Summarization: Evaluation

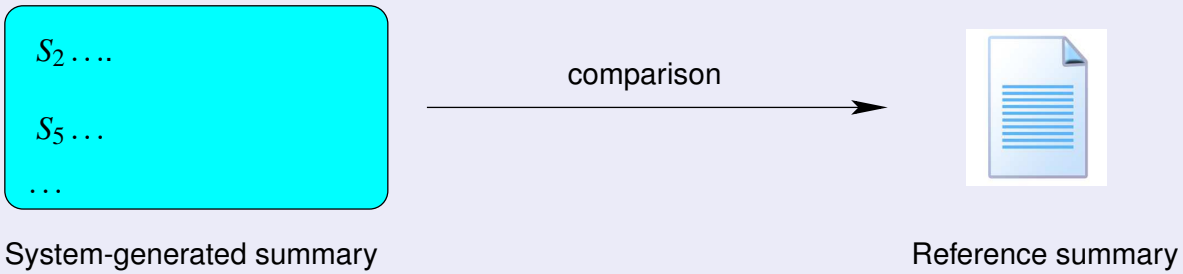
Pawan Goyal

CSE, IIT Kharagpur

Week 11, Lecture 3

System Evaluation

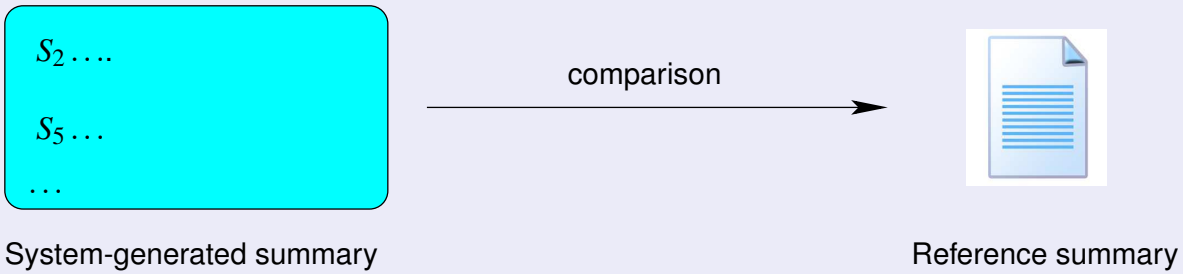
Evaluation Criteria



System Evaluation

System Evaluation

Evaluation Criteria



System Evaluation

ROUGE

Recall Oriented Understudy for Gisting Evaluation *Not as good as human evaluation but much more convenient*

Toolkit available for download.

ROUGE for evaluation

Given a document D, and an automatic summary X:

- Have N humans produce a set of reference summaries of D ($N \geq 1$)
- Run system, giving automatic summary X
- What percentage of the n-grams from the reference summaries appear in X?

$$ROUGE-2 = \frac{\sum_{S \in \{RefSums\}} \sum_{bi-gram \in S} Count_{match}(bi-gram)}{\sum_{S \in \{RefSums\}} \sum_{bi-gram \in S} Count(bi-gram)}$$

ROUGE Example

Reference Summaries

- **Human 1:** water spinach is a green leafy vegetable grown in the tropics.
- **Human 2:** water spinach is a semi-aquatic tropical plant grown as a vegetable.
- **Human 3:** water spinach is a commonly eaten leaf vegetable of Asia

System Summary

water spinach is a leaf vegetable commonly eaten in tropical areas of Asia.

ROUGE Example

Reference Summaries

- **Human 1:** water spinach is a green leafy vegetable grown in the tropics.
- **Human 2:** water spinach is a semi-aquatic tropical plant grown as a vegetable.
- **Human 3:** water spinach is a commonly eaten leaf vegetable of Asia

System Summary

water spinach is a leaf vegetable commonly eaten in tropical areas of Asia.

ROUGE-2

$$\frac{3+3+6}{10+10+9} = 12/29 = 0.413$$

Further Discussions

Further Discussions

- Multi-document summarization

Further Discussions

- Multi-document summarization
- Query-specific summarization

Further Discussions

- Multi-document summarization
- Query-specific summarization
- Abstractive summarization