Quality metrics in AI-Models' output: ROUGE

Saeed Siddik
Assistant Professor
IIT University of Dhaka

ROUGE (*Recall-Oriented Understudy for Gisting Evaluation*)

- A metric used to evaluate automatic summarization and machine translation by comparing system-generated text to human-written reference text.

- Common in machine translation and code generation tasks.
- Higher ROUGE→ closer match to reference output.
- Syntactic similarity

Example: ROUGR Score Calculation

- **Reference sentence (human output):**
- “The project manager approved the software release.”
- **Candidate sentence (AI-generated):**
- “The manager approved the release of the software.”

Step 1: Compute n-gram overlaps

- Unigram (1-word)
- Bigram (2-words)
- n-gram (n-words)

Output length

Type	Sentence (Tokenized)	Length (c) or (r)
Reference (R)	The project manager approved the software release.	r=7
Candidate (C)	The manager approved the release of the software.	c=8

Compute Unigram (1-word)

- Reference : the, project, manager, approved, the, software, release
- Candidate : the, manager, approved, the, release, of, the, software
- overlap tokens = the (min 2,3)=2, manager=1, approved=1, release=1, software=1
- total overlap = 6.
- Precision: overlap / |cand| = 6 / 8 = 0.75.
- Recall = overlap / |ref| = 6 / 7 ≈ 0.857.
- F1 = 2·P·R / (P+R) = 2×075×0.857/(075+0.857)
- **ROUGE-1 F1 ≈ 0.80**

Compute bigram

2-gram	Candidate Sentence (C) Count	Reference Sentence (R) Count	Clipped Count (Count clip)
The manager	1	0	0
manager approved	1	1	**1**
approved the	1	1	**1**
the release	1	0	0
release of	1	0	0
of the	1	0	0
the software	1	1	**1**
Total	**7**		**3**

Compute Bigram (2-word)

- Reference 2-grams: the project, project manager, manager approved, approved the, the software, software release
- Candidate 2-grams: the manager, manager approved, approved the, the release, release of, of the, the software
- Overlapping bigrams: (manager approved), (approved the), (the software) → 3 overlaps.
  Precision = 3 / 7 ≈ 0.4286.
  Recall = 3 / 6 = 0.5.
  F1 = 2·P·R / (P+R) = 2×0.4286×0.5 / (0.428+0.5)
- **ROUGE-2 F1 ≈ 0.462**

Class Task on Computing ROUGE Score

- Reference summary (human-written):

“**Sorts a list of numbers in ascending order.**”

- Candidate summary (AI-generated):

“**Sorts numbers in increasing order.**”

BLEU vs ROUGE

Feature	BLEU	ROUGE
Main Focus	Precision	Recall
What it measures	How much of the generated text is present in the reference text(s).	How much of the reference text(s) is captured by the generated text.
Goal	To assess the fluency and adequacy of the generated text.	To assess the content coverage how much key information from the reference is included.
Formula's Denominator	Primarily based on the length/count of n-grams in the candidate (generated) text.	Primarily based on the length/count of n-grams in the reference text.

End of ROUGE Score