# Achieving Accurate Conclusions in Evaluation of Automatic Machine Translation Metrics

**Yvette Graham**
School of Computing
Dublin City University
graham.yvette@gmail.com

**Qun Liu**
ADAPT Research Centre
Dublin City University
qliu@computing.dcu.ie

## Abstract

Automatic Machine Translation metrics, such as BLEU, are widely used in empirical evaluation as a substitute for human assessment. Subsequently, the performance of a given metric is measured by its strength of correlation with human judgment. When a newly proposed metric achieves a stronger correlation over that of a baseline, it is important to take into account the uncertainty inherent in correlation point estimates prior to concluding improvements in metric performance. Confidence intervals for correlations with human judgment are rarely reported in metric evaluations, however, and when they have been reported, the most suitable methods have unfortunately not been applied. For example, incorrect assumptions about correlation sampling distributions made in past evaluations risk over-estimation of significant differences in metric performance. In this paper, we provide analysis of each of the issues that may lead to inaccuracies before providing detail of a method that overcomes previous challenges. Additionally, we propose a new method of translation sampling that in contrast achieves genuine high conclusivity in evaluation of the relative performance of metrics.

## 1 Introduction

In empirical evaluation of Machine Translation (MT), automatic metrics are widely used as a substitute for human assessment for the purpose of measuring differences in MT system performance. The performance of a newly proposed metric is itself measured by the degree to which its automatic scores for a sample of MT systems correlate with human assessment of that same set of systems. A main venue for evaluation of MT metrics is the annual Workshop for Statistical Machine Translation (WMT) (Bojar et al., 2015) where large-scale human evaluation takes place, primarily for the purpose of ranking systems competing in the translation shared task, but additionally to use the resulting system rankings for evaluation of automatic metrics. Since 2014, WMT has used the Pearson correlation as the official measure for evaluation of metrics (Macháček and Bojar, 2014; Stanojević et al., 2015). Comparison of the performance of any two metrics involves the comparison of two Pearson correlation point estimates computed over a sample of MT systems, therefore. Table 1 shows correlations with human assessment of each of the metrics participating in the Czech-to-English component of WMT-14 metrics shared task, and, for example, if we wish to compare the performance of the top-performing metric, REDSYSSENT (Wu et al., 2014), with the popular metric BLEU (Papineni et al., 2001), this involves comparison of the correlation point estimate of REDSYSSENT, $r = 0.993$, with the weaker correlation point estimate of BLEU, $r = 0.909$, with both computed with reference to human assessment of a sample of 5 MT systems.

When a new metric achieves a stronger correlation with human assessment over a baseline metric, such as the increase achieved by REDSYSSENT over BLEU, it is important to consider the uncertainty surrounding the difference in correlation. Confidence intervals are very rarely reported in metric evaluations, however, and when attempts have been made,



*Proceedings of NAACL-HLT 2016*, pages 1–10,
San Diego, California, June 12-17, 2016. 

| Metric | $r$ | CI | UCL |
|---|---|---|---|
| REDSYSSENT | 0.993 | ± 0.018 | 1.011 |
| REDSYS | 0.989 | ± 0.021 | 1.010 |
| NIST | 0.983 | ± 0.025 | 1.008 |
| DISCOTK-PARTY | 0.983 | ± 0.025 | 1.008 |
| APAC | 0.982 | ± 0.026 | 1.008 |
| METEOR | 0.980 | ± 0.029 | 1.009 |
| TER | 0.976 | ± 0.031 | 1.007 |
| DISCOTK-PARTY-TUNED | 0.975 | ± 0.031 | 1.006 |
| WER | 0.974 | ± 0.033 | 1.007 |
| CDER | 0.965 | ± 0.035 | 1.000 |
| TBLEU | 0.957 | ± 0.040 | 0.997 |
| DISCOTK-LIGHT | 0.954 | ± 0.038 | 0.992 |
| UPC-STOUT | 0.948 | ± 0.040 | 0.988 |
| BLEU-NRC | 0.946 | ± 0.044 | 0.990 |
| ELEXR | 0.945 | ± 0.044 | 0.989 |
| LAYERED | 0.941 | ± 0.045 | 0.986 |
| VERTA-EQ | 0.938 | ± 0.048 | 0.986 |
| VERTA-W | 0.934 | ± 0.050 | 0.984 |
| BLEU | 0.909 | ± 0.054 | 0.963 |
| PER | 0.883 | ± 0.063 | 0.946 |
| UPC-IPA | 0.824 | ± 0.073 | 0.897 |
| AMBER | 0.744 | ± 0.095 | 0.839 |

