BLEU Score Evaluation Metric

🔹 What is BLEU Score?

BLEU (Bilingual Evaluation Understudy) is a quantitative metric to evaluate how close a machine-generated sentence is to a human reference translation. It’s one of the most widely used metrics in machine translation and text generation tasks.

🔹 Two Key Components of BLEU

1. N-gram Precision
Measures how many n-grams (sequences of n words) in the generated output match with the reference translation. It is computed for multiple n-gram levels (unigrams, bigrams, trigrams, etc.).

2. Brevity Penalty (BP)
Prevents models from cheating by generating shorter sentences to get high precision.

🔹 BLEU Score Formula

BLEU = BP × exp(∑ w⍰ × log(P⍰))

Where:
- P⍰ = modified precision for n-grams
- w⍰ = weight for n-gram (e.g., 0.25 for uniform 1–4 grams)
- BP = 1 if c > r, else exp(1 - r/c)
  where c = candidate length, r = reference length

🔹 Example

Reference Sentence: "a cat is sitting on the mat"

Candidate Sentence: "cat is sitting on mat"

🔹 Step-by-step Calculation

1. Unigram Precision

Reference unigrams: ["a", "cat", "is", "sitting", "on", "the", "mat"]

Candidate unigrams: ["cat", "is", "sitting", "on", "mat"]

Matched: 5 / 5 = 1.0

Precision P_1 = 1.0

2. Bigram Precision

Reference bigrams: ["a cat", "cat is", "is sitting", "sitting on", "on the", "the mat"]

Candidate bigrams: ["cat is", "is sitting", "sitting on", "on mat"]

Matched: 3 / 4 = 0.75

Precision P_2 = 0.75

3. Trigram Precision

Reference trigrams: ["a cat is", "cat is sitting", "is sitting on", "sitting on the", "on the mat"]

Candidate trigrams: ["cat is sitting", "is sitting on", "sitting on mat"]

Matched: 2 / 3 ≈ 0.6667

Precision P_3 ≈ 0.6667

4. Brevity Penalty

c = 5 (candidate length), r = 7 (reference length)

BP = exp(1 - r/c) = exp(1 - 7/5) = exp(-0.4) ≈ 0.6703

5. Final BLEU Score

Using n = 3 (up to trigrams), uniform weights w_n = 1/3

BLEU = 0.6703 × exp((1/3) × (log(1.0) + log(0.75) + log(0.6667)))

= 0.6703 × exp(-0.2311) = 0.6703 × 0.7938 ≈ 0.532

🔹 Python Libraries to Compute BLEU Score

- nltk.translate.bleu_score
- sacrebleu
- evaluate (from HuggingFace)

🔹 Applications of BLEU

- Machine Translation
- Text Summarization
- Dialogue Generation
- Image Captioning

Summary

BLEU helps benchmark your model’s linguistic precision, ensuring the generated sentences are not only accurate but also contextually complete.