Precision and Recall:

$$\boxed{P = \frac{m}{w_c}} \qquad \boxed{R = \frac{m}{w_R}}$$

→ m = # of n-grams in candidates also found in reference.

→ w_c = # of n-grams in candidate.

→ w_r = # of n-grams in reference.

Ex:

R: I work on machine learning.

C_A: I work

C_B: He works on machine learning.

$$P_A = \frac{2}{2} = 100\% \qquad P_B = \frac{3}{5} = 60\%$$

$$R_A = \frac{2}{5} = 40\% \qquad R_B = \frac{3}{5} = 60\%$$

→ But the precision and recall are not always proper measure.

BLEU:

→ Bilingual Evaluation Understudy.

→ $\boxed{\text{BLEU} = BP * \exp(\Sigma P_n)}$

→ BP = Brevity penalty

→ this adjusts the score when candidate is shorter than reference.

→ $BP = \begin{cases} 1 & \text{if } c > r \\ \exp(1 - \frac{c}{r}) & \text{otherwise} \end{cases}$

→ P_n = n-gram precision

🟢 → easy to calculate and interpret

🔴 → heavily relies on n-grams which may not capture overall meaning or fluency.

🔴 → always penalizes when transalation (candidate) is longer than reference.

ROUGE:

→ Recall oriented understudy for gisting evaluation.

→ mostly used for text summarization

→ based on Recall.

→ ROUGE = Σ (Recall of n-grams)

→ there are multiple measure-

1. ROUGE-N
2. ROUGE-L
3. ROUGE-S

2. ROUGE-L:

→ based on longest common subsequence.

→ $P = \frac{LCS(A,B)}{m}$ $R = \frac{LCS(A,B)}{n}$

m = candidate length
n = reference length
A = candidate
B = reference

then weighted harmonic mean (F_1):

$$\boxed{F_1 = \frac{(1+b^2)RP}{R^2+bP}}$$

- ROUGE-W: weighted LCS
- ROUGE-S: skipgram allowed.

METEOR:

→ having better correlation with human judgement.

→ $P = \frac{m}{w_c}$ $R = \frac{m}{w_R}$

$$F_{mean} = \frac{10PR}{R + 9P}$$

chunk penalty:

$$\boxed{p = \gamma \cdot \left(\frac{c}{u_m}\right)^{\beta}}$$

- $0 < \gamma \leq 1$
-

R: the cat sat on the mat

C: on the mat sat the cat

chunk: consecutive set of words in the candidate matching with the consecutive set of words in reference.

c = # of chunks in the candidate

u_m = length of the candidate.

here, $u_m = 6$

$$\begin{tikzcd} & \text{the cat} \\ c = 3 \arrow[ur] \arrow[r] \arrow[dr] & \text{the mat} \\ & \text{on the} \end{tikzcd}$$

$\gamma = 0.5$

$\beta = 3$

- $\text{METEOR} = F_{mean}(1-p)$