**Table 1:** WMT-14 Czech-to-English metrics shared task Pearson correlation ($r$) point estimates for metrics with human assessment (5 MT systems), reported confidence intervals (CI), and corresponding upper confidence limits (UCL).

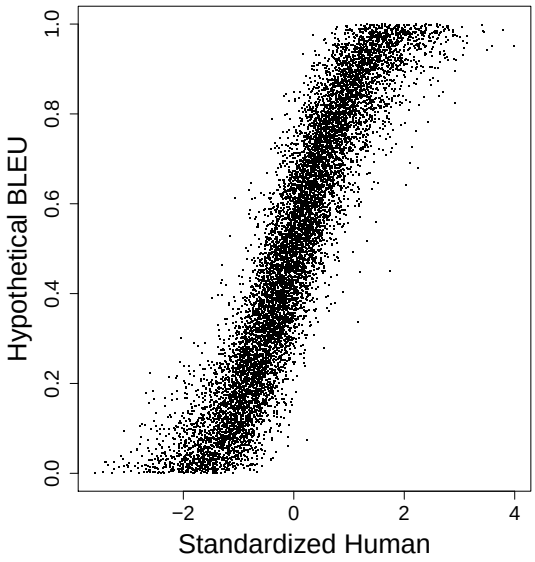


**Figure 1:** 10k simulated BLEU scores correlating with human assessment at $r = 0.91$ as in BLEU evaluation of Czech-to-English in WMT-14.

the most appropriate method has unfortunately not been applied. For example, although WMT constitutes a main authority on MT evaluation, and have made the best attempt to provide confidence intervals for metric correlations we could find, when we closely examine results of WMT-14 Czech-to-English metrics shared task, reproduced here in Table 1, a discrepancy can be identified. For the nine top-performing metrics participating in the shared task, upper confidence interval limits are reported to exceed 1.0.

Confidence intervals reported in the metrics shared task unfortunately also risk inaccurate conclusions about the relative performance of metrics for other less obvious reasons and risk conclusions that over-estimate the presence of significant differences. False-positives are problematic in metric evaluations because, if a given metric is mistakenly concluded to significantly outperform a competing metric, it is possible that had a larger sample of MT systems been employed in the evaluation, that the reverse conclusion should in fact be made. We demonstrate how this can occur for metrics, showing that in reality in current metric evaluation settings, it is only possible to identify a very small number of significant differences in performance. A main cause is the small number of MT systems employed in evaluations, and we propose a new sampling technique, hybrid super-sampling, that overcomes previous challenges and facilities the evaluation of metrics with reference to a practically unlimited number of MT systems.

## 2 WMT-style Evaluation

Alongside the correlation sample point estimates achieved by metrics, WMT reports confidence intervals for correlations that unfortunately risk overestimation of significant differences in metric performance, reasons for which we outline below (Macháček and Bojar, 2013; Macháček and Bojar, 2014; Stanojević et al., 2015).

### 2.1 Sampling Distribution Assumptions

As shown in Table 1, confidence intervals are reported for metric correlations using ± notation. The use of the ± notation implies that the sampling distribution is symmetrical. Since the sampling distribution of the Pearson correlation, $r$, is skewed, however, this means that, for a non-zero correlation, it is not possible for the portion of the confidence interval that lies above the correlation sample point estimate and the portion that lies below it to be equal. Ad-

