Quality metrics in AI-Models' output: BLEU Score

Saeed Siddik

Assistant Professor

IIT University of Dhaka

BLEU (Bilingual Evaluation Understudy)

- Measures precision : how much generated text overlaps with reference text.
- Common in machine translation and code generation tasks.
- Higher BLEU → closer match to reference output.
- Syntactic similarity

Example: BLEU Score Calculation

- **Reference sentence (human output):**
- “The project manager approved the software release.”
- **Candidate sentence (AI-generated):**
- “The manager approved the release of the software.″

Step 1: Compute n-gram overlaps

- Unigram (1-word)
- Bigram (2-words)
- n-gram (n-words)

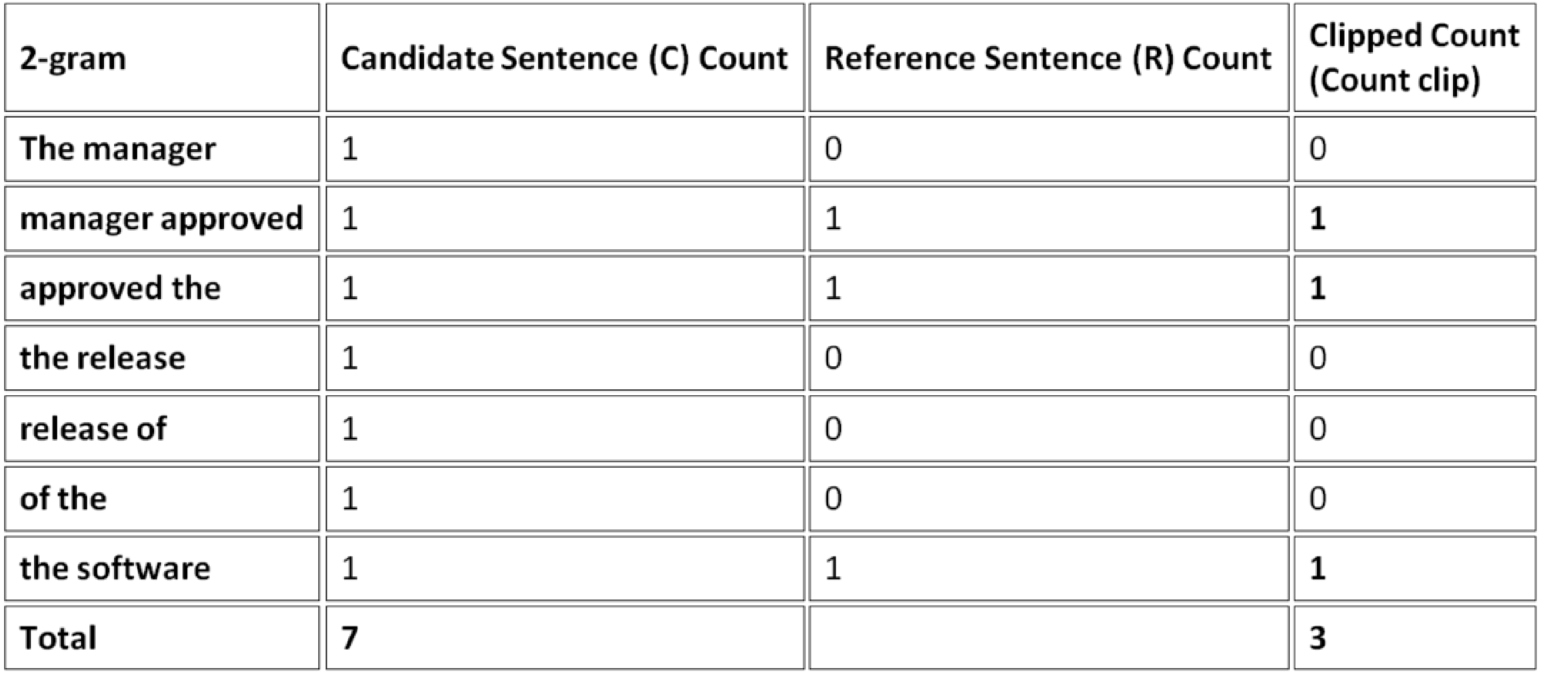
Output length

Type	Sentence (Tokenized)	Length (c) or (r)
Reference (R)	The project manager approved the software release.	r=7
Candidate (C)	The manager approved the release of the software.	c=8

Compute Unigram (1-word)

- Reference 1-gram: the, project, manager, approved, the, software, release
- Candidate 1-gram: the, manager, approved, the, release, of, the, software
- Overlaps: 6 matches out of 8
- Precision: `6/8 = 0.75`

Compute bigram



2-gram	**Candidate Sentence (C) Count**	**Reference Sentence (R) Count**	**Clipped Count (Count clip)**
The manager	1	0	0
manager approved	1	1	**1**
approved the	1	1	**1**
the release	1	0	0
release of	1	0	0
of the	1	0	0
the software	1	1	**1**
Total	**7**		**3**

Compute Bigram (2-word)

- Reference 2-grams: the project, project manager, manager approved, approved the, the software, software release
- Candidate 2-grams: the manager, manager approved, approved the, the release, release of, of the, the software
- Overlaps: 3 matches out of 7
- Precision: `3/7 ≈ 0.42`

Step 2: Apply brevity penalty (BP)

- BP (Brevity Penalty) in the context of the BLEU score calculation serves to penalize candidate sentences that are too short compared to the reference sentence(s).
- The primary goal is to counteract the inflation of n-gram precision that results from short outputs.
- A very short output can achieve a high precision score (if the few words it contains are perfect matches) but is clearly a poor translation, which the BP addresses.

Brevity Penalty formula

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

$$BP = 1$$

Step 2: Apply brevity penalty (BP)

- Candidate length = 8 words
- Reference length = 7 words
- BP=1 (since candidate ≥ reference length)

Step 3: Calculate the BLEU-2 Score

- The BLEU-2 score is the Brevity Penalty multiplied by the geometric mean of `p1` and `p2`. We use uniform weights `w1 = w2 = 1/2 = 0.5`

$$\text{BLEU-2} = BP \cdot \exp\left(\sum_{n=1}^{2} w_n \cdot \ln(p_n)\right)$$

$$\text{BLEU-2} = 1 \cdot \exp\left(0.5 \cdot \ln(p_1) + 0.5 \cdot \ln(p_2)\right)$$

Step 3: Calculate the BLEU-2 Score

$$\text{BLEU-2} = BP \cdot \exp\left(\sum_{n=1}^{2} w_n \cdot \ln(p_n)\right)$$

$$\text{BLEU-2} = 1 \cdot \exp\left(0.5 \cdot \ln(p_1) + 0.5 \cdot \ln(p_2)\right)$$

- p1 = 0.75
- p2 = 0.42
- w1 = w2 = 0.5
- BLEU-2

= 1 × exp (0.5 ln(0.75) + 0.5 ln(0.42))

= exp(-0.57) = **0.56**

BLEU-2 Score Interpretation

- So, BLEU ≈ `0.56 (56%),` meaning the AI’s sentence is moderately close to the reference, which is good structure but not identical phrasing

Class Task on Computing BLEU Score

- Reference summary (human-written):

“**`Sorts a list of numbers in ascending order.`**”

- Candidate summary (AI-generated):

“**`Sorts numbers in increasing order.`**”

End of BLUE Score