Precision and Recall of Machine Translation

I. Dan Melamed, Ryan Green, and Joseph P. Turian
Computer Science Department
New York University
{lastname}@cs.nyu.edu

Abstract

Machine translation can be evaluated using precision, recall, and the F-measure. These standard measures have significantly higher correlation with human judgments than recently proposed alternatives. More importantly, the standard measures have an intuitive interpretation, which can facilitate insights into how MT systems might be improved. The relevant software is publicly available.

1 Introduction

The benefits of objective evaluation have encouraged many researchers to seek reliable methods for automatically evaluating machine translation (MT) systems. Most efforts have involved some kind of similarity score between the output of an MT system and a "reference" translation. Early approaches to scoring a "candidate" text with respect to a reference text computed similarity in proportion to the number of matching words (e.g., Melamed, 1995). A more recent idea is that matching words in the right order should result in higher scores than matching words out of order (e.g., Rajman & Hartley, 2001). Papineni *et al.* (2002) recently described a simplification of this idea, which they call "BLEU." To measure the syntactic similarity between a candidate and a reference, BLEU counts the number of matching n-grams, for $1 \leq n \leq 4$.

Although BLEU is useful for comparing the relative quality of different MT outputs, it is difficult to gain insight from such a measure. What does a BLEU score of 0.317 mean? We show how MT can be evaluated in terms of well-understood measures such as precision and recall. These measures have significantly higher correlation with human judgments of translation quality than BLEU does. More importantly, these measures have an intuitive graphical interpretation, which can facilitate insights into how an MT system might be improved.

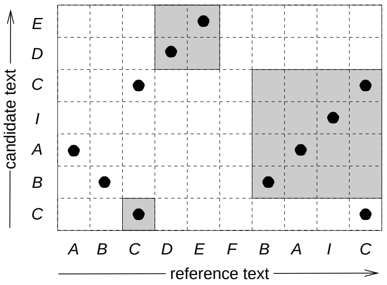

Figure 1: Computation of the maximum match size, using either unigrams or aligned blocks.

2 Precision and Recall of MT

Precision and recall are widely used to evaluate NLP systems. When comparing a set of "candidate" items Y to a set of "reference" items X, $\mathsf{precision}(Y|X) = \frac{|X \cap Y|}{|Y|}$ and $\mathsf{recall}(Y|X) = \frac{|X \cap Y|}{|X|}$. Both functions are proportional to the size of the set intersection in the numerator. The main problem in defining these measures for MT is finding a suitable way to compute $|X \cap Y|$, the intersection of a pair of texts.

2.1 Unigram-Based Measures

The solution is clear when we view the relationship between two texts in a bitext grid. Figure 1 has a hypothetical reference text on the X axis and a hypothetical candidate text on the Y axis. Every cell in the grid is the co-ordinate of some word in one text with some word in the other. Whenever a cell co-ordinates two words that are identical, we place a bullet in it, and call it a **hit**.

As a first approximation, suppose we were not interested in giving more credit for correct word order. A naive approach to computing the match size would be to count the number of hits in the grid. However, this algorithm runs the risk of double-counting, e.g., by awarding two hits for the B in the candidate in Figure 1.

To avoid double-counting, we borrow the concept of "maximum matching" from graph theory. A **matching** is a subset of the hits in the grid, such that no two hits are in the same row or column. The **match size** of a matching is the number of hits in the subset. A **maximum matching** is a matching of maximum possible size for a particular bitext. The **maximum match size (MMS)** of a bitext is the size of any maximum matching for that bitext. The MMS of the bitext in Figure 1 is 7. This definition of MMS guarantees that the MMS ranges from zero to the length of the shorter input text. Therefore, the MMS can be divided by the length of the candidate text (C) or the reference text (R), to derive recall or precision in the usual range between 0 and 1: $\mathsf{precision}(C|R) = \frac{\mathsf{MMS}(C,R)}{|C|}$; $\mathsf{recall}(C|R) = \frac{\mathsf{MMS}(C,R)}{|R|}$.