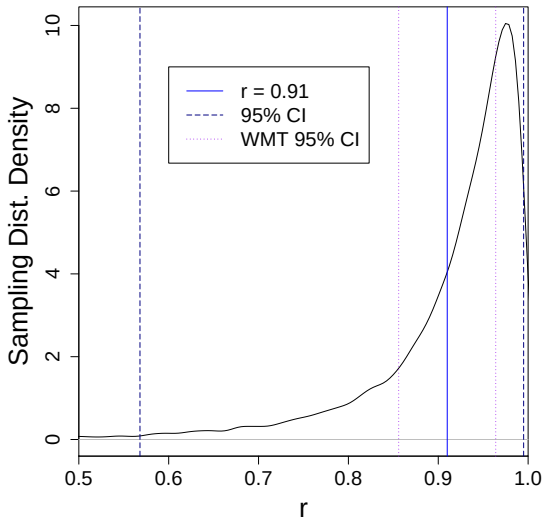


**Figure 2:** Sampling distribution of $r = 0.91$ and N = 5 for correlation of BLEU with human assessment for hypothetical "population" of MT systems in Figure 1.

ditionally, since the correlation sample statistic, $r$, cannot take on values greater than 1.0, the closer $r$ is to 1.0 the more extreme the skew of its sampling distribution becomes.[1]

To demonstrate how the skew of the sampling distribution of $r$ impacts on upper and lower confidence interval limits for metrics, in Figures 1 and 2, we simulate a possible population and sampling distribution for BLEU's correlation with human assessment, $r = 0.91$, achieved in WMT-14 Czech-to-English shared task, where the sample size, $n$, was 5 MT systems. Figure 1 depicts a hypothetical "population" of 10,000 MT systems and BLEU scores, where hypothetical BLEU scores for systems correspond with human assessment scores in such a way that a correlation of 0.91 is achieved. Figure 2 depicts the sampling distribution for $r$ yielded by repeatedly drawing sets of 5 systems at random from the larger "population" of 10,000 systems, where the negative skew can be clearly observed. Figure 2 also depicts the 95% confidence interval (CI), within which 95% of sampled correlations lie, where the width of the lower portion of confidence interval is substantially wider than the upper portion, and the overly conservative confidence interval reported for BLEU in WMT-14, where upper and lower portions of the confidence interval are incorrectly assumed to be equal in size.

### 2.2 Application of Bootstrap Resampling

A conventional approach to bootstrap resampling for the purpose of computing confidence intervals for a correlation sample point estimate is to create a correlation coefficient pseudo-distribution by sampling (at random with replacement) human and automatic scores for $n$ MT systems from the set of $n$ systems for which we have genuine metric and human scores. Instead, however, reported confidence intervals are the result of creating pseudo-distributions of human assessment scores for systems. The method unfortunately does not produce accurate confidence intervals for correlation sample point estimates, as confidence intervals produced in this way can unfortunately only inform us about the certainty surrounding human assessment scores for systems rather than the more relevant question of the certainty surrounding the correlation point estimates achieved by metrics. Confidence intervals computed in this way are substantially narrower than confidence intervals computed using the standard Fisher r-to-z transformation, that can also be directly applied to correlations of metrics with human assessment without application of randomized methods.

Table 2[2] includes reported confidence intervals of metric correlations for English-to-Czech in WMT-15, and those computed using the standard Fisher r-to-z transformation, where confidence intervals of the latter are substantially wider. An extreme example occurs for metric DREEM, where the difference between its original reported lower confidence interval limit and the correlation point estimate is 0.006, more than 34 times narrower than that computed with the Fisher r-to-z transformation, 0.206.

### 2.3 Difference in Dependent Correlations

When reporting the outcome of an empirical evaluation, along with sample point estimates, such as the mean or, in the case of metrics, correlation, we only

[1] It should be noted that the assumption of symmetry of the sampling distribution of $r$ is not explicitly made in any WMT report.

[2] WMT confidence intervals have been recomputed from the published data set to remove the previously described error with respect to the symmetry of $r$'s sampling distribution.

