Overlap Metrics: Mathematical Definitions

1 ROUGE-1

ROUGE-1 measures unigram (single word) overlap between a candidate response C and reference text R.

$$\text{ROUGE-1 Recall} = \frac{|\text{unigrams}(R) \cap \text{unigrams}(C)|}{|\text{unigrams}(R)|} \tag{1}$$

$$\text{ROUGE-1 Precision} = \frac{|\text{unigrams}(R) \cap \text{unigrams}(C)|}{|\text{unigrams}(C)|} \tag{2}$$

$$\text{ROUGE-1 F1} = \frac{2 \cdot P \cdot R}{P + R} \tag{3}$$

2 ROUGE-2

ROUGE-2 measures bigram (two consecutive words) overlap.

$$\text{ROUGE-2 Recall} = \frac{|\text{bigrams}(R) \cap \text{bigrams}(C)|}{|\text{bigrams}(R)|} \tag{4}$$

$$\text{ROUGE-2 Precision} = \frac{|\text{bigrams}(R) \cap \text{bigrams}(C)|}{|\text{bigrams}(C)|} \tag{5}$$

$$\text{ROUGE-2 F1} = \frac{2 \cdot P \cdot R}{P + R} \tag{6}$$

3 ROUGE-L

ROUGE-L uses the Longest Common Subsequence (LCS) to measure similarity, capturing sentence-level structure.

$$R_{\text{lcs}} = \frac{\text{LCS}(R, C)}{|R|} \tag{7}$$

$$P_{\text{lcs}} = \frac{\text{LCS}(R, C)}{|C|} \tag{8}$$

$$\text{ROUGE-L} = F_{\text{lcs}} = \frac{(1 + \beta^2) \cdot R_{\text{lcs}} \cdot P_{\text{lcs}}}{R_{\text{lcs}} + \beta^2 \cdot P_{\text{lcs}}} \tag{9}$$

where β controls the relative importance of recall vs. precision (typically $\beta = 1.2$).

4 BLEU

BLEU (Bilingual Evaluation Understudy) is a precision-based metric with a brevity penalty.

Modified n-gram precision:

$$p_n = \frac{\sum_{g \in \text{n-grams}(C)} \min\bigl(\text{count}_C(g), \max_{R_j} \text{count}_{R_j}(g)\bigr)}{\sum_{g \in \text{n-grams}(C)} \text{count}_C(g)} \tag{10}$$

Brevity Penalty:

$$\text{BP} = \begin{cases} 1 & \text{if } |C| > |R| \\ \exp\left(1 - \frac{|R|}{|C|}\right) & \text{if } |C| \leq |R| \end{cases} \tag{11}$$

BLEU Score:

$$\text{BLEU} = \text{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \tag{12}$$

where $w_n = \frac{1}{N}$ (uniform weights) and typically $N = 4$.

5 F1 Score

Token-level F1 score measuring the harmonic mean of precision and recall.

$$\text{Precision} = \frac{|T_C \cap T_R|}{|T_C|} \tag{13}$$

$$\text{Recall} = \frac{|T_C \cap T_R|}{|T_R|} \tag{14}$$

$$\text{F1} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{15}$$

where T_C and T_R are the sets of tokens in the candidate and reference texts, respectively.