**BLEU (Bilingual Evaluation Understudy)** is a very common metric for evaluating how well processes generate an expected sentence – originally devised for language translation, but applicable to other tasks with a comparable target sentence (such as image captioning)

Steps to calculating BLEU:

Precision = # n-grams in candidate translation that are also in correct text / total # n-grams in candidate translation

Modified Precision = “clips” the precision numerator so that it doesn’t exceed the total occurrences of an n-gram in the correct text

A brevity penalty is also applied if the candidate translation is shorter than the correct text.

Multiple “correct” translations can be used together, giving the candidate translation multiple chances to match.

By default, calculate separately for 1-grams, 2-grams, 3-grams and 4-grams and weight each equally – can also apply different settings.

Why do we do all this?: research demonstrates that this measure matches well with subjective human assessments of the quality of a translation.

Reference: https://towardsdatascience.com/bleu-bilingual-evaluation-understudy-2b4eab9bcfd1

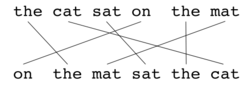
**METEOR (Metric for Evaluation of Translation with Explicit Ordering)**

The harmonic mean of 1-gram precision and recall, with 9x weight on precision.



Precision and recall are calculated after identifying an alignment of which words from the two sentences match to each other, allowing for the possibility that they use a different word ordering. Stemming and synonyms are incorporated in identifying the alignment.

Example alignment generated by the algorithm (which isn’t actually a correct alignment, but in this case it doesn’t affect the METEOR score)



Research indicates that this has an even higher correlation to human judgement than BLEU.

Reference: https://en.wikipedia.org/wiki/METEOR