| Metric | $r$ | Method | Low. CI (-) | Upper CI (+) |
|---|---|---|---|---|
| CHRF3 | 0.977 | WMT | 0.003 | 0.002 |
| | | Fisher | 0.046 | 0.015 |
| CHRF | 0.971 | WMT | 0.003 | 0.003 |
| | | Fisher | 0.059 | 0.020 |
| RATATOUILLE | 0.965 | WMT | 0.003 | 0.003 |
| | | Fisher | 0.071 | 0.024 |
| BEER | 0.962 | WMT | 0.004 | 0.003 |
| | | Fisher | 0.076 | 0.026 |
| METEORWSD | 0.953 | WMT | 0.004 | 0.003 |
| | | Fisher | 0.093 | 0.032 |
| LEBLEU-DEF. | 0.953 | WMT | 0.004 | 0.003 |
| | | Fisher | 0.091 | 0.031 |
| BS | 0.953 | WMT | 0.004 | 0.003 |
| | | Fisher | 0.032 | 0.092 |
| BLEU | 0.936 | WMT | 0.005 | 0.004 |
| | | Fisher | 0.123 | 0.043 |
| PER | 0.908 | WMT | 0.005 | 0.004 |
| | | Fisher | 0.168 | 0.062 |
| DREEM | 0.883 | WMT | 0.006 | 0.006 |
| | | Fisher | 0.206 | 0.078 |

**Table 2:** WMT and Fisher r-to-z (Fisher) confidence intervals (CI) for Pearson correlation, $\rho$, in WMT-15 sample of English-to-Czech metrics (15 MT systems).

ever have access to a *sample* of the actual data that would be needed to compute the corresponding true value for the *population*. Confidence intervals provide a way of estimating the range of values within which we believe with a specified degree of certainty that the corresponding true value lies. Generally speaking, they can also provide a mechanism for drawing conclusions about significant differences in sample statistics. If, for example, mean scores are used to measure system performance, and the confidence intervals of a pair of systems do not overlap, a significant difference in sample means and subsequently system performance can be concluded.

Although confidence intervals for individual correlations do provide an indication of the degree of certainty with which we should interpret reported correlation sample point estimates, they unfortunately cannot be used in the above described way to conclude significant differences in the performance of metrics, however. All we can gain from confidence intervals for *individual* correlations with respect to significance differences is the following: if the confidence interval of a correlation sample point estimate does not include zero, then it can be concluded (with a specified degree of certainty) that this *individual* correlation is significantly different from *zero*. Confidence intervals for individual metric correlations with human assessment do not inform us about the certainty surrounding a *difference* in correlation with human assessment, the relevant question for comparing performance of competing MT metrics.

When computing confidence intervals for a difference in correlation, it is important to consider the nature of the data. For MT metric evaluation, data used to compute correlation point estimates for a given pair of metrics is *dependent*, as it includes three variables (Human, Metric$_a$, Metric$_b$), and, for each MT system that is a member of the sample, there is a value corresponding to each of these three variables. Besides the two correlations we are interested in comparing, $r$(Human, Metric$_a$) and $r$(Human, Metric$_b$), there is a third correlation to consider, therefore, the correlation that exists directly between the metric scores themselves, $r$(Metric$_a$, Metric$_b$). Graham and Baldwin (2014) provide detail of Williams test, a test of significance of a difference in *dependent* correlations, suitable for evaluation of MT metrics. Confidence intervals are more informative than the binary conclusions that can be inferred from p-values produced by significance tests, however, and Zou (2007) presents a method of constructing confidence intervals for differences in *dependent* correlations also suitable for evaluation of MT metrics. We provide an implementation of Zou (2007) tailored to metric evaluation at `https://github.com/ygraham/MT-metric-confidence-intervals`.

Table 3 includes confidence intervals for differences in dependent correlations (Zou, 2007) for the seven top-performing German-to-English metrics in WMT-15. Besides providing an indication of the degree of certainty surrounding a given difference in correlation for a pair of metrics, confidence intervals that do not include *zero* can now be used to infer a significant difference in performance for a pair of metrics. For example, the 95% confidence interval for the difference in correlation between the top-performing metric, UPFCOBALT ($r = 0.981$) and METEORWSD ($r = 0.953$), [0.005, 0.123], in Table 3, does not include zero and subsequently implies a significant difference in performance.

