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 the.

Precision: Modified precision:

Bleu
bilingual evaluation understudy

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

Andrew Ng

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$_{clip}$
the cat	2 ←	1 ←
cat the	1 ←	0
cat on	1 ←	1 ←
on the	1 ←	1 ←
the mat	1 ←	1 ←

$$\frac{4}{6}$$

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

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. $(\hat{y})$

$p_1, p_2 = 1.0$

$$p_1 = \frac{\sum\limits_{unigram \in \hat{y}} count_{clip}(unigram)}{\sum\limits_{unigram \in \hat{y}} count(unigram)}$$

$$p_n = \frac{\sum\limits_{ngram \in \hat{y}} count_{clip}(ngram)}{\sum\limits_{ngram \in \hat{y}} count(ngram)}$$

n-gram

count_clip (n-gram)

n-grams ∈ ŷ

count (n-gram)

n-grams ∈ ŷ

unigram

Unigram ∈ ŷ

Count (unigram)

Unigram ∈ ŷ

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

Bleu details

p_n = Bleu score on n-grams only

P_1, P_2, P_3, P_4

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

$BP = brevity\ penalty$

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

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

Andrew Ng