# = BLEU =

BLEU ( bilingual evaluation understudy ) is an algorithm for evaluating the quality of text which has been machine @-@ translated from one natural language to another . Quality is considered to be the correspondence between a machine 's output and that of a human : " the closer a machine translation is to a professional human translation , the better it is " ? this is the central idea behind BLEU . [ 1 ] [ 2 ] BLEU was one of the first metrics to achieve a high correlation with human judgements of quality , [ 3 ] [ 4 ] and remains one of the most popular automated and inexpensive metrics .

Scores are calculated for individual translated segments ? generally sentences ? by comparing them with a set of good quality reference translations . Those scores are then averaged over the whole corpus to reach an estimate of the translation 's overall quality . Intelligibility or grammatical correctness are not taken into account .

BLEU is designed to approximate human judgement at a corpus level , and performs badly if used to evaluate the quality of individual sentences .

BLEU ? s output is always a number between 0 and 1 . This value indicates how similar the candidate and reference texts are , with values closer to 1 representing more similar texts . However , few human translations will attain a score of 1 . The candidate texts must be identical to a reference translation . For this reason , it is not necessary to attain a score of 1 . Because there are more opportunities to match , adding additional reference translations will increase the BLEU score . [ 5 ]

## = = Algorithm = =

BLEU uses a modified form of precision to compare a candidate translation against multiple reference translations . The metric modifies simple precision since machine translation systems have been known to generate more words than are in a reference text . This is illustrated in the following example from Papineni et al . ( 2002 ) ,

Of the seven words in the candidate translation , all of them appear in the reference translations . Thus the candidate text is given a unigram precision of ,

<formula>

where <formula> is number of words from the candidate that are found in the reference , and <formula> is the total number of words in the candidate . This is a perfect score , despite the fact that the candidate translation above retains little of the content of either of the references .

The modification that BLEU makes is fairly straightforward . For each word in the candidate translation , the algorithm takes its maximum total count , <formula> , in any of the reference translations . In the example above , the word " the " appears twice in reference 1 , and once in reference 2 . Thus <formula> .

For the candidate translation , the count <formula> of each word is clipped to a maximum of <formula> for that word . In this case , " the " has <formula> and <formula> , thus <formula> is clipped to 2 . <formula> is then summed over all words in the candidate . This sum is then divided by the total number of words in the candidate translation . In the above example , the modified unigram precision score would be :

<formula>

In practice , however , using individual words as the unit of comparison is not optimal . Instead , BLEU computes the same modified precision metric using n @-@ grams . The length which has the " highest correlation with monolingual human judgements " [ 6 ] was found to be four . The unigram scores are found to account for the adequacy of the translation , how much information is retained . The longer <formula> -gram scores account for the fluency of the translation , or to what extent it reads like " good English " .

Another problem with BLEU scores is that they tend to favor short translations , which can produce very high precision scores , even using modified precision . An example of a candidate translation for the same references as above might be :

the cat

In this example , the modified unigram precision would be ,

<formula>

as the word ' the ' and the word ' cat ' appear once each in the candidate , and the total number of words is two . The modified bigram precision would be <formula> as the bigram , " the cat " appears once in the candidate . It has been pointed out that precision is usually twinned with recall to overcome this problem [ 7 ] , as the unigram recall of this example would be <formula> or <formula> . The problem being that as there are multiple reference translations , a bad translation could easily have an inflated recall , such as a translation which consisted of all the words in each of the references . [ 8 ]

To produce a score for the whole corpus the modified precision scores for the segments are combined using the geometric mean multiplied by a brevity penalty to prevent very short candidates from receiving too high a score . Let <formula> be the total length of the reference corpus , and <formula> the total length of the translation corpus . If <formula> , the brevity penalty applies , defined to be <formula> . ( In the case of multiple reference sentences , <formula> is taken to be the sum of the lengths of the sentences whose lengths are closest to the lengths of the candidate sentences . However , in the version of the metric used by NIST evaluations prior to 2009 , the shortest reference sentence had been used instead . )

iBLEU is an interactive version of BLEU that allows a user to visually examine the BLEU scores obtained by the candidate translations . It also allows comparing two different systems in a visual and interactive manner which is useful for system development . [ 9 ]

## = = Performance = =

BLEU has frequently been reported as correlating well with human judgement , [ 10 ] [ 11 ] [ 12 ] and remains a benchmark for the assessment of any new evaluation metric . There are however a number of criticisms that have been voiced . It has been noted that although in principle capable of evaluating translations of any language , BLEU cannot in its present form deal with languages lacking word boundaries . [ 13 ]

It has been argued that although BLEU has significant advantages , there is no guarantee that an increase in BLEU score is an indicator of improved translation quality . [ 14 ] There is an inherent , systemic problem with any metric based on comparing with one or a few reference translations : in real life , sentences can be translated in many different ways , sometimes with no overlap . Therefore , the approach of comparing by how much any given translation result by a computer differs from just a few human translations is flawed . HyTER is another automated MT metric that compares to very many translations in a reference grammar defined by human translators ; the drawback is then that the human effort involved in correctly defining the combinatorially many ways to render the meaning of the translation in practice means HyTER also is only an approximation .