deeplearning.ai

Sequence to sequence models

Bleu score (optional)

If there are multiple great answers, how do you measure accuracy ?
The way this is done conventionally is through something called the Bleu score.

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

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

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

Evaluating machine translation

网易云课堂

French: Le chat est sur le tapis.

the word "the", gets credit up to twice.

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: $\frac{7}{7}$

Modified precision: $\frac{2}{7}$

$\text{Count}_{clip}(\text{"the"})$

$\text{Count}(\text{"the"})$

Bleu

bilingual evaluation understudy

2 appearances

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

Bleu score on bigrams

pairs of words appearing next to each other.

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]

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. ($\hat{y}$)

$p_1, p_2, \ldots = 1.0$

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

$$p_n = \frac{\sum_{n\text{-}gram \in \hat{y}} count_{clip}(n\text{-}gram)}{\sum_{n\text{-}gram \in \hat{y}} count(n\text{-}gram)}$$

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

Bleu details

p_n = Bleu score on n-grams only

p_1, p_2, p_3, p_4

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

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

BP = brevity penalty is an adjustment factor that penalizes translations systems that output translations that are too short

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

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]