

deeplearning.ai

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:

bilizard evaluation inderstudy

What Bleu does, is given a machine generated translation, it allows you automatically compute the score that measures how good is that machine translation.

Evaluating machine translation



French: Le chat est sur le tapis. the word "the", gets credit up to twice.

bilizand evaluation industribly

- 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:

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	Courtcuip	
the cat	Count 2 (
cat the	(<		et
cat on	(<	(—	
on the	(←	1 6	
the mat	←	(6	

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)

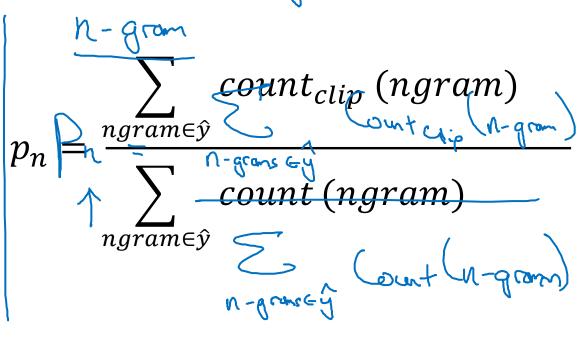
unigrames

unigrames

unigrames

unigrames

unigrames



Bleu details

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

P1, P2, P3, P4

Combined Bleu score:

We actually adjust this with one more factor called the BP penalty.

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

So the blue score is useful single real number evaluation metric to use whenever you want your algorithm to generate a piece of text. And you want to see whether it has similar meaning as a reference piece of text generate by humans.

This is not used in speech recognition because there is usually one ground truth. And you just use another measures to see if you got the speech transcription on pretty much exactly word for word correct.

But for image caption, for multiple captions for a picture could be about equally good or for machine translations, there are multiple translations but equally good.

The Bleu score give you a way to evaluate that automatically and therefore speed up your development.

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

