Blues for BLEU: Reconsidering the Validity of Reference-Based MT Evaluation

Arle Lommel
E-mail: arle.lommel@gmail.com

Abstract

This article describes experiments designed to test whether reference-based machine translation evaluation methods (represented by BLEU) (a) measure translation “quality” and (b) whether the scores they generate are reliable as a measure for systems (rather than for particular texts). It considers these questions via three methods. First, it examines the impact of changing reference translations and using them in combination on BLEU scores. Second, it examines the internal consistency of BLEU scores, the extent to which reference-based scores for a part of a text represent the score of the whole. Third, it applies BLEU to human translation to determine whether BLEU can reliably distinguish human translation from MT output. The results of these experiments, conducted on a Chinese-English news corpus with eleven human reference translations, bring the validity of BLEU as a measure of translation quality into question and suggest that the score differences cited in a considerable body of MT literature are likely to be unreliable indicators of system performance due to an inherent imprecision in reference-based methods. Although previous research has found that human quality judgments largely correlate with BLEU, this study suggests that the correlation is an artefact of experimental design rather than an indicator of validity.

Keywords: machine translation, BLEU, translation evaluation and assessment

1. Introduction

Determining the quality of translation is a fraught and complex task, in part due to the lack of any single, widely accepted definition of what “translation quality” is. In the human translation world, quality is generally assessed by bilingual subject-matter experts who review translations to identify errors. This defect-driven approach (exemplified by systems such as the LISA QA Model and SAE J2450) is quite common in translation-production environments, but it is relatively expensive and time-consuming. For MT developers who may need to rapidly test multiple system configurations, the time factor is a significant barrier.

Prior to the early 2000s, evaluation of machine translation (MT) in particular had been largely ad hoc and driven by the needs of particular evaluation tasks. Because human evaluation is expensive and time consuming there was a push to develop automatic methods of evaluation that could automatically provide results. The first prominent automatic method of this type was BLEU (Papineni, 2002). BLEU provided an automatic measure of similarity between a translation hypothesis and one or more reference translations. The assumption is that the more a translation is like a human reference the more likely to be of a higher quality.

BLEU has since become the most widely used reference-based method for evaluating MT quality. Other reference-based methods, such as METEOR, Word Error Rate (WER), and NIST, have appeared since this time, but BLEU has maintained a prominent role, with many MT-related papers using BLEU-score improvements to evaluate systems and changes.

2. Validity: Do Reference-Based Methods Evaluate “Quality”?

Reference-based methods for evaluation provide a mechanically determined score of string similarity between the translation hypothesis (the output from the system to be evaluated) and the reference translation. If the hypothesis contains the same tokens in the same order as the source, it will receive a high score. If the hypothesis contains other tokens or if they appear in a different order it will receive a lower score.

The implicit assumption is that quality can be measured based on similarity to human translation and that a mechanical measure of similarity is adequate to evaluate quality. Thus for reference-based methods, similarity to human output is assumed to be a reliable proxy for quality: Researchers recognize that it is not a direct measure.

One consequence of this mechanical model for evaluation is that reference-based models are sensitive to the particular references chosen. For example, consider the following source, reference, and hypothesis:

- Source (Hungarian): Ő a ferfi amit láttam
- Reference (English): That’s the man I saw.
- Hypothesis (English): The chap I saw is him

Three of the hypothesis tokens (the, I, and saw) appear in the reference and the word order is quite different. Consequently this hypothesis would receive a low reference-based score (around 50%), even though it is a good (albeit colloquial) translation of the Hungarian source.
Papineni et al. (2002) recognized the variability of human translation, and so BLEU from the beginning has allowed for the use of multiple reference translations. The expectation was that with multiple references, BLEU would be able to account for the variation of possible human translation and thereby not penalize a hypothesis just because it does not happen to look like a particular reference.

In practice, however, most BLEU scores are calculated against a single reference translation. Researchers justify this practice because Coughlin (2003) and subsequent research found that human adequacy and fluency judgments correlate quite well with the output of BLEU (and other reference-based methods), even when a single reference is used. If BLEU can reliably predict human judgment from a single reference, researchers do not need to solicit multiple human translations.

