# Re-evaluating Automatic Summarization with BLEU and 192 Shades of ROUGE

**Yvette Graham**
ADAPT Centre
School of Computer Science and Statistics
Trinity College Dublin
graham.yvette@gmail.com

## Abstract

In this paper, we provide an analysis of current evaluation methodologies applied to summarization metrics and identify the following areas of concern: (1) movement away from evaluation by correlation with human assessment; (2) omission of important components of human assessment from evaluations, in addition to large numbers of metric variants; (3) absence of methods of significance testing improvements over a baseline. We outline an evaluation methodology that overcomes all such challenges, providing the first method of significance testing suitable for evaluation of summarization metrics. Our evaluation reveals for the first time which metric variants significantly outperform others, optimal metric variants distinct from current recommended best variants, as well as machine translation metric BLEU to have performance on-par with ROUGE for the purpose of evaluation of summarization systems. We subsequently replicate a recent large-scale evaluation that relied on, what we now know to be, suboptimal ROUGE variants revealing distinct conclusions about the relative performance of state-of-the-art summarization systems.

## 1 Introduction

Automatic metrics of summarization evaluation have their origins in machine translation (MT), with ROUGE (Lin and Hovy, 2003), the first and still most widely used automatic summarization metric, comprising an adaption of the BLEU score (Papineni et al., 2002). Automatic evaluation in MT and summarization have much in common, as both involve the automatic comparison of system-generated texts with one or more human-generated reference texts, contrasting either *system-output translations* or *peer summaries* with *human reference translations* or *model summaries*, depending on the task. In both MT and summarization evaluation, any newly proposed automatic metric must be assessed by the degree to which it provides a good substitute of human assessment, and although there are obvious parallels between evaluation of *systems* in the two areas, when it comes to evaluation of *metrics*, summarization has diverged considerably from methodologies applied to evaluation of metrics in MT.

Since the inception of BLEU, evaluation of automatic metrics in MT has been by correlation with human assessment. In contrast in summarization, over the years since the introduction of ROUGE, summarization evaluation has seen a variety of different methodologies applied to evaluation of its metrics. Evaluation of summarization metrics has included, for example, the ability of a metric/significance test combination to distinguish between sets of human and system-generated summaries (Rankel et al., 2011), or by accuracy of conclusions drawn from metrics when combined with a particular significance test, Wilcoxon rank-sum (Owczarzak et al., 2012).

Besides moving away from well-established methods such as correlation with human judgment, previous summarization metric evaluations have been additionally limited by inclusion of only a small proportion of possible metrics and variants. For example, although the most commonly used metric ROUGE has a very large number of possible variants, it is common to include only a

small range of those in evaluations. This has the obvious disadvantage that superior variants may exist but remain unidentified due to their omission.

Despite such limitations, however, subsequent evaluations of state-of-the-art summarization systems operate under the assumption that recommended ROUGE variants are optimal and rely on this assumption to draw conclusions about the relative performance of systems (Hong et al., 2014). This forces us to raise some important questions. Firstly, to what degree was the divergence away from evaluation methodologies still applied to MT metrics today well-founded? For example, were the original methodology, by correlation with human assessment, to be applied, would a distinct variant of ROUGE emerge as superior and subsequently lead to distinct system rankings? Secondly, were *all variants* of ROUGE to be included in evaluations, would a variant originally omitted from the evaluation emerge as superior and lead to further differences in summarization system rankings? Furthermore, although methods of statistical significance testing are commonly applied to evaluation of summarization *systems*, attempts to identify significant differences in performance of *metrics* are extremely rare, and when they have been applied unfortunately have not used an appropriate test.

This motivates our review of past and current methodologies applied to the evaluation of summarization metrics. Since MT evaluation in general has its own imperfections, we do not attempt to indiscriminately impose all MT evaluation methodologies on summarization, but specifically revisit evaluation methodologies applied to one particular area of summarization, evaluation of metrics. Correlations with human assessment reveal an extremely wide range in performance among variants, highlighting the importance of an optimal choice of ROUGE variant in system evaluations. Since distinct variants of ROUGE achieve significantly stronger correlation with human assessment than previous recommended best variants, we subsequently replicate a recent evaluation of state-of-the-art summarization systems revealing distinct conclusions about the relative performance of systems. In addition, we include in the evaluation of metrics, an evaluation of BLEU for the purpose of summarization evaluation, and contrary to common belief, precision-based BLEU is on-par with recall-based ROUGE for evaluation of summarization systems.

## 2 Related Work

When ROUGE (Lin and Hovy, 2003) was first proposed, the methodology applied to its evaluation, in one respect, was similar to that applied to metrics in MT, as ROUGE variants were evaluated by correlation with a form of human assessment. Where the evaluation methodology diverged from MT, however, was with respect to the precise representation of human assessment that was employed. In MT evaluation of metrics, although experimentation has taken place with regards to methods of elicitation of assessments from human judges (Callison-Burch et al., 2008), human assessment is always aimed to encapsulate the overall quality of translations. In contrast in summarization, metrics are evaluated by the degree to which metric scores correlate with *human coverage scores* for summaries, a recall-based formulation of the number of peer summary units that a human assessor believed had the same meaning as model summaries. Substitution of overall quality assessments with a recall-based manual metric, unfortunately has the potential to introduce bias into the evaluation of metrics in favor of recall-based formulations.

One dimension of summary quality omitted from human coverage scores is, for example, the order in which the units of a summary are arranged within the summary. Despite unit order quite likely being something of importance to a human assessor, assessment of metrics by correlation with human coverage scores does not in any respect take into account the order in which the units of a summary appear, and evaluation by human coverage scores alone means that a summary with its units scrambled or even reversed in theory receives precisely the same metric score as the original. Given current evaluation methodologies for assessment of metrics, a metric that scores two such summaries *differently* would be unfairly penalized for it. Furthermore, when the *linguistic quality* of summaries has been assessed in parallel with annotations used to compute human coverage scores, it has been shown that the two dimensions of quality do not correlate with one another (no significant correlation) (Pitler et al., 2010), providing evidence that coverage scores alone do not fully represent human judgment of the overall quality of summaries.