2.2 Rewards for Longer Matches

Contiguous sequences of matching words show up in a bitext grid as diagonally adjacent hits, running parallel to the main diagonal. We shall refer to such sequences as **runs**. The unigram-based method already rewards a candidate text proportionally to run length, but it produces the same MMS if the hits are not contiguous or in the wrong order. To reward correct word order, it is necessary to reward runs *more* than linearly in their length. BLEU does so by double-counting all sub-runs. We propose to do so by generalizing the definition of match size.

We treat runs as atomic units. Each run's minimum enclosing square is one **aligned block**. A candidate text is rewarded in proportion to the *area* of non-conflicting aligned blocks, as illustrated in Figure 1. Specifically, we define the **weight of a run** to be the square of the run length. We then generalize the definition of **match size** as follows:

$$\mathsf{size}(\mathsf{M}) = \sqrt{\sum_{r\in M} \mathsf{length}(r)^2} \quad (1)$$

where each r is a run in the matching M.

A maximum matching and its size are determined as before.[1] Individual hits that are not part of a longer run are runs of length 1; if they are part of the maximum matching, then they still contribute a weight of $1^2 = 1$ to the MMS. For example, the run CDE in Figure 1 has a weight of 9, but the hits A, B, and C are in the wrong order, so their total weight is only 3. Note that when run r_1 partially conflicts with a longer run r_2, the non-conflicting remainder of r_1 can still participate in the maximum matching. E.g., although the C in run CDE of Figure 1 conflicts with a heavier run, DE still contributes a weight of 4 to the maximum matching.

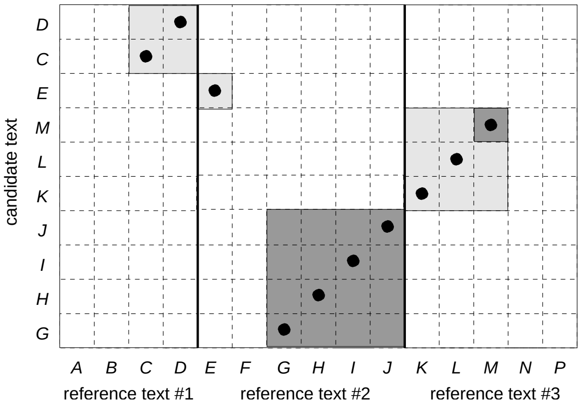

Figure 2: Using multiple references: The initial maximum matching (all shading) is capped by the mean reference length of 5, to arrive at the final matching (dark shading).

The purpose of the square root in Equation 1 is to allow the MMS to be normalized with respect to the lengths of the inputs. In the limiting case that a candidate text is identical to the reference text, the entire bitext grid is covered by one aligned block, and precision = recall = 1.0. Precision and recall scores computed in this manner can be manipulated to derive various other common measures. Their harmonic mean, the so-called "F-measure," has a particularly intuitive interpretation in the context of a bitext grid: It represents the fraction of the grid covered by aligned blocks.

2.3 Multiple References

One of the main sources of variance in MT evaluation measures is the multitude of ways to express any given concept in natural language. A candidate translation can be perfectly correct but very different from a given reference translation. One approach to reducing this source of variance, and thereby improving the reliability of MT evaluation, is to use multiple references (Thompson, 1991).

Figure 2 illustrates how to compute the MMS when multiple reference translations are available. Step 1 is to concatenate the relevant reference texts, in arbitrary order. Step 2 is to find a maximum matching as usual, except that a barrier between adjacent references prevents runs that start in one reference and end in another. Step 3 is to normalize the MMS with respect to the lengths of the input texts.

In the single-reference setting, the MMS is limited by the lengths of the candidate and the refer-

[1] In rare cases, we employ a randomized approximation for efficiency.

ence. In the multiple reference setting, we limit the MMS by the candidate length and the *mean* reference length. We enforce these limits by deleting hits from the maximum matching, until the number of hits is less than or equal to the lower of these two bounds. The hits are deleted in the order that maximizes the size of the remaining matching, i.e. from shorter runs first. After the maximum matching has been pared down, we continue with Equation 1 as before.

