Name – Dhruv Singhal **B.Tech. (C.S.E) AI&ML**

SAP ID – 500075346 **Roll No. - R177219074**

Lab Experiment 10

NMT-Automatic Evaluation-BLEU)

BLEU - Bilingual Evaluation Understudy Score

BLEU is a quality metric score for MT systems that attempts to measure the correspondence between a machine translation output and a human translation. The central idea behind BLEU is that the closer a machine translation is to a professional human translation, the better it is.

BLEU scores only reflect how a system performs on the specific set of source sentences and the translations selected for the test. As the selected translation for each segment may not be the only correct one, it is often possible to score good translations poorly. As a result, the scores don't always reflect the actual potential performance of a system, especially on content that differs from the specific test material.

BLEU does not aim to measure overall translation quality, but rather, focuses on strings. Over the years, people have come to interpret this as a safe measure of quality, but most experts would consider BLEU scores more accurate if comparisons were made at a corpus level rather than a sentence level.

How it works

To conduct a BLEU measurement the following is necessary:

1. One or more human reference translations. This should be data that has not been used in building the system (training data) and ideally should be unknown to the MT system developer.
2. It is generally recommended that 1,000 or more sentences be used to get a meaningful measurement. Too small a sample set can sway the score significantly with just a few sentences that match or do not match well.
3. Automated translation output of the exact same source data set.
4. A measurement utility that performs the comparison and score calculation.

Scores are given to individual MT-translated segments – usually sentences – by comparing them with one or a set of good quality human reference translations. When a sentence is translated by two different MT systems, one might produce a translation that matches 75% of the words of the reference correct translation, while the translation of the second MT system might match 55% of the words. Both MT translations might be 100% correct, but the one with the 75% match will be assessed as having provided higher quality, which might seem somewhat arbitrary.

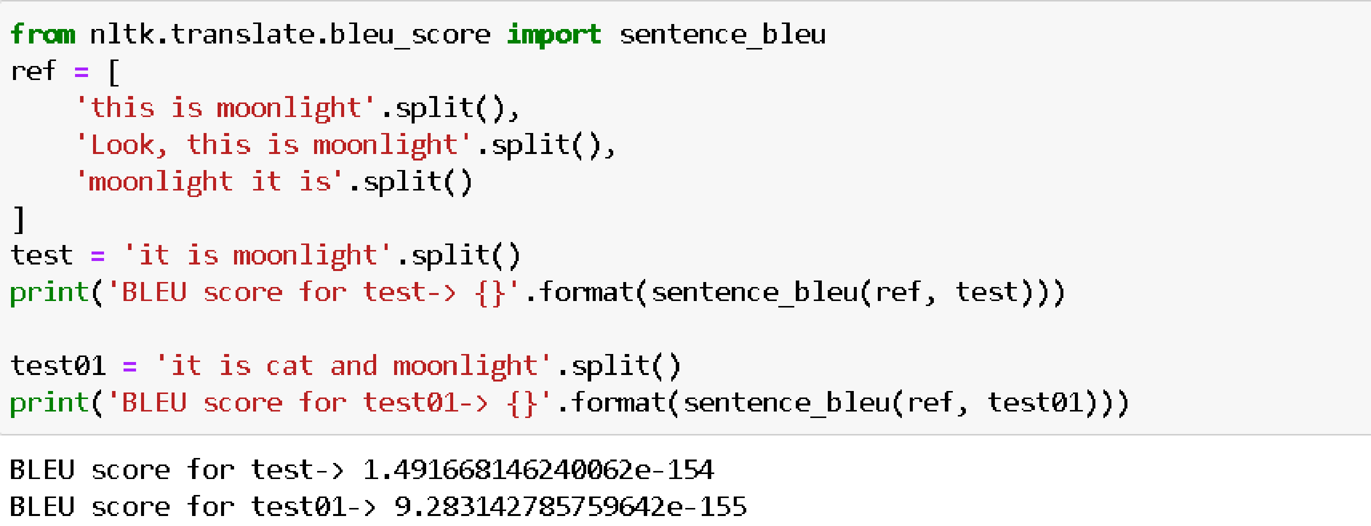
The BLEU metric scores a translation on a scale of 0 to 1, in an attempt to measure the adequacy and fluency of the MT output. The closer to 1 the test sentences score, the more overlap there is with their human reference translations and thus, the better the system is deemed to be. BLEU scores are often stated on a scale of 1 to 100 to simplify communication, but this should not be confused with the percentage of accuracy. The MT output would score 1 only if it is identical to the reference human translation. But even two competent human translations of the exact same material may only score in the 0.6 or 0.7 range as they are

likely to use different vocabulary and phrasing. We should be wary of very high BLEU scores (in excess of 0.7) as it is probably measuring improperly or overfitting.
The BLEU metric also gives higher scores to sequential matching words. That is, if a string of four words in the MT translation match the human reference translation in the same exact order, it will have more of a positive impact on the BLEU score than a string of two matching words will. This means that an accurate translation will receive a lower score if it uses different, but correct words or matching words in a different word order.

DISADVANTAGES:

1. BLEU only measures direct word-to-word similarity and the extent to which word clusters in two sentences are identical. Accurate translations that use different words may score poorly simply because they don't match the selected human reference.
2. There is no consideration of paraphrases or synonyms, so scores can be misleading in terms of overall accuracy. For example, "wander" doesn't get partial credit for "stroll," nor does "sofa" for "couch."
3. Nonsensical language that contains the right phrases in the wrong order can score high. For example, depending on the reference translation, "Appeared calm when he was taken to the American plane, which will to Miami, Florida" might get the very same score as "Was being led to the calm as he was would take carry him seemed quite when taken."

CODE SNIPPET:



```
from nltk.translate.bleu_score import sentence_bleu
ref = [
    'this is moonlight'.split(),
    'Look, this is moonlight'.split(),
    'moonlight it is'.split()
]
test = 'it is moonlight'.split()
print('BLEU score for test-> {}'.format(sentence_bleu(ref, test)))

test01 = 'it is cat and moonlight'.split()
print('BLEU score for test01-> {}'.format(sentence_bleu(ref, test01)))
```

```
BLEU score for test-> 1.491668146240062e-154
BLEU score for test01-> 9.283142785759642e-155
```