BLEU : A Method for Automatic Evaluation of Machine Translation

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

Association for Computational Linguistics, 2002

Problems Addressed

1. Human evaluations of machine translation are extensive but expensive, with evaluation possibly taking weeks to months.
2. The authors propose a method of automatic machine translation evaluation that is quick and language-independent, that correlates highly with human evaluation, and that has little marginal cost per run.
3. The method is presented as an automated understudy to skilled human judges which substitutes for them when there is need for quick or frequent evaluations.

Proposed Solution

1. The proposed MT evaluation system has two ingredients:
   - A numerical "translation closeness" metric.
   - A corpus of good quality human reference translations.
2. The closeness metric is inspired by the word error rate metric, appropriately modified for multiple reference translations and allowing for legitimate differences in word choice and word order.
3. The main idea is to use a weighted average of variable length phrase matches against the reference translations.
4. The primary programming task for a BLEU implementor is to compare n-grams of the candidate with the n-grams of the reference translation and count the number of matches.
5. These matches are position independent. The more the matches, the better the candidate translation is.

Modified n-gram precision

- To compute precision, the number of candidate translation words (unigrams) which occur in any reference translation are counted and then divided by the total number of words in the candidate translation.
- MT systems can overgenerate "reasonable" words, resulting in improbable, but high-precision, translations. Therefore, a reference word should be considered exhausted after a matching candidate word is identified.
- This is formalized as the modified unigram precision. It is computed by:
  - Counting the maximum number of times a word occurs in any single reference translation.
  - Clipping the total count of each candidate word by its maximum reference count.

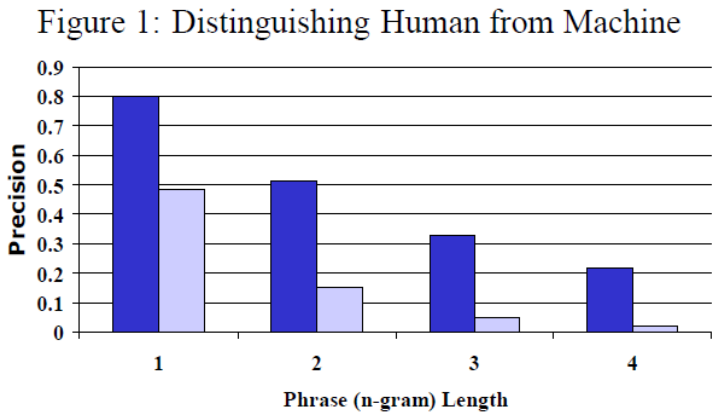


Figure 1: Output of computation of modified precision numbers on the output of a (good) human translator and a standard (poor) machine translation system using 4 reference translations for each of 127 source sentences.

  - Adding these clipped counts up.
  - Dividing by the total (unclipped) number of candidate words.

- The same procedure is generalized to n-grams.

  This sort of modified n-gram precision scoring captures two aspects of translation: adequacy and fluency.

- A translation using the same words (1-grams) as in the references tends to satisfy adequacy.

- The longer n-gram matches account for fluency.

  BLEU only needs to match human judgment when averaged over a test corpus; scores on individual sentences will often vary from human judgments. For example, a system which produces the fluent phrase "East Asian economy" is penalized heavily on the longer n-gram precisions if all the references happen to read "economy of East Asia." The key to BLEU's success is that all systems are treated similarly and multiple human translators with different styles are used, so this effect cancels out in comparisons between systems.

Modified n-gram precision on blocks of text

- First, the n-gram matches are computed sentence by sentence.

- Next, the clipped n-gram counts for all the candidate sentences are computed and are divided by the number of candidate n-grams in the test corpus to compute a modified precision score, p_n, for the entire test corpus.

$$p_n = \frac{\sum_{\mathbb{C} \in Candidates} \sum_{n-gram \in \mathbb{C}} Count_{clip}(n-gram)}{\sum_{\mathbb{C}' \in Candidates} \sum_{n-gram \in \mathbb{C}'} Count_{clip}(n-gram')}$$

