Decoding Strategies

Evaluation Metrics

Learning goals

- Learn about evaluation metrics for open-ended text generation
- Get to know the different metrics with- and without a gold reference
- Get to know potential issues with some evaluation metrics

HOW DO WE EVALUATE LLMs?

How to choose the appropriate evaluation metric?

- Does the task have a gold reference?
  - BLEU score ▸ Papineni et al., 2002
  - ROUGE score ▸ Lin, 2004
- Are we dealing with open ended text generation without a gold reference?
  - Diversity ▸ Su et al., 2022
  - Coherence ▸ Su et al., 2022
  - MAUVE ▸ Pillutla et al., 2021
- If you have the proper resources choose human evaluation

BLEU SCORE (1)

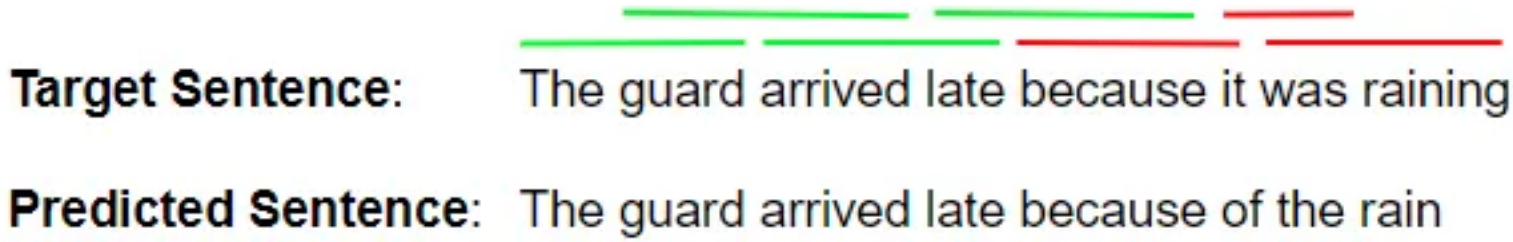
Given a task with a gold reference, e.g machine translation or text summarization, you compare the generated output with the given source reference to compute the BLEU score:

▸ Towards Data Science, Ketan Doshi

Five out of eight 1-grams are correctly predicted:

$\rightarrow p_1 = 5/8$

BLEU SCORE (2)



▸ Towards Data Science, Ketan Doshi

Four out of seven 2-grams are correctly predicted:

$\rightarrow p_2 = 4/7$

You keep doing this procedure until N n-grams and compute a weighted geometric average over the precision scores with weights w_n:

$$exp\left(\sum_{n=1}^{N} w_n \cdot log(p_n)\right)$$

BLEU SCORE - BREVITY PENALTY

In order to penalize very short predictions (it's more likely for shorter sentences to achieve a good precision score) the BLEU score additionally has a brevity penalty term:

$$BP = \begin{cases} 1, & \text{if } c > r \\ e^{(1-r/c)}, & \text{if } c \leq r \end{cases}$$

- With r being the **reference corpus length** and c the **candidate corpus length**
- The final formula is then:

$$BLEU = BP \cdot exp\left(\sum_{n=1}^{N} w_n \cdot log(p_n)\right)$$

ROUGE SCORE

- The ROUGE (Recall-Oriented Understudy for Gisting Evaluation) is a metric commonly used for evaluating the quality of machine-generated text, particularly summaries
- ROUGE measures the similarity between the generated summary and one or more reference (human-written) summaries
- ROUGE includes multiple metrics, such as ROUGE-N (for n-grams), ROUGE-L (for longest common subsequence), and ROUGE-W (for weighted n-grams). Depending on the task, these metrics capture different aspects of summary quality, allowing a more comprehensive evaluation

EXAMPLE: ROUGE-1 PRECISION

Consider the following source sentence S and candidate summary C:

- **S:** The cat is on the mat.
- **C:** The cat and the dog.

Using the ROUGE-N precision score with $N = 1$ *you get:*

- Three correctly predicted unigrams
- Total of number of unigrams in *C* is 5

$\rightarrow$ ROUGE-1 precision $= 3/5 = 0.6$

There are more ROUGE scores as mentioned earlier. You can find more details here: ▸ Medium, Fabio Chiusano

METRICS WITHOUT A GOLD REFERENCE

- BLEU and ROUGE are both used for tasks that have a gold reference you can compare your prediction to
- In open ended text generation you just have a prompt and an output generated by the model
- You don't have any gold reference to compare your output to
- Therefore you have to get a bit more creative with the choice of evaluation metrics

DIVERSITY

COHERENCE

MAUVE