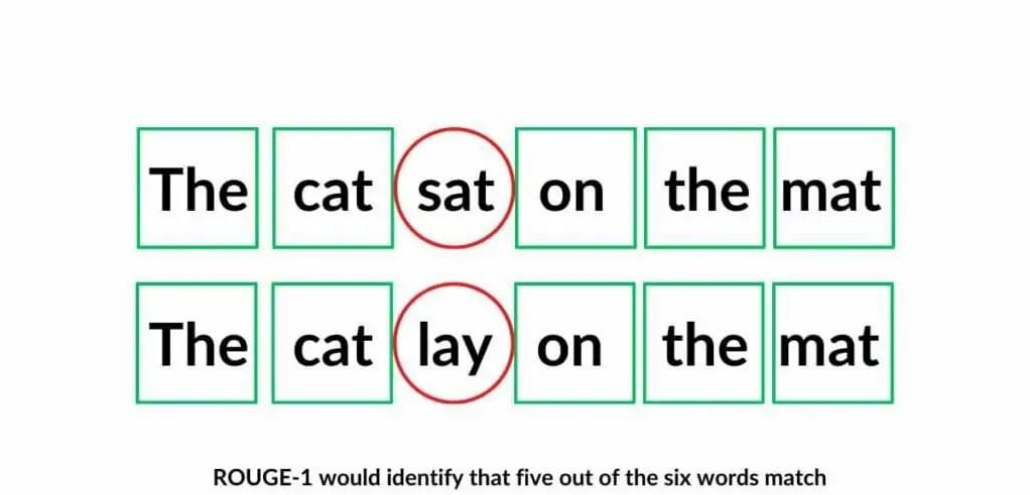
Matching Skip-Bigrams $(\mathbf{6})$

(the, cat),(the, on),(the, mat),(cat, on),(cat, mat),(on, mat)

$$ROUGE - S = \frac{6}{15} = 0.4$$

ROUGE

- Sentence-level structure
- Word order (loosely)
- Good for fluency
- Still surface-based



https://spotintelligence.com/2024/08/12/rouge-metric-in-nlp/

BLEU and ROUGE: Limitations

Example

Reference: The government passed a climate bill to reduce emissions.

Candidate A: The government passed a climate law.

Candidate B: A bill was approved to reduce emissions.

- **BLEU** may penalize both for wording differences
- **ROUGE** captures overlap but still misses semantic similarity like “law” ≈ “bill”

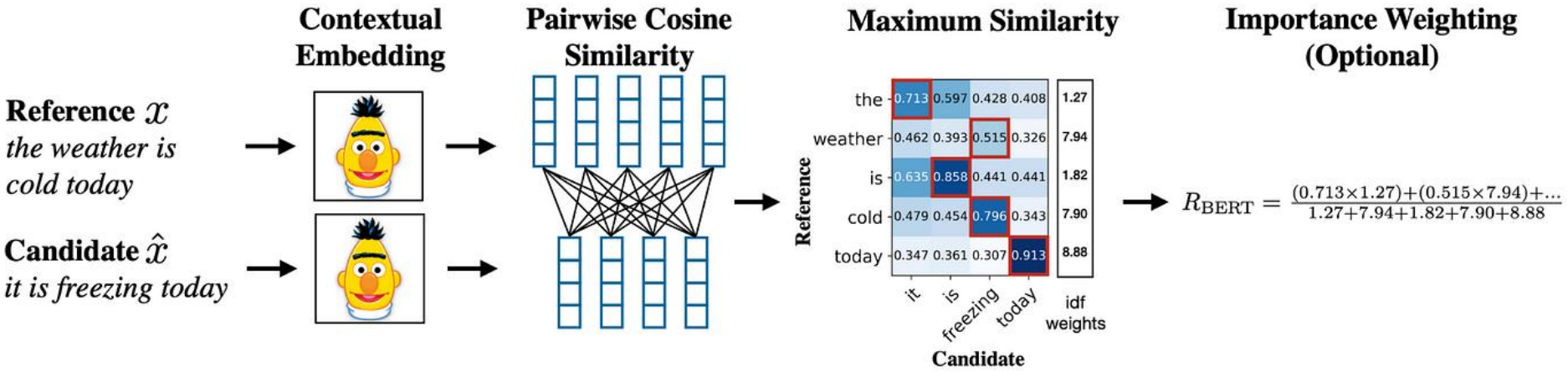
Modern Improvements - **BERTScore**

- **Semantic similarity metric** for evaluating generated text (summaries, translations, captions)

$$\text{Sentence} \rightarrow \text{tokens} \rightarrow \text{embeddings}$$

- **Contextual embeddings** from BERT

$$cat \rightarrow e_{cat} = [0.12, -0.55, 0.33, \dots]$$
$$feline \rightarrow e_{feline} = [0.10, -0.52, 0.35, \dots]$$
$$\text{sim}(cat, feline) = \cos(e_{cat}, e_{feline})$$



Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. BERTScore: Evaluating Text Generation with BERT. In *International Conference on Learning Representations*.

BERTScore

- Candidate sentence $(c_1, c_2, \dots, c_m)$, Reference sentence $(r_1, r_2, \dots, r_n)$
- Each token has an embedding
- Similarity $\boldsymbol{s_{ij} = cos(c_i, r_j)}$
- **Precision:** Each candidate token finds the **best matching reference token**

$$P = \frac{1}{m}\sum_{i=1}^{m} \max_j s_{ij}$$

- **Recall:** Each reference token finds the **best matching candidate token**.

$$R = \frac{1}{n}\sum_{j=1}^{n} \max_i s_{ij}$$

$$\boxed{BERTScore(F_1) = \frac{2PR}{P+R}}$$

- **IDF Weighting:** Important words should matter more

$$P = \frac{\sum_i IDF(c_i) \max_j s_{ij}}{\sum IDF(c_i)}$$

LLM-as-a-judge

```
Document: <source text>

Generated summary: <model
output>

Reference summary: <human
summary>

Evaluate the generated
summary on:
1. Faithfulness
2. Coverage
3. Fluency

Give a score from 1–5 and
explanation.
```

Is the output consistent with the source?

Does the output answer the question or summarize the main ideas?

Is the text grammatically correct and readable?

References

Speech and Language Processing
Authors: Daniel Jurafsky & James H. Martin

Topic	Chapter
NLP evaluation metrics	Chapter 13 – Evaluation
Machine translation metrics (BLEU)	Chapter 13
Text summarization	Chapter 23