Evaluation for Machine Translation

The closer the BLEU score is to one, the better your model is. The closer to zero, the worse it is.

To get the BLEU score, the candidates and the references are usually based on an average of uni, bi, tri or even four-gram precision. To demonstrate, I'll use uni-grams as an example. Look at the following table:

Candidate	I	I	am	I	I
Reference 1	Younes	said	I	am	hungry
Reference 2	He	said	I	am	hungry

To calculate the BLEU score you can do the following.

"I" appears at most once in both, so clip to one: $m_w = 1$

$$\frac{\text{(Sum over unique n-gram counts in the candidate)}}{\text{(total \# of words in candidate)}}$$

You would sum over the unique n-gram counts in the candidate and divide by the total number of words in the candidate. The same concept could apply to unigrams, bigrams, etc. One issue with the BLEU score is that it does not take into account semantics, so it does not take into account the order of the n-grams in the sentence.

Another similar method for evaluation is the ROUGE score which calculates precision and recall for machine texts by counting the n-gram overlap between the machine texts and a reference text. Here is an example that calculates recall:

Recall in ROUGE

Model	The	cat	had	striped	orange	fur
Reference	The	cat	had	orange	fur	

$$\frac{\text{(Sum of overlapping unigrams in model and reference)}}{\text{(total \# of words in reference)}} \quad \frac{5}{5} \quad \text{Recall} = 1$$

Rouge also allows you to compute precision as follows:

Precision in ROUGE

Model	The	cat	had	striped	orange	fur
Reference	The	cat	had	orange	fur	

$$\frac{\text{(Sum of overlapping unigrams in model and reference)}}{\text{(total \# of words in model)}} \quad \frac{5}{6} \quad \text{Precision} = 0.83$$

Mark as completed