



deeplearning.ai

# Sequence to sequence models

---

## Bleu score (optional)

# Evaluating machine translation

French: Le chat est sur le tapis.

Bleu  
bilingual evaluation understudy

Reference 1: The cat is on the mat. ←

Reference 2: There is a cat on the mat. ←

MT output: the the the the the the the.

Precision:

Modified precision:

# Bleu score on bigrams

Example: Reference 1: The cat is on the mat. ←

Reference 2: There is a cat on the mat. ←

MT output: The cat the cat on the mat. ←

	Count	Count <sub>clip</sub>	
the cat	2 ←	1 ←	
cat the	1 ←	0	4
cat on	1 ←	1 ←	<hr/>
on the	1 ←	1 ←	6
the mat	1 ←	1 ←	
	↑		

# Bleu score on unigrams

Example: Reference 1: The cat is on the mat.

Reference 2: There is a cat on the mat.

→ MT output: The cat the cat on the mat. ( $\hat{y}$ )

$$p_1, p_2 = \underline{1.0}$$

$$p_1 = \frac{\sum_{unigram \in \hat{y}} \text{count}_{clip}(unigram)}{\sum_{unigram \in \hat{y}} \text{count}(unigram)}$$

*Handwritten notes:* "unigram" with an arrow pointing to the numerator's variable; "count(unigram)" written below the denominator's term.

$$p_n = \frac{\sum_{ngram \in \hat{y}} \text{count}_{clip}(ngram)}{\sum_{ngram \in \hat{y}} \text{count}(ngram)}$$

*Handwritten notes:* "n-gram" written above the numerator's sum; "count(n-gram)" written below the denominator's term.

# Bleu details

$p_n$  = Bleu score on n-grams only

$p_1, p_2, p_3, p_4$

Combined Bleu score:  $BP \exp\left(\frac{1}{4} \sum_{n=1}^4 p_n\right)$

$BP$  = brevity penalty

$$BP = \begin{cases} 1 & \text{if } \underline{MT\_output\_length} > \underline{reference\_output\_length} \\ \exp(1 - MT\_output\_length/reference\_output\_length) & \text{otherwise} \end{cases}$$