Stating that BLEU corresponds to human judgment, however, runs into a fundamental issue: judgment about what? BLEU is correlated not against the evaluations of professional translators (or even bilinguals) who understand the languages and subject matters under consideration. Coughlin’s experiment (and most subsequent research) relied on the judgment of monolingual evaluators, individuals who could evaluate adequacy only with respect to a reference translation:

We suggest that when [monolingual] human evaluators are forced to make decisions without sufficient context or domain expertise, they fall back on strategies that are not unlike determining n-gram precision. (2003:23, emphasis added)

This finding is not surprising. The human evaluators could base their decisions only on how similar the hypotheses were to the single reference translation: They did not have the linguistic or domain skill to evaluate the hypotheses as translations independent of the particular references they had. Accordingly those translations that were most similar to the reference would be evaluated as the most adequate (i.e., they convey the same information as the reference).

Although Coughlin is very clear that it appears that humans utilized BLEU-like strategies, the claim frequently heard is that BLEU corresponds to human judgment, not that human judgment can correspond to BLEU in a certain (rather artificial) experimental setting.

Understanding the limitations of claims based on monolingual adequacy is important. BLEU and similar methods are predictive of human quality assessments, if human quality assessments are determined in a fashion that encourages BLEU-like assessment. The claim then is circular unless it can be confirmed that both also correspond to other methods of human evaluation.

Although concerns about BLEU and similar metrics have long been voiced (e.g., Callison-Burch et al., 2006), issues of practicality and professional convention have kept them central to the field and they remain in widespread use. They remain in use because there is no alternative.

Nevertheless, the validity of reference-based methods has yet to be demonstrated. Reference-based methods assume that similarity to a given reference is a valid measure of quality and the tests designed to demonstrate that validity bias the results because they use a similar method with human evaluators who cannot independently evaluate the translations without the references that are under consideration.

If “quality” can be measured as similarity to a reference then reference-based methods evaluate quality. How well those judgments correlate to what bilinguals or translators would understand as “quality” is another issue that has not been fully explored. As will be seen below, however, there are reasons to believe that reference-based methods do not measure what would generally be understood as translation quality.

3. Reliability: Are Reference-Based Methods a Reliable Measure of Quality?

Setting aside the issues raised in the last section, let us assume that what BLEU measures is actually “quality.” The next question is whether BLEU is reliable in measuring it.

First off, BLEU is not an absolute measure of quality. Researchers are well aware that a score of 43.1 for one system, for instance, does not mean that the system performs better than one that obtains a score of 42.0. Instead they use them as a relative measure of change with respect to a given set of references. I do not question this usage at one level, but as I will demonstrate, the devil is in the details. Researchers assume that relatively small changes (as little as a quarter point on a 100-point scale) are meaningful. However, this paper demonstrates that such changes are not reliable as an indicator of improvement. Larger changes may be, but small changes with respect to a particular reference cannot be so easily interpreted.

At one level, BLEU is obviously reliable. Because it is mechanical, for a given set of references and a hypothesis BLEU will always generate the exact same score. When the hypothesis changes the score will perfectly reflect the differences. BLEU does not depend on the judgment of an annotator. This reliability is very attractive to researchers who want a way to measure change. So it is clearly reliable in measuring something that has some relationship to translation quality, however complex that relationship may be.
However, at another level, this reliability is highly contingent. If the references change the score will change. If a different portion of an engine’s output is evaluated, even if taken from the same source text, the score will change. Reference-based methods evaluate a particular text, not an engine. To some extent this problem is unavoidable because it reflects the inconsistency of MT engines: An engine may translate one piece of text very well but perform poorly for another, and a measure of quality should distinguish between the two.

But if an engine is working on an internally coherent body of text, the scores for one part should be relatively similar to the scores for another part. If they are not then the scoring method is unreliable as a way to evaluate the engine. This issue is crucial because it reflects the inconsistency of MT engines: An engine may translate one piece of text very well but perform poorly for another, and a measure of quality should distinguish between the two.

