**ROUGE-N**

**ROUGE-N** measures the number of matching [n-grams](https://en.wikipedia.org/wiki/N-gram) between the model-generated text and a human-produced reference.

Consider the reference *R* and the candidate summary *C*:

* *R*: The cat is on the mat.
* *C*: The cat and the dog.

**ROUGE-1**

Using *R* and *C*, we are going to compute the [precision, recall, and F1-score](https://blog.exsilio.com/all/accuracy-precision-recall-f1-score-interpretation-of-performance-measures) of the matching n-grams. Let’s start computing ROUGE-1 by considering 1-grams only.

ROUGE-1 precision can be computed as the ratio of the number of unigrams in *C* that appear also in *R* (that are the words “the”, “cat”, and “the”), over the number of unigrams in *C*.

ROUGE-1 precision = 3/5 = 0.6

ROUGE-1 recall can be computed as the ratio of the number of unigrams in *R* that appear also in *C* (that are the words “the”, “cat”, and “the”), over the number of unigrams in *R*.

ROUGE-1 recall = 3/6 = 0.5

Then, ROUGE-1 F1-score can be directly obtained from the ROUGE-1 precision and recall using the standard F1-score formula.

ROUGE-1 F1-score = 2 \* (precision \* recall) / (precision + recall) = 0.54

**ROUGE-2**

Let’s try computing the ROUGE-2 considering 2-grams.

Remember our reference *R* and candidate summary *C*:

* *R*: The cat is on the mat.
* *C*: The cat and the dog.

ROUGE-2 precision is the ratio of the number of 2-grams in *C* that appear also in *R* (only the 2-gram “the cat”), over the number of 2-grams in *C*.

ROUGE-2 precision = 1/4 = 0.25

ROUGE-1 recall is the ratio of the number of 2-grams in *R* that appear also in *C* (only the 2-gram “the cat”), over the number of 2-grams in *R*.

ROUGE-2 recall = 1/5 = 0.20

Therefore, the F1-score is:

ROUGE-2 F1-score = 2 \* (precision \* recall) / (precision + recall) = 0.22

**ROUGE-L**

**ROUGE-L** is based on the [longest common subsequence (LCS)](https://en.wikipedia.org/wiki/Longest_common_subsequence_problem) between our model output and reference, i.e. the longest sequence of words (not necessarily consecutive, but still in order) that is shared between both. A longer shared sequence should indicate more similarity between the two sequences.

We can compute ROUGE-L recall, precision, and F1-score just like we did with ROUGE-N, but this time we replace each n-gram match with the LCS.

Remember our reference *R* and candidate summary *C*:

* *R*: The cat is on the mat.
* *C*: The cat and the dog.

The LCS is the 3-gram “the cat the” (remember that the words are not necessarily consecutive), which appears in both *R* and *C*.

ROUGE-L precision is the ratio of the length of the LCS, over the number of unigrams in *C*.

ROUGE-L precision = 3/5 = 0.6

ROUGE-L precision is the ratio of the length of the LCS, over the number of unigrams in *R*.

ROUGE-L recall = 3/6 = 0.5

Therefore, the F1-score is:

ROUGE-L F1-score = 2 \* (precision \* recall) / (precision + recall) = 0.55

**ROUGE-S**

**ROUGE-S** allows us to add a degree of leniency to the n-gram matching performed with ROUGE-N and ROUGE-L. ROUGE-S is a skip-gram concurrence metric: this allows to search for consecutive words from the reference text that appear in the model output but are separated by one-or-more other words.

Consider the new reference *R* and candidate summary *C*:

* *R*: The cat is on the mat.
* *C*: The gray cat and the dog.

If we consider the 2-gram “the cat”, the ROUGE-2 metric would match it only if it appears in *C* exactly, but this is not the case since *C* contains “the gray cat”. However, using ROUGE-S with unigram skipping, “the cat” would match “the gray cat” too.

We can compute ROUGE-S precision, recall, and F1-score in the same way as the other ROUGE metrics.

**Pros and Cons of ROUGE**

This is the tradeoff to take into account when using ROUGE.

* *Pros*: it correlates positively with human evaluation, it’s inexpensive to compute and language-independent.
* *Cons*: ROUGE does not manage different words that have the same meaning, as it measures syntactical matches rather than semantics.

In general:

* ROUGE focuses on recall: how much the words (and/or n-grams) in the human references appear in the candidate model outputs.

These results are complementing, as is often the case in the precision-recall tradeoff.