Subsequent summarization metric evaluations depart from correlation with human judgment further by evaluating metrics according to the ability of a metric/significance test combination to identify a significant difference between the quality of human and system-generated summaries (Rankel et al., 2011). Unfortunately, the evaluation of metrics with respect to how well they distinguish between *high-quality human summaries* and *all system-generated summaries*, does not provide insight into the task of metrics, to score *better quality system-generated summaries* higher than *worse quality system-generated summaries*, however. This is in contrast to evaluation of MT metrics by correlation with human judgment, where metrics *only* receive credit for their ability to appropriately score system-output documents relative to other system-output documents. Since differences in quality levels between pairs of system-generated summaries are likely to be far smaller than differences in system and human-generated summaries, the methodology unfortunately sets too low a bar for summarization metrics to meet.

Furthermore, the approach to metric evaluation unfortunately does not work in the long-term, as the performance of summarization systems improves and approaches or achieves the quality of a human, a metric that accurately identifies this achievement would be unfairly penalized for it. Separate from the evaluation of metrics, Rankel et al. (2011) make the highly important recommendation of *paired* tests for identification of significant differences in performance of summarization systems. Since data used in the evaluation of summarization systems is not independent, paired tests are more appropriate and more powerful.

Owczarzak et al. (2012) diverge further from correlation with human judgment for evaluation of metrics by assessing the accuracy of metrics to identify significant differences between pairs of systems when combined with a significance test. Although the approach to evaluation of metrics provides insight into the accuracy of conclusions drawn from metric/test combinations, the evaluation is limited by inclusion of only six variants of ROUGE, fewer than 4% of possible ROUGE variants. Despite such limitations, however, subsequent evaluations relied on recommended ROUGE variants to rank state-of-the-art systems (Hong et al., 2014).

Although methods of identifying significant differences in performance are commonly applied to the evaluation of *systems* in summarization, the application of significance tests to the evaluation of summarization *metrics* is extremely rare, and when attempts have been made, unfortunately appropriate tests have not been applied. Computation of confidence intervals for individual correlation with human coverage scores, for example, unfortunately does not provide insight into whether or not a *difference* in correlation with human coverage scores is significant.

## 3 Summarization Metric Evaluation

When large-scale human evaluation of summarization systems takes place, human evaluation commonly takes the form of annotation of whether or not system-generated summary units express the meaning of model summary units, annotations subsequently used to compute *human coverage scores*. In addition, an evaluation of the *linguistic quality* of summaries is commonly carried out. As described in Section 2, when used for the evaluation of metrics, linguistic quality is commonly omitted, however, with metrics only assessed by the degree to which they correlate with human coverage scores. In contrast, we include all available human assessment data for evaluating metrics.

### 3.1 Combining Quality and Coverage

In DUC-2004 (Over et al., 2007), human annotations used to compute summary *coverage* are carried out by identification of matching peer units (PUs), the units in a peer summary that express content of the corresponding model summary. In addition, an overall coverage estimate (E) is provided by the human annotator, the proportion of the corresponding model summary or collective model units (MUs) expressed overall by a given peer summary. Human coverage scores (CS) are computed by combining Matching PUs with coverage estimates as follows:

$$\text{CS} = \frac{|Matching\ PUs| \cdot E}{|MUs|} \quad (1)$$

In addition to annotations used to compute human coverage scores, human assessors were asked to rate the *linguistic quality* of summaries under 7 different criteria, providing ratings from $A$ to $E$, with $A$ denoting highest and $E$ least quality rating.

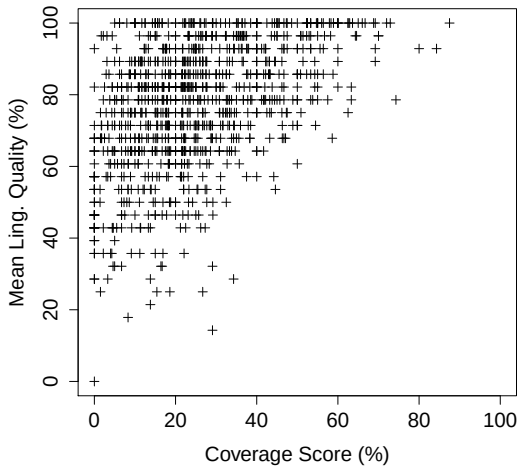


Figure 1: Scatter-plot of mean linguistic quality and coverage scores for human assessments of summaries in DUC-2004

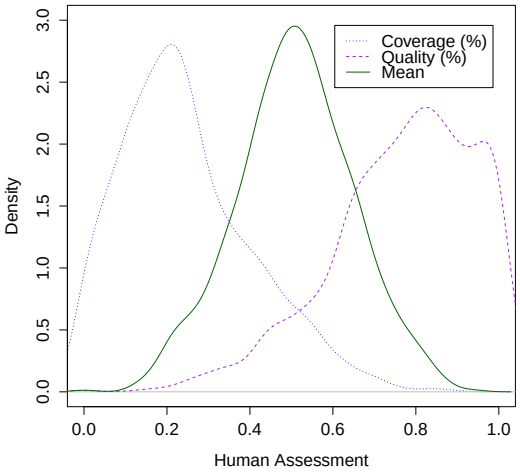


Figure 2: Combining linguistic quality and coverage scores provided by human assessors in DUC-2004

Figure 1 is a scatter-plot of human coverage scores and corresponding linguistic quality scores for all human-assessed summaries from DUC-2004, where, for the purpose of comparison, each of the 7 linguistic quality ratings are converted to a corresponding percentage quality (A= 100%; B= 75%; C= 50%; D= 25%; E= 0%). The location of all points almost exclusively within the upper left corner of the plot in Figure 1 indicates that the linguistic quality of almost all summaries reaches at least as high a level as its corresponding coverage score. This follows the intuition that a summary is unlikely to obtain high coverage without sufficient linguistic quality, while the same cannot be said for the converse, that a high level of linguistic quality necessarily leads to high coverage. More importantly, however, linguistic quality scores provide an additional dimension of human assessment, allowing greater discriminatory power between the quality of summaries than was possible with coverage scores alone.

