Explanation - Automatic Evaluation Metrics Used

1. BLEU (Bilingual Evaluation Understudy)

- **Purpose:** Measures how many n-grams (sequences of words) in the machine translation also appear in the human reference translation.
- **Formula:**

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

Where:

- p_n= precision for n-grams (1-gram, 2-gram, …, N-gram)
- w_n= weight (usually equal, e.g. 0.25 each for up to 4-grams)
- BP= brevity penalty (penalizes translations that are too short)

$$BP = \begin{cases} 1 & \text{if } c >; r \\ e^{1-r/c} & \text{if } c \leq r \end{cases}$$

where (c) = candidate length, (r) = reference length.

BLEU is good at measuring **literal overlap**, but it misses synonyms and meaning equivalence.

2. chrF++ (Character n-gram F-score)

- **Purpose:** Works at the **character level**, making it better for morphologically rich or agglutinative languages (like iTaukei).
- **Formula:**

$$\text{chrF++} = (1 + \beta^2) \cdot \frac{\text{Precision} \cdot \text{Recall}}{\beta^2 \cdot \text{Precision} + \text{Recall}}$$

Where:

- Precision = proportion of system n-grams also found in the reference.
- Recall = proportion of reference n-grams found in the system output.
- β (usually 2) gives more weight to recall.

chrF++ balances precision and recall at the character level, which helps catch partial matches.

3. TER (Translation Edit Rate)

- **Purpose:** Counts how many edits are needed to turn the machine translation into the reference translation.
- **Formula:**

$$\text{TER} = \frac{\text{Number of edits}}{\text{Average reference length}}$$

Edits include insertions, deletions, substitutions, and shifts.

Lower TER means better translation (fewer edits). It directly measures post-editing effort but can penalize harmless word order differences.

4. Levenshtein Similarity (Edit Distance Ratio)

- **Purpose:** Measures how similar two sequences (system vs reference) are based on edit operations.
- **Formula:**

$$\text{Levenshtein Ratio} = 1 - \frac{D(s,r)}{\max(|s|,|r|)}$$

Where:

- $D(s,r)$= Levenshtein edit distance (minimum number of insertions, deletions, substitutions).
- $|s|, |r|$ = lengths of system and reference strings.

A higher ratio = more similar. Unlike TER, this is normalized to [0,1].

5. COMET (Crosslingual Optimized Metric for Evaluation of Translation)

- **Purpose:** A **neural network–based metric** that uses multilingual embeddings to compare translation, reference, and source sentence meaning.
- **Formula (conceptual):**

$$\text{COMET}(src, mt, ref) = f_\theta(\text{Enc}(src),\text{Enc}(mt),\text{Enc}(ref))$$

Where:

- Enc = neural encoder (e.g., XLM-R) that produces embeddings.
- f_θ= trained regression model predicting human judgment scores.

Unlike BLEU or chrF, COMET captures semantic similarity, not just surface form.

Summary

- **BLEU**: Word overlap (precise but surface-level).
- **chrF++**: Character overlap (better for morphologically rich languages).
- **TER**: How many edits needed (lower is better).
- **Levenshtein Ratio**: Normalized string similarity (higher is better).
- **COMET**: AI-based semantic similarity (best at matching human judgment).

Comparison of Automatic MT Evaluation Metrics

Metric	What it measures	Strengths	Weaknesses
BLEU	Word n-gram overlap with reference	Simple, widely used; good for surface ove	Insensitive to synonyms, word order, meani
chrF++	Character-level n-gram overlap	Handles morphology; better for agglutinat	Still surface-based; may over-penalize varia
TER	Number of edits to match reference	Directly linked to post-editing effort	Penalizes harmless reorderings; not semant
Levenshtein Ratio	Normalized string similarity	Normalized score (0-1); intuitive similarity	Too simplistic; ignores deeper meaning
COMET	Semantic similarity via neural embedding	Captures meaning, aligns with human jud	Requires trained model; resource intensive