Model Evaluation

ROGUE SCORE

ROGUE 1

A screenshot of a white board

Description automatically generated

Stop for a moment and imagine that the sentence generated by the model was different by just one word. Not, so it is not cold outside. The scores would be the same. You can get a slightly better score by taking into account bigrams or collections of two words at a time from the reference and generated sentence.

ROGUE 2

A screenshot of a computer

Description automatically generated

You'll notice that the scores are lower than the ROUGE-1 scores. With longer sentences, they're a greater chance that bigrams don't match, and the scores may be even lower.

ROGUE – L

Here L Means Longest Common Subsequence.

Longest Common Subsequence is the Longest Sequence of Words that appear in both sentences in same order but not necessarily consecutively.

A screenshot of a computer

Description automatically generated

A screenshot of a white board

Description automatically generated

ROGUE Hacking or ROGUE Clipping

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

ROGUE vs BLEU ( bilingual evaluation under study)

A diagram of a comparison between a red and a white circle

Description automatically generated with medium confidence

BLEU

A white and purple rectangle with black text

Description automatically generated

The BLEU score quantifies the quality of a translation by checking how many n-grams in the machine-generated translation match those in the reference translation.

A screenshot of a white card with black text

Description automatically generated