Figure 2 includes linguistic quality and coverage score distributions from DUC-2004 human evaluation, where each distribution is skewed in opposing directions, in addition to the distribution of the average of the two scores for summaries.

For the purpose of metric evaluation, we combine human coverage and linguistic quality scores using the average of the two scores, and use this as a gold standard human score for evaluation of metrics:

$$\text{Human Assessment Score} = \frac{CS+MLQ}{2}$$

### 3.2 ROUGE

ROUGE includes a large number of distinct variants, including eight choices of n-gram counting method (ROUGE-1; 2; 3; 4; S4; SU4; W; L), binary settings such as word-stemming of summaries and an option to remove or retain stop-words. Additional configurations include the use of precision, recall or f-score to compute individual summary scores. Finally, options for computation of the overall score for a system is by computation of the mean or median of that system's summary score distribution. In total, therefore, when employing ROUGE for the evaluation of summarization systems, there are 192 (8 x 2 x 2 x 3 x 2) possible system-level variants to choose from.

The fact that final overall ROUGE scores for systems are comprised of the mean or median of ROUGE scores of individual summaries, is, again, a divergence from MT evaluation, as n-gram counts used to compute BLEU scores are computed at the document as opposed to sentence-level. However, in this respect, ROUGE has a distinct advantage over BLEU, as the fact that ROUGE comprises the mean or median score of individual summary scores makes possible the application of standard methods of significance testing differences in system-level ROUGE scores, while BLEU is restricted to the application of randomized methods (Koehn, 2004; Graham et al., 2014).

For this purpose, differences in median ROUGE scores can be tested for statistical significance using, for example, Wilcoxon signed-rank test, while paired t-test can be applied to difference of mean ROUGE scores for systems.

### 3.3 Metric Evaluation by Pearson's $r$

Moses (Koehn et al., 2007) multi-bleu[1] was used to compute BLEU (Papineni et al., 2002) scores for summaries and prepare4rouge[2] applied to summaries before running ROUGE (Lin and Hovy, 2003). Table 1 shows the Pearson correlation of each variant of ROUGE with human assessment, in addition to BLEU's correlation with the same human assessment of summaries from DUC-2004. Somewhat surprisingly, BLEU MT evaluation metric achieves strongest correlation with human assessment overall, $r = 0.797$, with performance of ROUGE variants ranging from $r = 0.786$, just below that of BLEU, to as low as $r = 0.293$. For many pairs of metrics, differences in correlation with human judgment are small, however, and prior to concluding superiority in performance of one metric over another, significance tests should be applied.

## 4 Metric Significance Testing

In MT, recent work has identified the suitability of Williams significance test (Williams, 1959) for evaluation of automatic MT metrics (Graham and Baldwin, 2014; Graham et al., 2015; Graham, 2015), and, for similar reasons, Williams test is suited to significance testing differences in performance of competing summarization metrics which we detail further below. Williams test has additionally been used in evaluation of systems that automatically assess spoken and written language quality (Yannakoudakis et al., 2011; Yannakoudakis and Briscoe, 2012; Evanini et al., 2013).

Evaluation of a given summarization metric, $M_{new}$, by Pearson correlation takes the form of quantifying the correlation, $r(M_{new}, H)$, that exists between metric scores for systems and corresponding human assessment scores, and contrasting this correlation with the correlation for some baseline metric, $r(M_{base}, H)$.

One approach to testing for significance that may seem reasonable is to apply a significance test separately to the correlation of each metric with human assessment, with the hope that the new metric will achieve a significant correlation where the baseline metric does not. The reasoning here is flawed however: the fact that one correlation is significantly higher than zero ($r(M_{new}, H)$) and that of another is not, does not necessarily mean that the *difference* between the two correlations is significant. Instead, a specific test should be applied to the difference in correlations. For this same reason, confidence intervals for individual correlations with human assessment are also not useful.

If samples that data are drawn from are independent, and differences in correlations are computed on independent data sets, the Fisher $r$ to $z$ transformation is applied to test for significant differences in correlations. Data used for the evaluation of summarization metrics are not independent, as evaluations comprise three sets of scores for precisely the same set of summaries (corresponding to variables $X_1$, $X_2$ and $X_3$ below), and subsequently three correlations: $r(M_{base}, H)$, $r(M_{new}, H)$ and $r(M_{new}, M_{base})$. If $r(M_{base}, H)$ and $r(M_{new}, H)$ are both $> 0$, then the third correlation, between metric scores themselves, $r(M_{base}, M_{new})$, must also be $> 0$. The strength of this correlation, directly between scores of pairs of summarization metrics, should be taken into account using a significance test of the difference in correlation between $r(M_{base}, H)$ and $r(M_{new}, H)$.