4. Experimental Setup

In the experiment described in this paper, I used the standard multi-bleu.perl script\(^1\) to examine whether BLEU is valid and reliable as a measure of translation quality. To do so I performed three experiments, two of which focused on reliability and one of which focused on validity. The experiments used a Chinese-English corpus containing eleven reference translations, each with 993 segments of news data. It was derived from the Multi-Translation Chinese Corpus (Huang et al. 2002). To prepare it I tokenized the text, changed the encoding to UTF-8, and checked alignment.

(In addition, I performed smaller-scale experiments with English>German translations taken from the QTLeap project and text&form, a Berlin-based language service provider. The results from these experiments are not included here but correlated well with the larger Chinese>English corpus study.)

The three experiments examined the following:

1. the impact of using multiple combinations of references on BLEU score
2. the internal consistency of BLEU scores for translations
3. how well BLEU serves to evaluate human translations.

\(^1\) https://github.com/moses-smt/mosesdecoder/blob/master/scripts/generic/multi-bleu.perl

Figure 1. Range of BLEU (max – min) scores for \(n\) references (top) and increases in average BLEU score from adding the \(n\)th reference (bottom)

The first two experiments address reliability. The third address validity.

The hypotheses used in these experiments were generated using two online SMT systems and one online RbMT system. (The smaller-scale English>German test had used five systems—two SMT, two RbMT, and one hybrid, but not all of these systems covered Chinese>English.)

The particular setup and results for each experiment are described below.

5. Impact of Multiple References

In this experiment, I compared the hypotheses against every possible combination of the available references, from a single reference up to all possible references. To make this comparison I created a shell script that ran through all the combinations and fed them into multi-bleu.perl and recorded the resulting BLEU scores. Based on the results I then calculated the range of scores for a given number of references and the average impact of adding the \(n\)th reference.

As can be seen, adding additional references results in a substantial BLEU score increase. The average increase for the \(n\)th reference and the span in scores for \(n\) references are shown in Figure 1. The actual scores
The magnitude of the score change that comes from adding each additional reference is striking. Adding a second reference increase average BLEU score across all systems by 8.28 points. Adding each additional reference provides a declining increase, but adding the 11th reference still provides an average BLEU score increase across all systems of 1.07 points.

On the one hand, the substantial increases here are particularly troubling for BLEU. System changes that show an increase of .25 points are often considered worthy of publication. Adding references translations, even when there are already more references than any normal evaluation would use, creates a larger score increase than most system changes. It is thus not an exaggeration to say that the fastest route to BLEU score improvements would be to simply use more references.

<table>
<thead>
<tr>
<th>No. Refs</th>
<th>MT1 (online SMT)</th>
<th>MT2 (online SMT)</th>
<th>MT3 (online RhMT)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>18.11</td>
<td>7.64</td>
<td>2.39</td>
</tr>
<tr>
<td>2</td>
<td>28.13</td>
<td>7.88</td>
<td>1.88</td>
</tr>
<tr>
<td>3</td>
<td>34.55</td>
<td>7.06</td>
<td>1.50</td>
</tr>
<tr>
<td>4</td>
<td>39.11</td>
<td>6.17</td>
<td>1.20</td>
</tr>
<tr>
<td>5</td>
<td>42.55</td>
<td>5.20</td>
<td>0.96</td>
</tr>
<tr>
<td>6</td>
<td>45.27</td>
<td>4.45</td>
<td>0.76</td>
</tr>
<tr>
<td>7</td>
<td>47.47</td>
<td>3.57</td>
<td>0.60</td>
</tr>
<tr>
<td>8</td>
<td>49.32</td>
<td>2.30</td>
<td>0.47</td>
</tr>
<tr>
<td>9</td>
<td>50.88</td>
<td>1.27</td>
<td>0.55</td>
</tr>
<tr>
<td>10</td>
<td>52.24</td>
<td>0.58</td>
<td>0.23</td>
</tr>
<tr>
<td>11</td>
<td>53.42</td>
<td>1.18</td>
<td>0.71</td>
</tr>
</tbody>
</table>

Table 1. Average BLEU scores by number of references used, with range of scores (highest – lowest scores), standard deviation of scores, and difference from adding the nth reference.

Figure 2. Impact of adding multiple references on various online MT systems (MT1 and MT2: SMT; MT3: RhMT)

appear in Figure 2 and a summary of the data behind these graphs is in Table 1.
This finding means that specifying a “BLEU score” without also specifying the number of references results in a meaningless figure. BLEU can only be interpreted with respect to the number of references used.

