Machine Translation Evaluation: N-grams to the Rescue

Kishore Papineni

IBM T.J. Watson Research Center
Yorktown Heights, NY 10598, U.S.A
papineni@us.ibm.com

Abstract

Human judges weigh many subtle aspects of translation quality. But human evaluations are very expensive. Developers of Machine Translation systems need to evaluate quality constantly. Automatic methods that approximate human judgment are therefore very useful. The main difficulty in automatic evaluation is that there are many correct translations that differ in choice and order of words. There is no single gold standard to compare a translation with. The closer a machine translation is to professional human translations, the better it is. We borrow precision and recall concepts from Information Retrieval to measure closeness. The precision measure is used on variable-length n-grams. Unigram matches between machine translation and the professional reference translations account for adequacy. Longer n-gram matches account for fluency. The n-gram precisions are aggregated across sentences and averaged. A multiplicative brevity penalty prevents cheating. The resulting metric correlates highly with human judgments of translation quality. This method is tested for robustness across language families and across the spectrum of translation quality. We discuss BLEU, an automatic method to evaluate translation quality that is cheap, fast, and good.

1. Introduction

Evaluating translation quality is considered difficult because there is no single gold standard or ground truth for translation. There are many possible correct translations of a given source text, differing in word choice and word order. These differences must be accounted for when judging the quality of a translation. Human judges of translation quality take these and many more subtle aspects into consideration. Collective human judgment of translation quality is therefore the gold standard of evaluation itself. However, such human evaluations are very expensive, and they take a long time to finish. Nor do we benefit from the past human effort when a new system must be evaluated. For MT system developers there is a constant need to evaluate MT quality so that they can weed out bad ideas from good ones. They need automatic evaluation of translation quality that is cheap, fast, and good.

How to measure the goodness of an automatic metric? The grand objective of any automatic metric is to approximate human judgment. Then we can view automatic metrics as predictors of human judgment. Prediction error of a metric is then a natural measure of goodness of a metric: the higher its correlation with human judgment, the better the metric is. We discuss BLEU, a method for automatic evaluation of translation quality that correlates highly with human judgment across language pairs from different language families.

2. Closeness to many reference translations

The central thesis of BLEU is that the closer a machine translation is to professional human translations, the better it is. The closeness measure, to be described later, is inspired by the precision and recall concepts from Information Retrieval and the word error rate in Speech Recognition that has driven the progress in speech technology for over a decade. However, these concepts are modified to take the multiplicity of gold standards into account. If there were a single gold standard for translation, then the traditional word error rate would be sufficient to judge the quality of a translation.

BLEU does not eliminate human effort altogether. Instead, it shifts the effort from expert judges to professional translators in that it requires one or more high quality reference translations. This up-front one-time cost is shared across all system evaluations. The marginal cost of evaluating a new system is negligible. The evaluation itself takes only seconds.

BLEU has two component scores. One is a precision score derived by counting the number of n-gram matches between the candidate translation and the reference translations. We typically count n-gram matches for n from 1 upto 4. Lower-size n-gram matches account for adequacy of the translation while longer n-gram matches account for fluency. The n-gram match counts are first turned into precision numbers and then averaged to get the precision score. The second component of BLEU is a brevity penalty that acts like a cheating detector. Translations that are brief compared to the reference translations incur a penalty that depends on the comparative brevity. So, in order to score high, a translation must match the reference translations in length as closely as possible. Once the length is approximately the same as the references, a translation must produce the same words in roughly the same order as the references to get high precision score. BLEU score is the product of the brevity penalty and the precision score. It is normalized to give a score of 1 to a translation that is identical to any of the reference translations.

Clearly, target sentences that do not share words with reference translations get a BLEU score of 0 — no matter how fluent or grammatical they are. Those that get high scores will match many long n-grams with references and tend to fluently splice reference translation snippets together. The n-gram matching simultaneously accounts for fluency as well as fidelity, assuming that the reference translations fluent and faithful. In summary, to score high on BLEU, a translation must match references in length, in word choice, and in word order.

3. Experimental Results

To measure BLEU's correlation with human judgment, we obtained judgments of translation quality by a pool of judges. An automatic metric ideally predicts human judgment robustly across the spectrum of translation quality and across language families. To assess the robustness across the quality spectrum, we mixed human and machine translations in the set of translations that the humans judged. To test the robustness across several language families, we considered translations from Arabic, Chinese, French, and Spanish into English. The BLEU score correlates highly with human judgments. On Chinese-English, it attains a correlation (R) of 0.99 . On Arabic-English, the correlation is 0.98. On French-English (DARPA-94 evaluation data), the correlation with Adequacy judgment is 0.94 and with Fluency is 0.99. On Spanish-English (DARPA-94 evaluation data), the corresponding numbers are 0.98 and 0.96.