Williams test [3] (Williams, 1959) evaluates the significance of a difference in dependent correlations (Steiger, 1980). It is formulated as follows as a test of whether the population correlation between $X_1$ and $X_3$ equals the population correlation between $X_2$ and $X_3$:

$$t(n-3) = \frac{(r_{13} - r_{23})\sqrt{(n-1)(1+r_{12})}}{\sqrt{2K\frac{(n-1)}{(n-3)} + \frac{(r_{23}+r_{13})^2}{4}(1-r_{12})^3}},$$

where $r_{ij}$ is the correlation between $X_i$ and $X_j$, $n$ is the size of the population, and:

$$K = 1 - {r_{12}}^2 - {r_{13}}^2 - {r_{23}}^2 + 2r_{12}r_{13}r_{23}$$

Since the power of Williams test increases when the third correlation, $r(M_{base}, M_{new})$, between

[1] https://github.com/moses-smt/mosesdecoder/commits/master/scripts/generic/multi-bleu.perl

[2] http://kavita-ganesan.com/content/prepare4rouge-script-prepare-rouge-evaluation

[3] Also known as Hotelling-Williams.

| Metric | Stemming | RSW | Ave./Med | P/R/F | $r$ | Metric | Stemming | RSW | Ave./Med | P/R/F | $r$ | Metric | Stemming | RSW | Ave./Med | P/R/F | $r$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| BLEU | | | | | 0.797 • | R-2 | Y | N | M | F | 0.706 | R-L | Y | Y | A | F | 0.638 |
| R-2 | Y | Y | A | P | 0.786 • | R-3 | N | Y | M | P | 0.704 • | R-1 | N | N | A | F | 0.637 |
| R-3 | N | N | A | F | 0.785 • | R-1 | N | Y | A | P | 0.704 • | R-S4 | Y | N | M | F | 0.634 |
| R-2 | N | Y | A | P | 0.783 • | R-4 | N | N | M | R | 0.703 • | R-4 | Y | N | M | P | 0.634 |
| R-3 | N | Y | A | P | 0.781 • | R-L | N | Y | A | P | 0.700 • | R-1 | N | N | M | F | 0.634 |
| R-3 | Y | N | A | F | 0.779 • | R-W | Y | Y | A | P | 0.700 • | R-SU4 | N | Y | A | R | 0.633 |
| R-3 | N | N | A | R | 0.777 • | R-4 | N | Y | A | R | 0.700 • | R-L | Y | Y | M | P | 0.633 |
| R-4 | N | N | A | F | 0.771 • | R-1 | Y | N | M | P | 0.699 • | R-SU4 | Y | Y | M | R | 0.631 |
| R-3 | N | N | A | P | 0.771 • | R-S4 | N | Y | M | P | 0.698 | R-1 | Y | N | A | F | 0.630 |
| R-3 | Y | N | A | R | 0.770 • | R-1 | Y | Y | A | P | 0.698 • | R-1 | Y | Y | M | F | 0.629 |
| R-2 | N | Y | A | F | 0.769 • | R-3 | N | Y | M | F | 0.697 • | R-S4 | Y | Y | M | R | 0.626 |
| R-4 | N | N | A | R | 0.768 • | R-W | N | N | A | P | 0.696 • | R-S4 | N | N | A | R | 0.626 |
| R-2 | Y | Y | A | F | 0.768 • | R-W | Y | N | A | P | 0.695 • | R-SU4 | Y | N | M | F | 0.625 |
| R-3 | Y | N | A | P | 0.767 • | R-4 | N | N | M | F | 0.695 • | R-S4 | Y | N | A | R | 0.624 |
| R-3 | N | N | M | F | 0.766 • | R-S4 | N | Y | M | F | 0.693 | R-L | N | Y | M | F | 0.623 |
| R-3 | N | Y | A | F | 0.764 • | R-S4 | N | Y | A | F | 0.691 | R-SU4 | Y | Y | A | R | 0.622 |
| R-3 | Y | Y | A | P | 0.764 • | R-SU4 | N | Y | M | P | 0.690 | R-1 | Y | N | M | F | 0.617 |
| R-4 | Y | N | A | F | 0.763 • | R-1 | N | N | M | P | 0.690 • | R-1 | N | Y | M | R | 0.615 |
| R-4 | N | N | A | P | 0.762 • | R-2 | N | N | M | R | 0.689 | R-W | N | Y | A | R | 0.613 |
| R-4 | Y | N | A | R | 0.761 • | R-L | Y | Y | A | P | 0.688 • | R-S4 | N | N | M | R | 0.611 |
| R-3 | N | N | M | P | 0.760 • | R-3 | N | Y | M | R | 0.687 • | R-L | N | Y | M | R | 0.609 |
| R-4 | Y | Y | A | P | 0.759 • | R-S4 | N | N | M | P | 0.687 | R-1 | N | Y | A | R | 0.604 |
| R-2 | Y | N | A | P | 0.759 • | R-S4 | Y | N | A | F | 0.687 | R-L | N | Y | A | R | 0.601 |
| R-4 | N | Y | A | P | 0.758 • | R-S4 | N | N | A | F | 0.687 | R-W | N | N | M | F | 0.600 |
| R-2 | N | N | A | P | 0.757 • | R-4 | N | N | M | P | 0.687 • | R-L | N | N | M | F | 0.599 |
| R-3 | N | N | M | R | 0.753 • | R-L | N | N | A | P | 0.686 • | R-W | Y | Y | A | R | 0.598 |
| R-4 | Y | N | A | P | 0.752 • | R-SU4 | N | N | M | P | 0.686 | R-W | N | Y | M | R | 0.597 |
| R-3 | Y | Y | A | F | 0.748 • | R-L | Y | N | A | P | 0.683 • | R-1 | Y | Y | A | R | 0.595 |
| R-2 | N | N | A | F | 0.747 • | R-W | N | N | M | P | 0.682 • | R-1 | Y | Y | M | R | 0.591 |
| R-2 | Y | N | A | F | 0.747 • | R-W | Y | N | M | P | 0.680 • | R-L | N | N | A | F | 0.586 |
| R-3 | N | Y | A | R | 0.746 • | R-SU4 | Y | N | M | P | 0.678 | R-W | Y | Y | M | F | 0.586 |
| R-3 | Y | N | M | P | 0.744 • | R-SU4 | N | Y | A | F | 0.678 | R-W | Y | N | M | F | 0.585 |
| R-2 | N | Y | M | P | 0.743 • | R-S4 | Y | Y | A | F | 0.676 | R-L | Y | Y | A | R | 0.583 |
| R-3 | Y | N | M | F | 0.743 • | R-SU4 | N | Y | M | F | 0.676 | R-L | Y | Y | M | F | 0.582 |
| R-2 | N | Y | A | R | 0.742 • | R-SU4 | N | N | A | F | 0.673 | R-L | Y | Y | M | R | 0.579 |
| R-2 | Y | Y | M | P | 0.741 • | R-1 | N | Y | M | P | 0.673 | R-L | Y | N | A | F | 0.579 |
| R-2 | N | Y | M | F | 0.740 • | **R-2** | **Y** | **N** | **M** | **R** | 0.672 | R-W | N | N | A | F | 0.579 |
| R-3 | Y | N | M | R | 0.739 • | R-SU4 | Y | N | A | F | 0.671 | R-SU4 | N | N | M | R | 0.576 |
| R-2 | Y | Y | A | R | 0.737 • | R-S4 | N | Y | M | R | 0.670 | R-W | Y | N | A | F | 0.576 |
| R-2 | Y | Y | M | F | 0.735 • | R-S4 | Y | N | M | P | 0.670 | R-SU4 | N | N | A | R | 0.574 |
| R-2 | N | N | M | P | 0.734 • | R-SU4 | Y | Y | A | F | 0.668 | R-SU4 | Y | N | A | R | 0.571 |
| R-3 | Y | Y | M | P | 0.733 • | R-S4 | N | N | M | F | 0.666 | R-L | Y | N | M | F | 0.569 |
| R-3 | Y | Y | A | R | 0.730 | R-W | N | Y | M | P | 0.664 | R-W | Y | Y | M | R | 0.567 |
| R-4 | Y | Y | A | F | 0.729 • | R-S4 | Y | Y | M | P | 0.664 | R-S4 | Y | N | M | R | 0.566 |
| R-3 | Y | Y | M | F | 0.726 • | R-SU4 | Y | Y | M | P | 0.663 | R-SU4 | Y | N | M | R | 0.525 |
| R-S4 | Y | N | A | P | 0.725 • | R-L | N | N | M | P | 0.661 • | R-1 | N | N | M | R | 0.488 |
| R-SU4 | N | N | A | P | 0.724 • | R-SU4 | N | N | M | F | 0.658 | **R-1** | **Y** | **N** | **M** | **R** | 0.477 |
| R-2 | Y | N | M | P | 0.724 | R-1 | N | Y | A | F | 0.656 | R-W | Y | N | M | R | 0.477 |
| R-S4 | N | Y | A | P | 0.724 | R-W | Y | Y | M | P | 0.656 | R-1 | N | N | A | R | 0.470 |
| R-SU4 | Y | N | A | P | 0.723 • | R-S4 | N | Y | A | R | 0.656 | R-W | N | N | M | R | 0.470 |
| R-S4 | N | N | A | P | 0.723 • | R-L | Y | N | M | P | 0.656 • | R-L | N | N | M | R | 0.470 |
| R-2 | N | Y | M | R | 0.722 • | R-W | N | Y | A | F | 0.655 | R-1 | Y | N | A | R | 0.459 |
| R-4 | N | Y | A | F | 0.721 • | R-1 | N | Y | M | F | 0.653 | R-W | N | N | A | R | 0.456 |
| R-1 | N | N | A | P | 0.720 • | R-L | N | Y | A | F | 0.652 | R-W | Y | N | A | R | 0.452 |
| R-2 | N | N | M | F | 0.719 • | R-1 | Y | Y | M | P | 0.651 | R-L | Y | N | M | R | 0.423 |
| R-SU4 | N | Y | A | P | 0.719 | R-S4 | Y | Y | M | F | 0.649 | R-L | N | N | A | R | 0.416 |
| R-1 | Y | N | A | P | 0.714 • | R-1 | Y | Y | A | F | 0.649 | R-L | Y | N | A | R | 0.406 |
| R-2 | Y | Y | M | R | 0.714 • | R-SU4 | Y | Y | M | F | 0.649 | R-4 | Y | Y | M | P | 0.307 |
| R-3 | Y | Y | M | R | 0.713 • | R-SU4 | N | Y | M | R | 0.646 | R-4 | Y | Y | M | F | 0.302 |
| R-4 | Y | Y | A | R | 0.712 • | R-L | N | Y | M | P | 0.645 | R-4 | N | Y | M | P | 0.301 |
| R-S4 | Y | Y | A | P | 0.711 | R-W | N | Y | M | F | 0.642 | R-4 | Y | Y | M | R | 0.297 |
| R-SU4 | Y | Y | A | P | 0.710 | R-W | Y | Y | A | F | 0.642 | R-4 | N | Y | M | F | 0.296 |
| R-2 | N | N | A | R | 0.710 • | **R-4** | **Y** | **N** | **M** | **R** | 0.641 | R-4 | N | Y | M | R | 0.293 |
| R-W | N | Y | A | P | 0.709 • | R-S4 | Y | Y | A | R | 0.641 | | | | | | |
| R-2 | Y | N | A | R | 0.707 • | R-4 | Y | N | M | F | 0.639 | | | | | | |

