Evaluating machine translation

Bleu score:

french: Le chat est sur le tapis

Reference 1: The cat is on the mat

Reference 2: There is a cat on the mat

bilingual evaluation understudy → actor studies senior actor, in case required to replace

↓

bleu is understudy for human reference.

MT o/p: the the the the the the the

precision: $\frac{7}{7}$ (w.r.t the reference sentence)

modified precision: $\frac{2}{7}$ ← clipped count

Papineni et al. 2002

Bleu: A method for automatic evaluation of machine translation

→ bleu score on bigrams

MT o/p: The cat the cat on the mat

the cat, cat the, cat on, on the, the mat

count:	2	1	1	1	1	$\frac{4}{6}$ $\frac{2}{3}$
$count_{clip}$:	1	0	1	1	1	

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

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

BP → Brevity Penality

↓

penalizes short translation.

blue details:

p_n = Blue score on n-grams only

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

- 1 → if the MT output is > RT
- $\exp(1 - RT/MT)$ → otherwise.

implementations available.

↓

used for image captioning