Figure 3 depicts the contrast in conclusions for

| | DPMFCOMB ($r = 0.973$) | DPMF ($r = 0.960$) | UOWLSTM ($r = 0.960$) | RATATOUILLE ($r = 0.958$) | CHRF3 ($r = 0.956$) | METEORWSD ($r = 0.953$) |
|---|---|---|---|---|---|---|
| ($r = 0.981$) UPFCOBALT | $[-0.023, 0.061]$ | $[-0.004, 0.101]$ | $[-0.013, 0.106]$ | $[-0.010, 0.109]$ | $[-0.001, 0.114]$ | [ **0.005, 0.123]** |
| DPMFCOMB | - | $[-0.025, 0.087]$ | $[-0.032, 0.092]$ | $[-0.026, 0.093]$ | $[-0.024, 0.101]$ | $[-0.017, 0.109]$ |
| DPMF | - | - | $[-0.070, 0.073]$ | $[-0.067, 0.075]$ | $[-0.061, 0.079]$ | $[-0.054, 0.087]$ |
| UOWLSTM | - | - | - | $[-0.071, 0.077]$ | $[-0.069, 0.084]$ | $[-0.066, 0.094]$ |
| RATATOUILLE | - | - | - | - | $[-0.072, 0.082]$ | $[-0.064, 0.088]$ |
| CHRF3 | - | - | - | - | - | $[-0.067, 0.081]$ |

**Table 3:** Pairwise 95% confidence intervals for differences in correlation for seven top-performing metrics for German-to-English in WMT-15 (13 MT systems), confidence intervals not including zero imply a significant difference and are highlighted in bold.

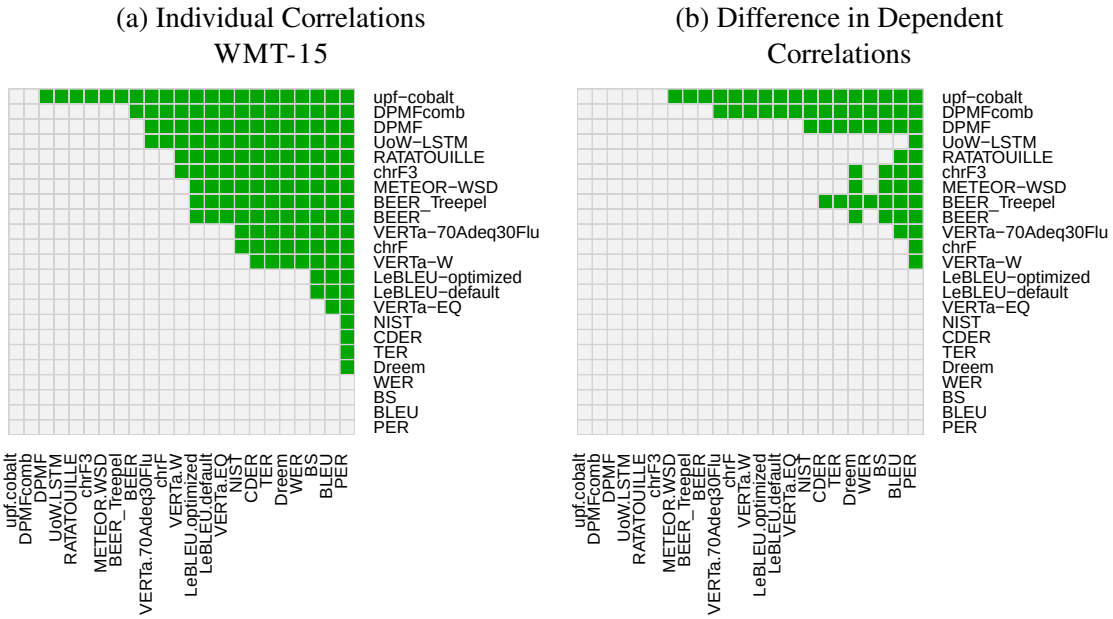


**Figure 3:** Conclusions of significant differences in correlation for WMT-15 German-to-English metrics (13 MT systems) drawn from the (a) non-overlap of individual correlation confidence intervals originally reported in WMT and from (b) confidence intervals of a difference in dependent correlations not including zero, green cells imply a significant win for the metric in that row over the metric in that column.