Table 1: Pearson correlation ($r$) of BLEU and 192 variants of ROUGE (R-*) with human assessment in DUC-2004, with (Y) and without (N) stemming, with (Y) and without (N) removal of stop words (RSW), aggregated at the summary level using precision (P), recall (R) or f-score (F), aggregated at the system level by average (A) or median (M) summary score, correlations marked with • signify a metric/variant whose correlation with human assessment is not significantly weaker than that of any other metric/variant (an optimal variant) according to pairwise Williams significance tests, variants employed in Hong et al. (2014) are in bold.

metric scores is stronger, metrics should not be ranked by the number of competing metrics they outperform, as a metric that happens to correlate strongly with a higher number of competing metrics in a given competition would be at an unfair advantage. This increased power also means, somewhat counter-intuitively, it can happen for a pair of competing metrics for which the correlation between metric scores is strong, that a small difference in competing correlations with human assessment is significant, while, for a different pair of metrics with a larger difference in correlation, the difference is not significant, because $r(M_{base}, M_{new})$ is weak. For example, in Table 1 the difference in correlation with human assessment of BLEU and that of median ROUGE-L precision with stemming and stop-words retained, 0.141 (0.797 − 0.656), is not significant, while the smaller difference in correlation with human assessment between correlations of BLEU and average ROUGE-3 recall with stemming and stop-words removed, 0.067 (0.797 − 0.73) is significant, since scores of the latter pair of metrics correlate with one another with more strength.