This dependence of the score on the number of references should not be surprising since multiple references provide a larger number of potential targets for what the hypothesis can match against. Although the magnitude of the increase is striking, one response would be that as long as the same number of references is used across calculation, BLEU scores should still correlate with human evaluation.

Unfortunately, however, when the range of scores in Figure 1 is considered, it is also apparent that BLEU scores are highly dependent on which references are used: A translation that would score very highly against one set of references might score poorly against another set. The problem is particularly acute when the number of references is small because it is impossible to know where in the spread of potential scores a particular score stands. Consider that for MT1, there is a range of 7.88 BLEU points between the highest- and lowest-scoring combinations of two references. If one system’s output happens to achieve a low score against a particular reference (or even two or three) while another system’s output happens to score well, a researcher might conclude that the first system outperforms the second system, even though a different set of references could produce a different result.

The average standard deviation in BLEU scores for all three systems tested in this experiment when using a single reference is 1.92. This result indicates that BLEU increases of less than this number cannot be considered significant increases for determining how systems perform in general. (They can be considered significant with respect to a particular set of references, but as this experiment shows, significance against particular references does not demonstrate real-world performance increase.)

The results of this study call the reliability of reference-based methods into question, at least when a small number of references are available. Reliability increases with the number of references, but even with 11 references the inherent imprecision in BLEU is larger than the effects observed in many MT experiments.

One possible response to this criticism is that BLEU can be used for ranking systems with respect to one another. Certainly, the averages shown in Figures 1 and 2 would rank the systems in a certain order in every case (whether that order reflects their quality is, as noted above, very much in question). However, note that the spread of scores for MT 1 and MT 2 overlap for up to three references: for a single reference in 8 of the 11 cases MT 2 scores higher than the minimum for MT 1; for two references MT 2 outscores the minimum for MT 1 in 23 of 54 cases; and for three references in 14 of 165 cases. Because it is impossible to know where particular scores fall in the possible range, close rankings (where they differ by 1 or 2 points) may reflect chance rather than actual performance differences.

6. Internal Consistency

In the second experiment, I wrote a script that took a random slice of one half of the segments of each hypothesis and calculated the BLEU score for that half versus the remaining half. I then found the difference between the scores to see how different the two halves were. Because the segments in each half were selected at random they should not be biased by the particular news sources used in particular portions of the corpus.

I repeated this procedure 10,000 times for two online SMT engines. This test allowed me to see how consistent BLEU was in ranking a particular engine’s output (versus in ranking the complete translation). If BLEU is really providing a quality evaluation of the engine rather than the particular translation, the difference in scores between the halves should be quite low; by contrast, if the difference is high, it indicates either that (a) BLEU is limited for evaluating engines rather than particular outputs, or (b) the engines are very inconsistent in their output.

Figure 3 (overleaf) shows the results categorized into bands of 0.1 BLEU point difference. The red column marks the average and the orange the difference within the standard deviation.

For MT1 the average difference between the halves was 1.92 BLEU points (standard deviation = .50) with a minimum difference of 0.24 and a maximum of 3.95.

For MT2 the curve is rather different. It is much more likely to show absolute differences closer to 0 than was MT1. At the same time the range of difference was considerably greater. The average difference was 0.97 (with a standard deviation of = .78). The minimum of 0.00 and a maximum of 5.00.

These results suggest that reference-based scores are actually not terribly consistent at evaluating system performance. Instead they evaluate a particular set of strings consistently, but selecting a different set of strings for evaluation, even from the same corpus, can result in substantial changes in BLEU score. This finding is important when evaluating changes in scores that result from system tweaks: A tweak that results in a relatively large positive change (e.g., a full BLEU point) for one text might result in a negative change for another text, even if taken from the same corpus with a reference from the same translator. Larger corpora should reduce the variability, but would not eliminate it.
This result also shows that some engines are more consistent than others (at least in terms of BLEU). MT2 was more consistent (by almost a full BLEU points) on average than MT1. We see that an absolute change in BLEU score of less than roughly 2.4 (the average + the standard deviation) for MT1 and 1.8 (for MT2) cannot reliably reflect system change because it is within the inherent “noise” of the system with respect to BLEU. A smaller change in BLEU score indicates that the translation changed, but cannot indicate with any certainty whether the change is an actual systematic change in system performance absent additional evidence.

