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)

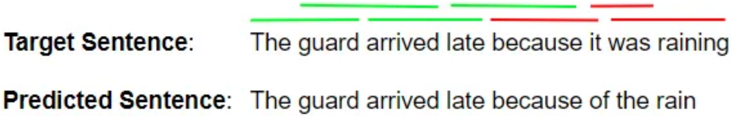
Given a task with a gold reference, e.g machine translation or text summarization, you compare the generated output with the given source reference to compute the BLEU score:

▸ 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

▸ Su et al., 2022 *define diversity in their paper, where they introduce contrastive search, as the generation repetition at different n-gram levels*:

- **Generation Repitition:**
  - Measures the sequence-level repitition as the portion of duplicate *n*-grams in the generated text
  - For a generated text continuation x_{cont} the repitition at *n*-gram level is defined as:

$$\text{rep-n} = 100 \times \left(1.0 - \frac{|\text{unique } n\text{-grams}(x_{cont})|}{|\text{total } n\text{-grams}(x_{cont})|}\right)$$

DIVERSITY (2)

- **Diversity:**
  - Repitition at different *n*-gram levels:

$$\mathrm{DIV} = \prod_{n=2}^{4} \left(1.0 - \frac{\text{rep-n}}{100}\right)$$

  - Plugging in rep-n from the previous slide, this expression simplifies to:

$$\mathrm{DIV} = \prod_{n=2}^{4} \frac{|\text{unique } n\text{-grams}(x_{cont})|}{|\text{total } n\text{-grams}(x_{cont})|}$$

  - A low diversity score suggests the model suffers from repitition, and a high diversity score means the model-generated text is lexically diverse

COHERENCE

This measure was also proposed by ▸ *Su et al., 2022* *as the cosine similarity between the sentence embeddings of the prompt* x_{prompt} *and a generated text conitinuation* x_{cont}:

- They use pre-trained SimCSE sentence embeddings $EMB(x)$ proposed by ▸ Gao et al., 2022:

$$COH(x_{cont}, x_{prompt}) = \frac{EMB(x_{prompt}) \cdot EMB(x_{cont})}{\|EMB(x_{prompt})\| \cdot \|EMB(x_{cont})\|}$$

- The higher the coherence-score the better the model-generated text fits to the given prompt

MAUVE

Pillutla et al., 2021

- A language model is an estimate $\hat{P}(x)$ of the probability distribution over sequences of text $x = (x_1, ..., x_{|\mathbf{x}|})$, consisting of tokens x_t belonging to a fixed vocabulary
- Given a context $x_{1:t}$, a language model $\hat{P}$ and a stochastic decoding strategy we generate text by sampling $\hat{x}_{t+1} \sim \hat{P}(\cdot|x_{1:t}), \hat{x}_{t+2} \sim \hat{P}(\cdot|x_{1:t}, \hat{x}_{t+1})$, etc.
- The decoding strategy and the language model taken together define a distribution Q over text, which we call *model distribution*
- The goal of MAUVE is to measure the gap between the model distribution Q and the target distribution P

SOURCES OF ERROR IN TEXT GENERATION

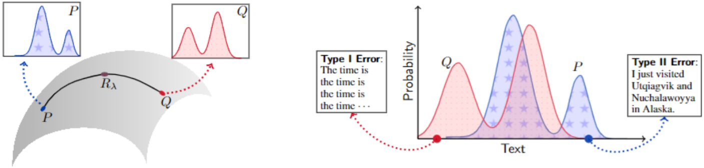


Figure 1: **Left**: MAUVE compares the machine text distribution Q to that of human text P by using the family of mixtures $R_\lambda = \lambda P + (1-\lambda)Q$ for $\lambda \in (0, 1)$. **Right**: Illustration of *Type I errors*, where Q produces degenerate, repetitive text which is unlikely under P, and, *Type II errors*, where Q cannot produce plausible human text due to truncation heuristics [26]. MAUVE measures these errors softly, by using the mixture distribution R_λ. Varying λ in $(0, 1)$ gives a divergence curve and captures a spectrum of soft Type I and Type II errors. MAUVE summarizes the entire divergence curve in a single scalar as the area under this curve.

- The gap between Q and P arises from two sources of error
- Type I error: Q places high mass on text which is unlikely under P
- Type II error: Q cannot generate text which is plausible under P

SOURCES OF ERROR IN TEXT GENERATION

- They formalize the two errors through the Kullback-Leibler divergence:
  - $KL(Q|P)$ penalizes Q if there is a text x that leads to a high $Q(x)$ but a low $P(x)$, which is the Type I error
  - Similarly the Type II error is defined by $KL(P|Q)$
- Issue: both KL divergences are infinite if the supports of Q and P are not identical
- The authors overcome this issue by *softly* measuring the two errors with a mixture distribution:

$$R_\lambda = \lambda P(1 - \lambda)Q \quad \text{for} \quad \lambda \in (0, 1)$$

- (soft) Type I error: $KL(Q, R_\lambda)$
- (soft) Type II error: $KL(P, R_\lambda)$

DIVERGENCE CURVE

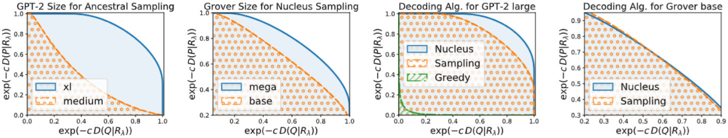


Figure 2: Divergence curves for different models (GPT-2 [45], Grover [61]) and decoding algorithms (greedy decoding, ancestral and nucleus sampling). MAUVE is computed as the area of the shaded region, and larger values of MAUVE indicate that Q is closer to P. In general, MAUVE indicates that generations from larger models and nucleus sampling are closer to human text. **Rightmost**: Nucleus sampling has a slightly smaller Type I error than ancestral sampling but a higher Type II error, indicating that ancestral sampling with Grover base produces more degenerate text while nucleus sampling does not effectively cover the human text distribution.

- To capture all the possible values of the mixture weight λ they vary λ between 0 and 1 to generate a *divergence curve*:

$$C(P, Q) = \{(exp(-cKL(Q|R_\lambda)), exp(-cKL(P|R_\lambda)) : R_\lambda = \lambda P + (1-\lambda)Q, \lambda \in (0,1)\}$$

- $MAUVE(P, Q)$ is the area under this divergence curve, it is a summary of the trade-off between Type I and II errors and lies in (0,1] (more details can be found in the paper ▸ Pillutla et al., 2021)

HUMAN EVALUATION

Why Human Evaluation?

- **Subjectivity of Quality**: Human judgments are essential for evaluating the nuanced quality of text that automatic metrics might miss, such as humor, creativity, and relevance

Key Considerations

- **Evaluators**: Use domain experts or crowdworkers, depending on the task complexity
- **Evaluation Criteria**:
  - **Fluency**: Is the generated text grammatically correct and natural-sounding?
  - **Coherence**: Does the text make logical sense?
  - **Diversity**: Is the output lexically diverse?
- **Challenges**:
  - **Subjectivity**: Different evaluators might have varying opinions, leading to inconsistency

HUMAN EVALUATION

- **Cost and Time**: Human evaluation is resource-intensive
- **Bias**: Evaluators might bring in their biases, which can skew results