As part of this research, we have made available an open-source implementation of statistical tests for evaluation of summarization metrics, at `https://github.com/ygraham/nlp-williams`.

### 4.1 Significance Test Results

In Table 1, • identifies variants of ROUGE not significantly outperformed by any other variant. Figure 3 shows pairwise Williams significance test outcomes for BLEU, the top ten ROUGE variants, as well as current recommended ROUGE variants (Owczarzak et al. (2012)) used to compare systems in Hong et al. (2014). Current recommended best variants of ROUGE are shown to be *significantly outperformed* by several other ROUGE variants.

Although BLEU achieves strongest correlation with human assessment overall, Figure 3 reveals the difference between BLEU's correlation with human assessment and that of the best-performing ROUGE variant as *not statistically significant*, and since ROUGE holds the distinct advantage over BLEU of facilitating standard methods of significance testing differences in scores for systems, for this reason alone we recommend the use of the best-performing ROUGE variant over BLEU, average ROUGE-2 precision with stemming and stop-words removed.

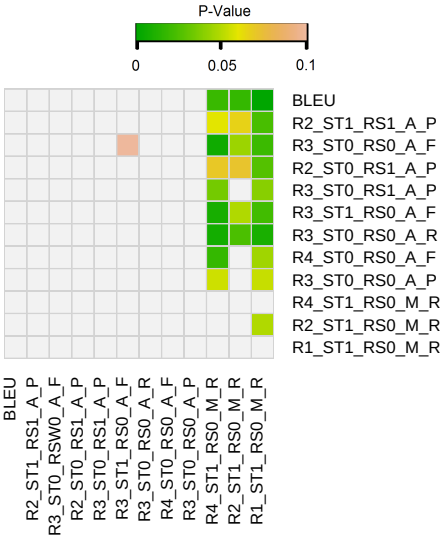


Figure 3: Pairwise significance test outcomes for BLEU, best-performing ROUGE (rows 2-9), and ROUGE applied in Hong et al. (2014) (bottom 3 rows), with (ST1) and without (ST0) stemming, with (RS1) and without (RS0) removal of stop words, for average (A) or median (M) ROUGE precision (P), recall (R) or f-score (F), colored cells denote significant win for row $i$ metric over column $j$ metric with Williams test.

Table 2 shows proportions of optimal ROUGE variants that can be attributed to each of ROUGE's configuration options. Contrary to prior belief, the vast majority of optimal ROUGE variants are precision-based, showing that the assumption that recall-based metrics are superior for evaluation of summarization systems to be inaccurate, and a likely presence of bias in favor of recall-based metrics in evaluations by correlation with human coverage scores alone. Furthermore, since there exists a vast number of possible formulations that could potentially be applied to evaluation of summaries that are neither purely precision nor recall-based, evaluation methodologies should avoid reliance on assumptions that either precision or recall is superior and instead base conclusions on empirical evidence where possible.

## 5 Summarization System Evaluation

Since we have established that the variants of ROUGE used to rank state-of-the-art and baseline summarization systems in Hong et al. (2014) have significantly weaker correlations with human assessment than several other ROUGE variants, this

| N-gram Count | |
|---|---|
| R-3 | 28.7 |
| R-2 | 25.0 |
| R-4 | 18.8 |
| R-1 | 7.5 |
| R-L | 7.5 |
| R-W | 7.5 |
| R-S4 | 2.5 |
| R-SU4 | 2.5 |

| Stemming | |
|---|---|
| Not Stemmed | 53.8 |
| Stemmed | 46.2 |

| Stop-words | |
|---|---|
| Not Rem. | 56.2 |
| Removed | 43.8 |

| Summary-level Agg. | |
|---|---|
| Prec. | 52.5 |
| F-score | 25.0 |
| Recall | 22.5 |

| System-level Agg. | |
|---|---|
| Average | 63.7 |
| Median | 36.3 |

Table 2: Proportions of optimal ROUGE variants attributed to each ROUGE configuration option (%).

| System | ROUGE Best | ROUGE Original |
|---|---|---|
| DPP | 8.498 | 9.62 |
| ICSISumm | 8.317 | 9.78 |
| RegSum | 8.187 | 9.75 |
| Submodular | 8.047 | 9.35 |
| CLASSY11 | 7.717 | 9.20 |
| CLASSY04 | 7.690 | 8.96 |
| OCCAMS_V | 7.643 | 9.76 |
| GreedyKL | 6.918 | 8.53 |
| FreqSum | 6.838 | 8.11 |
| TsSum | 6.671 | 8.15 |
| Centroid | 6.660 | 7.97 |
| LexRank | 6.655 | 7.47 |

Table 3: Summarization systems originally included in Hong et al. (2014) evaluated with the best-performing ROUGE variant (Best): average ROUGE-2 precision with stemming and stop words removed; and evaluated with original suboptimal variant (median ROUGE-2 recall with stemming and without removal of stop-words)

motivates our replication of the evaluation. We evaluate systems using the variant of ROUGE that achieves strongest correlation with human assessment, average ROUGE-2 precision with stemming and stop-words removed.

Table 3 shows ROUGE scores for summarization systems originally presented in Hong et al. (2014). System rankings diverge considerably from those of the original evaluation. Notably, the system now taking first place had originally ranked in fourth position.