As with the previous experiment, this result suggests that reliability, particularly for changes of BLEU score of less than 2 points, is a major concern and that changes evaluate particular texts rather than particular engines. (However, larger differences for the same text between engines are likely to be significant, so these results do not suggest that BLEU is inherently useless for comparing engines, but rather that its precision and reliability for low values is limited.)

7. Evaluating Human Translation

The final experiment was designed to test validity. It replicated the first experiment but used each of the human reference translations as a hypothesis, treating them in the same way MT would normally be evaluated. Accordingly each professional human translation was compared against possible all combinations of \( n \) references \((n=1 \text{ to } 10, \text{ with a maximum of } 10 \text{ because one reference was always set aside for testing})\).

Figure 4 (overleaf) shows the results for one of the reference sets (due to space constraints, the results for only one translation can be shown).

Not surprisingly, the curves shown look very similar to those from the first experiment. What is surprising, however, is that the BLEU scores are so low: the maximum BLEU score for any of the translations against a single reference was 25.23, a BLEU score that would normally indicate relatively poor performance. In other words BLEU scores seem to indicate that human translations are worse than many MT systems’ output.

And in fact, if we compare the results of the first experiment with this experiment, we find that, in fact, the BLEU scores do seem to indicate that MT may be better than human translation. As shown in Figure 5 (overleaf), one of the SMT systems (MT1) outperformed eight of the eleven reference translations in terms of BLEU and the other (MT2) outperformed one of them. Only the RbMT system (MT3) fell below the human references in each case.

If BLEU determines translation quality, the developer of MT1 could say that it has created a system that outperforms human translators 73% of the time and we would have to conclude that one of the professional human translators just barely managed to exceed the quality of the lowest-performing MT system.

This conclusion is clearly nonsense (as even a casual perusal of the MT hypotheses demonstrates). Rather, it demonstrates that BLEU scores, no matter how well they...
correlate to the judgment of monolingual evaluators comparing MT output to a reference translation, are not determining quality in a sense that is meaningful for comparison with human translation. If a human reference translation, which is considered the gold standard in MT evaluation, can score as low as 11.25 (the lowest score for a human translation against one other reference), then whatever BLEU may be evaluating, it is not useful for determining how MT will perform in any circumstances outside of experimental conditions.

8. Does Human Evaluation Perform Any Better?
If BLEU cannot reliably measure something that can be reasonably understood to be “quality,” what is the alternative? Human evaluators have their own major problems with reliability. In the QTLaunchPad (http://qt21.eu/launchpad) and QT21 (http://qt21.eu) projects, investigators found major disagreement in the number and type of errors. Human reviewers frequently disagree with each other about how good particular translations are. They are inconsistent with themselves from one day to the next. So reliability problems are hardly unique to reference-based methods.

This paper is not meant to suggest that human evaluation can replace BLEU. It is expensive and inconsistent. Although projects like QT21 are trying to learn from human annotation, relatively little work has been completed in this area and we do not yet know how well MT can learn from human annotation and scoring. So I certainly do not mean to indicate that all is gloom and doom for reference-based methods and sunshine and butterflies for human evaluation.

Part of the problem is that reference-based methods provide consistent and seemingly precise scores. While MT researchers are aware of the limits of reference-based assessment, they do not always convey this awareness. When they present a 0.5 BLEU-point increase as significant, others may interpret the research as being more precise and reliable than it is. If researchers claimed that a 0.5 point difference (on a 100 point scale) as determined by human translators were significant, the problem would be clear: Humans are simply not that precise. But, as shown here, reference-based methods are not that precise either, even if they appear to be because that can repeatedly generate the same result.

9. What Are the Alternatives?
If reference-based methods are problematic, what is the alternative? Unfortunately there are no good alternatives. This paper points out severe limitations in how MT researchers use and understand reference-based scores, but it cannot suggest a replacement.