3 Experimental Results

We used a corpus of six English translations of 728 Arabic sentences. Two were machine ("candidate") translations and four were human ("reference") translations. The candidates ranged in length from 2 to 116 words (mean 37.1, std.dev. 18.1). Each candidate sentence was manually evaluated on Adequacy and Fluency, on a scale of 1-5.[2]

The reliability of any MT evaluation method will vary with the length of the input. To measure this effect, we created pseudo-documents by concatenating between 1 and 25 randomly chosen candidate sentences. For each candidate "document" we created 4 corresponding reference documents, using each of the available references once. We then computed BLEU scores and F-measures for each candidate and its 4 references. For each of the resulting 2 sets of 4 scores, we computed the Pearson and Spearman correlation with the 2 manual evaluations of Adequacy. The entire procedure was then repeated using multiple references, i.e. all possible combinations of 2 and 3 reference translations per candidate.

Figure 3 shows how BLEU and the F-measure correlate with Adequacy. The graph reveals several results. First, the Pearson coefficient seems inappropriate for this purpose. This is not surprising, because the manual evaluation instructions all but guaranteed that the scores would not be on a linear scale.[2] Second, the Spearman coefficient is not ideal either, because it presumes no ties: The ceiling effect visible in Figure 3 is an artifact of information lost during tie-breaking. Despite this dampening effect, our experiments show that the F-measure can be more than twice as reliable as BLEU in the single-reference setting. Both measures gain reliability from multiple references. However, for very short documents, the F-measure is more reliable with one reference than BLEU is with three.

We performed the same experiments using Fluency instead of Adequacy, and also using precision or recall instead of F-measure. Correlations with Fluency were qualitatively identical, but uniformly lower, in all experiments. Correlations of recall were also qualitatively identical but uniformly lower, and precision was lower still, but higher than BLEU.

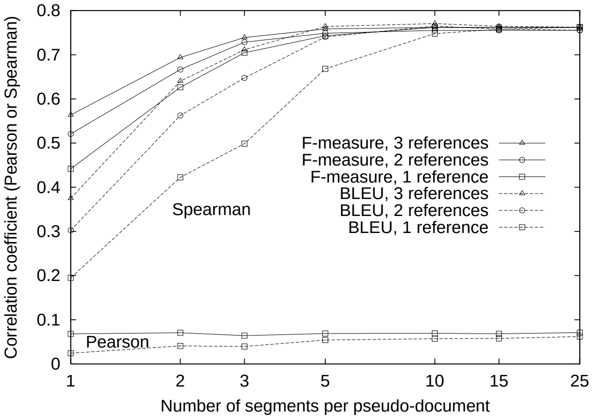

Figure 3: Spearman and Pearson correlation with Adequacy. All correlation differences of 0.03 or more are statistically significant using the Wilcoxon signed ranks test with $\alpha = 0.95$ and $n = 100$ samples.

4 Conclusion

Machine translation can be evaluated using well-known evaluation measures. The standard measures are significantly more reliable than BLEU. Our techniques can be used to compute standard evaluation measures for other NLP tasks where reference texts are available, such as text generation and summarization. The relevant software can be downloaded from `http://nlp.cs.nyu.edu/eval/`. We are also developing tools for visualizing maximum matchings.

Acknowledgment: This research was supported by the DARPA TIDES program.

References

I. Melamed (1995) "Automatic Evaluation and Uniform Filter Cascades for Inducing N-best Translation Lexicons" *Third Workshop on Very Large Corpora (WVLC3).* Boston, MA.

K. Papineni, S. Roukos, T. Ward, and W. Zhu (2002) "BLEU: a Method for Automatic Evaluation of Machine Translation" *Proceedings of the ACL.* Philadelphia, PA.

M. Rajman and T. Hartley (2001) "Automatically predicting MT systems rankings compatible with Fluency, Adequacy or Informativeness scores" *MT Summit VIII.* Santiago de Compostela, Spain.

H. Thompson (1991) "Automatic evaluation of translation quality: Outline of methodology and report on pilot experiment" *Proceedings of the Evaluators' Forum.* ISSCO, Geneva.

[2] See `http://www.ldc.upenn.edu/TIDES/` for details about the corpus and the manual evaluation method.