Since the best variant of ROUGE is based on average ROUGE scores as opposed to median ROUGE scores, a difference of means significance test is

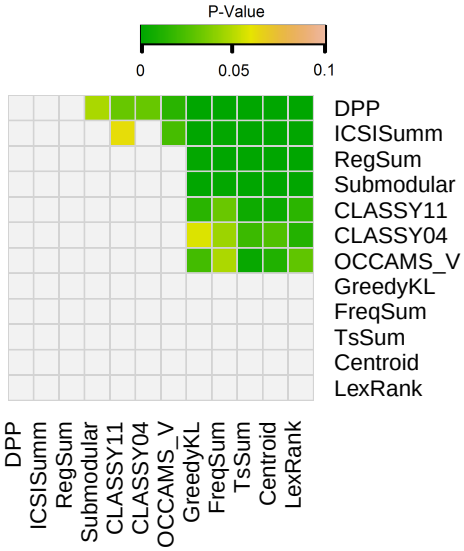


Figure 4: Summarization system pairwise significance test outcomes (paired t-test) for state-of-the-art (top 7 rows) and baseline systems (bottom 5 rows) of Hong et al. (2014) evaluated with best-performing ROUGE variant: average ROUGE-2 precision with stemming and stop words removed, colored cells denote a significant greater mean score for row $i$ system over column $j$ system according to paired t-test.

appropriate provided the normality assumption of score distributions for systems is not violated. In addition, since data used to evaluate systems are not independent, *paired* tests are also appropriate (Rankel et al., 2011). ROUGE score distributions for systems were tested for normality using the Shapiro-Wilk test (Royston, 1982) where score distributions for none of the included systems were shown to be significantly non-normal.

Figure 4 shows outcomes of paired t-tests for summary score distributions of each pair of systems, revealing three summarization systems not significantly outperformed by any other as DPP, ICSISUMM and REGSUM. In addition, as expected, all state-of-the-art systems significantly outperform all baseline systems.

## 6 Human Assessment Combinations

In order to evaluate metrics by correlation with human assessment, it is necessary to obtain a single human evaluation score per system. For example, in the evaluation of metrics in Section 3, we combined linguistic quality and coverage into a single score using the mean of the two scores. Other combinations are of course possible, but without

| Metric | Stem. | RSW | Ave/Med | P/R/F | Mean | Geometric Mean | Harmonic Mean | Coverage Only | Ling. Qual. Only |
|---|---|---|---|---|---|---|---|---|---|
| BLEU | | | | | **0.797•** | **0.901•** | **0.936•** | **0.944•** | 0.642• |
| ROUGE-2 | Y | Y | A | P | 0.786• | 0.870• | 0.887• | 0.878 | 0.660• |
| ROUGE-3 | N | N | A | F | 0.785• | 0.869• | 0.893 | 0.894 | 0.650• |
| ROUGE-2 | N | Y | A | P | 0.783• | 0.868• | 0.885• | 0.876 | 0.658• |
| ROUGE-3 | N | Y | A | P | 0.781• | 0.836• | 0.840 | 0.826 | **0.682•** |
| ROUGE-3 | Y | N | A | F | 0.779• | 0.866• | 0.891 | 0.893 | 0.643• |
| ROUGE-3 | N | N | A | R | 0.777• | 0.871• | 0.901 | 0.907 | 0.632• |
| ROUGE-4 | N | N | A | F | 0.771• | 0.843• | 0.863 | 0.866 | 0.645• |
| ROUGE-3 | N | N | A | P | 0.771• | 0.837• | 0.849 | 0.843 | 0.658• |
| ROUGE-3 | Y | N | A | R | 0.770• | 0.867• | 0.899 | 0.905 | 0.624• |
| ROUGE-2 | N | Y | A | F | 0.769• | 0.877• | 0.909• | 0.910• | 0.619• |
| ROUGE-2 | Y | Y | A | F | 0.768• | 0.875• | 0.908• | 0.908• | 0.618• |
| ROUGE-3 | Y | N | A | P | 0.767• | 0.835• | 0.849 | 0.843 | 0.652• |
| ROUGE-3 | Y | Y | A | P | 0.764• | 0.825• | 0.832 | 0.821 | 0.660• |
| ROUGE-4 | N | N | A | P | 0.762• | 0.815• | 0.824 | 0.819 | 0.657• |
| ROUGE-4 | Y | Y | A | P | 0.759• | 0.794 | 0.790 | 0.774 | 0.678• |
| ROUGE-4 | N | Y | A | P | 0.758• | 0.793 | 0.789 | 0.772 | 0.678• |
| ROUGE-4 | Y | N | A | P | 0.752• | 0.809 | 0.819 | 0.815 | 0.646• |
| ROUGE-2 | N | N | A | F | 0.747• | 0.867• | 0.907• | 0.910• | 0.587• |
| ROUGE-2 | Y | N | A | F | 0.747• | 0.868• | 0.908• | 0.912• | 0.586• |
| ROUGE-2 | N | Y | A | R | 0.742• | 0.862• | 0.904• | 0.912• | 0.578• |
| ROUGE-2 | N | Y | M | F | 0.740• | 0.855• | 0.894• | 0.898• | 0.584• |
| ROUGE-2 | Y | Y | A | R | 0.737• | 0.858• | 0.900• | 0.908• | 0.575• |
| ROUGE-2 | N | Y | M | R | 0.722• | 0.848• | 0.895• | 0.905• | 0.553• |
| ROUGE-2 | N | N | M | R | 0.689 | 0.828 | 0.884• | 0.901• | 0.508 |

Table 4: Correlation of top-ten metric variants for each alternate combination of linguistic quality and coverage, • denotes a metric not significantly outperformed by any other under that particular human evaluation combination, highest correlations highlighted in bold font.

any additional human evaluation data, it is challenging to identify the combination that best represents an overall human assessment for a given summary. One possibility would be to search for optimal weights for combining quality and coverage, but there is a risk with this approach that we will not find the most representative combination but simply the combination that best describes the metrics.

