ROUGE

ROUGE is a common evaluation indicator in the fields of machine translation, automatic summarization, question and answer generation, etc. ROUGE is calculated by comparing the text generated by the model with the reference text to obtain the corresponding score. ROUGE is very similar to BLEU, and both can be used to measure the matching degree of generated text and reference text. The difference is that ROUGE is based on recall rate, and BLEU pays more attention to accuracy.
The following introduces ROUGE-N, ROUGE-L, ROUGE-W, ROUGE-S.

1 Definition of ROUGE

1.1 ROUGE-N

$$Rouge-N=\frac{\sum_{S\in ReferenceSummaries}\sum_{gram_n\in S}Count_{match}(gram_n)}{\sum_{S\in ReferenceSummaries}\sum_{gram_n\in S}Count(gram_n)}$$

Among them, n stands for $n-gram$. The numerator is the sum of the number of $n-gram$ matching the generated text and a set of reference texts, and the denominator is the sum of the number of $n-gram$ occurrences of the reference text.

1.2 ROUGE-L

ROUGE-L stands for Longest Common Subsequence, which computes the longest common subsequence (LCS) between the generated text and the reference text.

$$R_{lcs} = \frac{LCS(X,Y)}{m}$$

$$P_{lcs} = \frac{LCS(X,Y)}{n}$$

$$F_{lcs} = \frac{(1+\beta^2)R_{lcs}P_{lcs}}{R_{lcs}+\beta^2 P_{lcs}}$$

Among them, X represents the reference text, and Y represents the generated text. m represents the length of X, and n represents the length of Y. $LCS(X,Y)$ represents the longest common subsequence of X and Y. β is a hyperparameter that needs to be set by yourself, and this value is generally relatively large.

1.3 ROUGE-S

The S of Rouge-S means: Skip-Bigram Co-Occurrence Statistics, which is actually an extension of Rouge-N. Skip-bigram is any 2 words in a sentence.

$$R_{skip2} = \frac{SKIP2(X,Y)}{C(m,2)}$$

$$P_{skip2} = \frac{SKIP2(X,Y)}{C(n,2)}$$

$$F_{skip2} = \frac{(1+\beta^2)R_{skip2}P_{skip2}}{R_{skip2}+\beta^2 P_{skip2}}$$

Among them, $C(m,2)$ represents the number of skip-bigrams in the reference text. $C(n,2)$ indicates the number of skip-bigrams in the generated text. $SKIP2(X,Y)$ is the number of skip-bigram matches between the reference text and the generated text. β is a

hyperparameter that controls the relative importance of P_{skip2} and R_{skip2}.

2 Advantages and disadvantages of ROUGE

2.1 Advantages of ROUGE

- The calculation method of ROUGE is relatively simple, and the text similarity is measured by calculating N-gram matching. This makes the results somewhat interpretable.
- ROUGE was originally designed for text summarization tasks, but has since been applied in other text generation tasks, such as machine translation.

2.2 Disadvantages of ROUGE

- ROUGE mainly focuses on the recall rate of the text, that is, whether the key information is covered, while ignoring other important factors such as the precision and fluency of the text.
- ROUGE uses N-gram matching to evaluate text similarity, but N-grams may in some cases be too small to capture long-distance semantic relationships, or too large to be prone to inaccurate matching.
- ROUGE mainly matches from the perspective of vocabulary and phrases, but cannot capture the higher-level syntactic structure and semantic complexity.

References:

https://aclanthology.org/W04-1013.pdf