Paper: BLEU : a Method for Automatic Evaluation of Machine Translation

Authors: Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu

Reference: Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2001. Bleu. Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL 02(2001). DOI:http://dx.doi.org/10.3115/1073083.1073135

**What they were trying to do?**

They were trying to evaluate the generated responses of machine translation systems

**How did they do it?**

Counting the number of n-gram matches of candidates with reference translation. A higher number of matches means a higher quality translation. They used a precision measure for the matches. Precision means: Number of words you got that are correct/Number of words you got and in the case of BLEU: Number of candidate words that appear in translation/Total number of candidate words.

Sometimes machine translation systems would generate more words that can be found in the reference text leading to high precision. Example from paper:

Candidate: the the the the the the the.

Reference 1: The cat is on the mat.

Reference 2: There is a cat on the mat.

This would give an abnormally high precision. Hence they use modified precision to handle it.

They modified the precision by clipping the number of candidate words that appear in the translation at the maximum number of times that word appears in the reference text. So if the word ‘the’ appears in the reference text twice, they will clip the numerator (No of candidate words that appear in translation/Total no of candidate words) at 2. This can be done for all n-grams. The greater the n-gram matches the better the quality. This can be done for an entire corpus by going sentence by sentence and summing all the matched (clipped) n-gram translations with total number of candidate n-grams in the corpus.

**Verifying the power of using modified n-gram precision**

To verify the power and accuracy in using modified n-gram precision, they compared good human translation with poor machine translation. For each n-gram the good human translation had higher precision. The difference between the two became greater as N increased. Also, precision generally decreased as the N increased. The ability to show the difference showed the use of modified n-gram precision was valid

They also wanted to check if it could distinguish between quality so they applied it to five different translations using different subjects. The understanding was that as machine translations become better, BLEU would then need to move from identifying between poor and excellent to ranking quality. They did this by taking five subject enumerated below:

Individual lacking native proficiency in both Chinese and English H1

Individual who is a native English Speaker H2

Three commercial systems.

Totalling 2 humans and 3 machines

Professional humans were used to judge and rank the translations produced by the subjects. Modified n-gram produced the same ranking as human judges.

They then sough to combining for a single figure measure to represent all the n-gram sizes

They believed an averaging scheme should take the reduction in modified precision over the increasing n-grams into account. The solution was to use a weighted average of the logarithm of modified precisions. This represent a geometric sequence of the modified precisions.

How sentence length factors in the evaluation (Too short or too long)

Ngrams penalizes longer sentences with excess words (Words that do not appear dont get counted and have lower precision).

However this is not sufficient given the below example from the book

Candidate: of the

Reference 1: It is a guide to action that ensures that the military will forever heed Party commands.

Reference 2: It is the guiding principle which guarantees the military forces always being under the command of the Party.

Reference 3: It is the practical guide for the army always to heed the directions of the party.

Modified unigram precision is 2/2

Modified bigram precision is 1/1

Hence some other penalty measure is necessary to effectively account for sentence length.

How recall factors in

Recall: Number of candidate words you match/Total number of candidate

Recall performed over the set of all words would not be useful

Penalties for sentence brevity

Candidates longer than their references are already penalized

Brevity Penalty factor: A high scoring candidate must match in length, word choice and word order.

Brevity penalty is conducted over the whole corpus. If it was done sentence by sentence, shorter sentences would be penalized harshly.

BLEU Calculation:

Find the geometric mean (log) of the test corpus modified precision scores and multiply by exponential brevity factor. Result ranges from 0 to 1. 1 means translation is identical.

Even humans don’t score 1

They broke up a 500 sentence corpus into 20 blocks of 25 sentences each. They found out the the BLEU metric difference between them was statistically significant. BLEU Tracks human judgement quite well