However, it does suggest some ways to improve how researchers use BLEU:

1. Use multiple references. Single references are too variable and misleading. While using 11 references for most research is impractical, using the average of three or four would improve reliability and would prevent situations in which one system happens to perform well against a particular reference and another does not from skewing results.
2. Do not over-interpret small differences. MT researchers should take care not to convey the idea that BLEU values have a precision of less than a few points. Touting 0.5 or even 1.0 point score increases as significant simply overstates what BLEU can actually tell us.
3. Use multiple texts and generate independent scores for them. If score differences are consistent across multiple texts (and not just as an aggregate) that will indicate consistent performance and help increase confidence that score differences matter.

Ultimately, however, what we need is a better understanding of what translation quality is and how to understand and measure it. Until we have a sound
theoretical understanding of what we are trying to measure, it is likely that any alternative measures will fall short as well.

10. Conclusion
The findings detailed in this paper should be troubling. They call into question the significance of a considerable body of MT research that relies on the use of small differences in BLEU (or other reference-based) scores to demonstrate system improvement or to compare systems. Because BLEU is simply a measure of string similarity to a particular reference, it is not evaluating "quality" in any sense that really corresponds to human understanding (even if we see some correlation in judgments). If changing the reference or adding references can change scores so dramatically, then the scores are too sensitive to input. If an increase of two or even three BLEU points falls within the inherent noise of BLEU, then BLEU is unreliable as a measure of quality as it is used today.

These results also suggest that the oft-cited correlation between human judgment and scores from reference-based systems is epiphenomenal to the experimental setup used to measure such correlation. As Coughlin herself suggested in her seminal paper, the correlation in judgment is likely due to the monolingual reviewers considering whether the same words seem to be found in the translation hypothesis and the reference. Because they cannot evaluate the hypothesis on its own terms, they are effectively recreating the judgments of the reference-based approach and their results correlate to BLEU rather than BLEU correlating to any real understanding of translation quality.

These results call into question fundamental approaches in MT development and should be replicated rather than relied upon as is. Unfortunately, there are few corpora with sufficient numbers of references to be used in such studies. Generating reference translations is expensive. Nevertheless, if MT research is to claim that it can be demonstrated. We do not know what shape evaluation will take in the future, but it is clear that reference-based methods on their own provide us imprecise and at-times misleading guidance.

Can BLEU and other similar methods be used to produce valid results that withstand scrutiny? My results suggest that they can be used if the magnitude of change exceeds that of their inherent. An increase of 10 BLEU points, for example, would almost certainly indicate a real quality improvement in MT output. However, given that a change of 0.5 is within the standard deviation for seven reference translations and a change of 2.0 is within the standard deviation for a single reference, it is clear that changes smaller than a few points, no matter what p value is obtained, are not likely to represent real changes in how humans will perceive quality.

Adding additional references helps as well by reducing the likelihood that a change is relative only to a single reference translation. If three or four references are used and a system shows an improvement against each of them individually and in combination, it is likely to represent a real change. But in most cases a system change that shows a score increase with respect to certain references would show a decrease with respect to others. We do not yet have the tools to interpret such results in order to tell if apparent changes are meaningful.

I would also suggest that approaches that combine the judgment of professional human translators with machine evaluation are the only way to be certain about the meaning of changes. Such approaches are being pioneered in the QT21 project, but there is considerable scepticism about their value and utility among researchers. Because human evaluation is time-consuming (and noisy in its own right) researchers have sought more consistent and practical methods. But consistency and practicality are not enough if validity and reliability cannot be demonstrated. We do not know what shape evaluation will take in the future, but it is clear that reference-based methods on their own provide us imprecise and at-times misleading guidance.

11. Acknowledgements
Early stages of this research were funded by the QT21 (http://qt21.eu) project, which has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No. 645452. The conclusions of this research, however, are the author’s own and do not represent the opinions or positions of the QT21 consortium.

Special thanks for Maja Popović (Humboldt Universität zur Berlin), who ran early versions of the multi-reference experiments, and to Aljoscha Burchardt, Kim Harris, and Hans Uszkoreit (DFKI Berlin) for their valuable feedback. Any errors are, of course, my own.

12. Bibliographical References