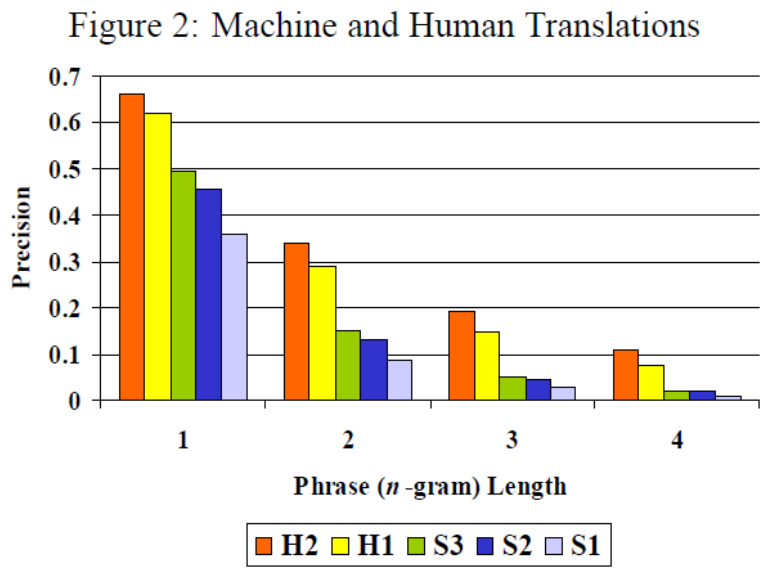
- Results are given in figure 1.

A metric must also reliably distinguish between:

- Translations that do not differ so greatly in quality.

- Two human translations of differing quality.

Combining modified n-gram precisions



- The modified n-gram precision can be seen to decay roughly exponentially with n.
- Therefore, BLEU uses the average logarithm with uniform weights, which is equivalent to using the geometric mean of the modified n-gram precisions.
  - The geometric average is harsh if any of the modified precisions vanish, but this should be an extremely rare event in test corpora of reasonable size (for $N_{max} \leq 4$).
  - Using the geometric average also yields slightly stronger correlation with human judgments than our best results using an arithmetic average.

Sentence Length

- An evaluation metric should enforce that a candidate translation is neither too long nor too short.
  - N-gram precision penalizes spurious words in the candidate that do not appear in any of the reference translations.
  - Modified precision is penalized if a word occurs more frequently in a candidate translation than its maximum reference count.

  Problematic 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.

  Because this candidate is so short compared to the proper length, one expects to find inflated precisions: the modified unigram precision is 2/2, and the modified bigram precision is 1/1.

- A multiplicative brevity penalty factor takes care of this, thus ensuring that a high-scoring candidate translation must now match the reference translations in length, in word choice, and in word order.

  Note that neither this brevity penalty nor the modified n-gram precision length effect directly considers the source length; instead, they consider the range of reference translation lengths in the target language.

- The aim is to make the brevity penalty 1.0 when the candidate's length is the same as any reference translation's length. The closest reference sentence length is termed the "best match length."
- The brevity penalty is not calculated sentence by sentence and then averaged as it would punish the length deviations on short sentences harshly.
- Hence, the brevity penalty is calculated over the entire corpus to allow some freedom at the sentence level.
  - First, the test corpus' effective reference length, r, is computed by summing the best match lengths for each candidate sentence in the corpus.
  - The brevity penalty is chosen to be a decaying exponential in $\frac{r}{c}$, where c is the total length of the candidate translation corpus.

BLEU Evaluation

- Let the geometric average of the modified n-gram precisions be p_n (using n-grams up to length N), w_n be the positive weights (summing to one), c be the length of the candidate translation, and r be the effective reference corpus length.

  Then, we have:

$$\log BLEU = \min(1 - \frac{r}{c}, 0) + \sum_{n=1}^{N} w_n \log p_n$$

  In baseline, $N = 4$ and $w_n = \frac{1}{N}$
- The BLEU metric ranges from 0 to 1.
- The more reference translations per sentence there are, the higher the score BLEU is.

Bibliography

1. E.H. Hovy. 1999. Toward finely differentiated evaluation metrics for machine translation. In Proceedings of the Eagles Workshop on Standards and Evaluation, Pisa, Italy.

2. Kishore Papineni, Salim Roukos, Todd Ward, John Henderson, and Florence Reeder. 2002. Corpus-based comprehensive and diagnostic MT evaluation: Initial Arabic, Chinese, French, and Spanish results. In Proceedings of Human Language Technology 2002, San Diego, CA. To appear.

3. Florence Reeder. 2001. Additional mt-eval references. Technical report, International Standards for Language Engineering, Evaluation Working Group.
http://isscowww. unige.ch/projects/isle/taxonomy2/

4. J.S. White and T. O'Connell. 1994. The ARPA MT evaluation methodologies: evolution, lessons, and future approaches. In Proceedings of the First Conference of the Association for Machine Translation in the Americas, pages 193–205, Columbia, Maryland.