An additional variation of human assessment scores is by combining coverage and quality with a variant of the arithmetic mean, such as the harmonic or geometric mean. Table 4 shows correlations of BLEU and the top ten performing variants of ROUGE when evaluated against the arithmetic (mean), harmonic and geometric mean of quality and coverage scores for summaries. In addition, Table 4 includes correlations of metric scores with coverage alone, as well as linguistic quality scores alone, although linguistic quality scores alone do not provide a sufficient evaluation of metrics – since it is possible to generate summaries with perfect linguistic quality without inclusion of any relevant content whatsoever. We include linguistic quality correlations for metrics simply to provide additional insights.

BLEU MT metric achieves highest correlation across all human evaluation combinations and highest again when evaluated against human coverage scores alone, and BLEU's brevity penalty, that essentially represents recall, is a probable cause of the metric overcoming the recall-based bias of an evaluation based on coverage scores alone. In addition, our recommended variant, ave. ROUGE-2 prec. with stemming and stop words removed is not significantly outperformed by BLEU or any other variant of ROUGE for any of the three combined mean human assessment scores.

## 7 Conclusions

An analysis of evaluation of summarization metrics was provided with an evaluation of BLEU and 192 variants of ROUGE. Detail of the first suitable summarization metric significance test, Williams test, was provided. Results reveal superior variants of metrics distinct from previously best recommendations. Replication of a recent evaluation of state-of-the-art summarization systems also revealed contrasting conclusions about the relative performance of systems. In addition, BLEU achieves strongest correlation with human assessment overall, but does not significantly outperform the best-performing ROUGE variant.

### Acknowledgements

We wish to thank the anonymous reviewers. This research is supported by Science Foundation Ireland through the CNGL Programme (Grant 12/CE/I2267) in the ADAPT Centre (`www.adaptcentre.ie`) at Trinity College Dublin.

## References


C. Callison-Burch, C. Fordyce, P. Koehn, C. Monz, and J. Schroeder. 2008. Further meta-evaluation of machine translation. In *Proc. 3rd Wkshp. Statistical Machine Translation*, pages 70–106, Columbus, Ohio. Association for Computational Linguistics.

K. Evanini, S. Xie, and K. Zechner. 2013. Prompt-based content scoring for automated spoken language assessment. In *Proc. 2013 Conf. North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 157–162, Atlanta, GA. Association for Computational Linguistics.

Y. Graham and T. Baldwin. 2014. Testing for significance of increased correlation with human judgment. In *Proc. 2014 Conf. Empirical Methods in Natural Language Processing*, pages 172–176, Doha, Qatar. Association for Computational Linguistics.

Y. Graham, N. Mathur, and T. Baldwin. 2014. Randomized significance tests in machine translation. In *Proc. 9th Wkshp. Statistical Machine Translation*, pages 266–274, Baltimore, MD. Association for Computational Linguistics.

Y. Graham, N. Mathur, and T. Baldwin. 2015. Accurate evaluation of segment-level machine translation metrics. In *Proc. 2015 Conf. North American Chapter of the Association for Computational Linguistics - Human Language Technologies*, pages 1183–1191, Denver, CO. Association for Computational Linguistics.

Y. Graham. 2015. Improving evaluation of machine translation quality estimation. In *Proc. Fifty-Third Annual Meeting of the Association for Computational Linguistics*, pages 1804–1813, Beijing, China. Association for Computational Linguistics.

K. Hong, J.M. Conroy, B. Favre, A. Kulesza, H. Lin, and A. Nenkova. 2014. A repository of state of the art and competitive baseline summaries for generic news summarization. In *Proc. 9th edition of the Language Resources and Evaluation Conference*, pages 1608–1616, Reykjavik, Iceland.

P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proc. 45th Annual Meeting of the Association for Computational Linguistics*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

P. Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proc. 2004 Conf. Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain. Association for Computational Linguistics.

C.Y. Lin and E. Hovy. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proc. 2003 Conf. North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pages 71–78, Edmonton, Canada. Association for Computational Linguistics.

P. Over, H. Dang, and D. Harman. 2007. Duc in context. *Information Processing & Management*, 43(6):1506–1520.

K. Owczarzak, J.M. Conroy, H.T. Dang, and A. Nenkova. 2012. An assessment of the accuracy of automatic evaluation in summarization. In *Proc. Wkshp. on Evaluation Metrics and System Comparison for Automatic Summarization*, pages 1–9, Quebec, Canada. Association for Computational Linguistics.

K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proc. 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, PA. Association for Computational Linguistics.

E. Pitler, A. Louis, and Ani A. Nenkova. 2010. Automatic evaluation of linguistic quality in multi-document summarization. In *Proc. 48th Annual Meeting of the Association for Computational Linguistics*, pages 544–554, Uppsala, Sweden. Association for Computational Linguistics.

P. Rankel, J.M. Conroy, E.V Slud, and D.P. O'Leary. 2011. Ranking human and machine summarization systems. In *Proc. 2011 Conf. Empirical Methods in Natural Language Processing*, pages 467–473, Edinburgh, Scotland. Association for Computational Linguistics.

P. Royston. 1982. Algorithm as 181: The W test for normality. *Applied Statistics*, 31:176–180.

J.H. Steiger. 1980. Tests for comparing elements of a correlation matrix. *Psychological Bulletin*, 87(2):245–251.

E.J. Williams. 1959. *Regression analysis*, volume 14. Wiley New York.

H. Yannakoudakis and T. Briscoe. 2012. Modeling coherence in ESOL learner texts. In *Proc. Seventh Wkshp. on Building Educational Applications Using NLP*, pages 33–43, Montreal, Canada. Association for Computational Linguistics.

H. Yannakoudakis, T. Briscoe, and B. Medlock. 2011. A new dataset and method for automatically grading ESOL texts. In *Proc. 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 180–189, Portland, OR. Association for Computational Linguistics.