

Sequence to sequence models

Bleu score (optional)

Evaluating machine translation

French: Le chat est sur le tapis.

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.

Precision: Modified precision:

Dilingual evaluation understudy

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	Courtclip	
26		
(<		4
(<	(←	
1 ←	1 6	
←	(6	

[Papineni et. al., 2002. Bleu: A method for automatic evaluation of machine translation]

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.

count (unigram)

unigrames

count (unigram)

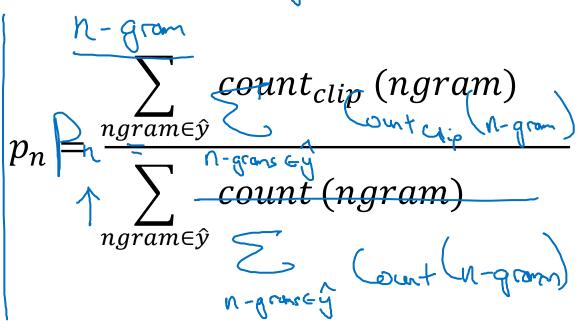
unigrames

unigrames

unigrames

unigrames

unigrames



Bleu details

 $p_n = \text{Bleu score on n-grams only}$

Combined Bleu score:
$$\mathbb{R}^p \exp\left(\frac{1}{4} \sum_{n=1}^4 \mathbb{P}^n\right)$$

$$BP = \begin{cases} 1 & \text{if MT_output_length} > \text{reference_output_length} \\ \exp(1 - \text{MT_output_length}/\text{reference_output_length}) & \text{otherwise} \end{cases}$$