WMT-15 German-to-English metrics drawn from (a) a likely interpretation of confidence intervals originally reported in WMT, where the non-overlap of individual correlation confidence intervals of a pair of metrics is used to infer a significant difference, and (b) those drawn from the non-overlap of confidence intervals for differences in dependent correlations with zero (Zou, 2007), highlighting the over-estimation of significant differences in metric performance risked by current WMT confidence intervals. For example, for German-to-English with interpretation (a) a total of 91 significant differences are implied that are not identified according to our corresponding approach. For instance, the non-overlap of confidence intervals of the top-performing metric, UPFCOBALT, with those of all but one other metrics in the original report risks the interpretation of a significant increase in performance for that metric with all but one other competing metrics, but with the more appropriate method of Zou (2007), however, confidence intervals of this metric's difference in correlation with four of those competing metrics in fact include zero, with no significant difference identified. It is worth noting that original WMT reports *do not* state that the confidence intervals they provide *should* be interpreted in the way we have done here, where the non-overlap of individual correlation confidence intervals of a pair of metrics implies a significant difference, but this is nonetheless a very likely interpretation.

## 3 Accurate and Conclusive Metric Evaluations

Results of past metric evaluations have been highly inconclusive with relatively few significant differences in performance possible to identify for metrics.[3] The lack of conclusivity in metric evaluations is mainly caused by the small number of systems used to evaluate metrics. For example, in the original experiments used to justify the use of automatic metric BLEU, reported correlations with human assessment were for a sample size of as small as 5, comprising three automatic systems and two human translators (Papineni et al., 2001). WMT have improved on this for some language pairs at least, as in the past four evaluations sample sizes have ranged from 5 (Czech-to-English WMT-14) to 22 systems (German-to-English WMT-12/WMT-13). Even at the maximum sample size of 22 systems, however, correlation point estimates are computed with a high degree of uncertainty.

### 3.1 Hybrid Super-Sampling

In an ideal world, MT metric evaluations would employ a much larger sample of systems than those relied upon in past evaluations, subsequently yielding correlation sample point estimates that can be relied on with more certainty. Although not immediately obvious, data sets currently used to evaluate MT metrics potentially contain data for a very large number of systems. If we consider the fact that, given the output of as little as two MT systems, there exists a very large number of possible ways of combining their translated segments to form a hybrid system, this opens up the evaluation of metrics to a vastly larger pool of systems. For example, even if we restrict the creation of hybrid systems to combinations of *pairs* of the $n$ MT systems competing in a translation shared task (as opposed to hybrids created by sampling translations from *several* different MT systems at once), the number of potential hybrid systems is exponential in size of the test set, $m$:

$$n(n-1)/2 \cdot 2^m \quad (1)$$

For instance, even for a language pair for which human scores are available for as few as 5 MT systems, by *super-sampling* translations from every pair of competing systems, this results in 10 x $2^{3,000}$ hybrid systems. Including all possible hybrid systems is of course not necessary, and to make the approach feasible, we sample a large but manageable subset of 10,000 MT systems.

Obtaining automatic metric scores for this larger number of MT systems is feasible for any metric that is expected to be useful in practice, since automatic metrics must already be highly efficient to be used for optimizing systems. Obtaining human assessment of this large set of hybrid systems may seem

[3] Due to space limitations, it was only possible to include confidence intervals for differences in correlation for a subset of German-to-English WMT-15 metrics (Figure 3). Confidence intervals for the the remaining metrics and language pairs are available at `https://github.com/ygraham/MT-metric-confidence-intervals` for which very few significant differences in performance are identified.

| Metric | $r$ | CI of Difference in $r$ with next best metric | $r$ | CI of Difference in $r$ with next best metric |
|---|---|---|---|---|
| TERRORCAT | 0.971 | [ −0.019 , 0.155 ] | 0.960 | [ **0.028 , 0.030** ] |
| SAGANSTS | 0.942 | [ −0.120 , 0.136 ] | 0.932 | [ **0.006 , 0.011** ] |
| METEOR | 0.938 | [ −0.086 , 0.172 ] | 0.923 | [ **0.028 , 0.032** ] |
| POSF | 0.919 | [ −0.134 , 0.184 ] | 0.893 | [ **0.004 , 0.008** ] |
| SPEDE07FP | 0.907 | [ −0.138 , 0.162 ] | 0.887 | [ −0.001 , 0.003 ] |
| • SPEDE08FP | 0.897 | [ −0.142 , 0.202 ] | 0.886 | [ **0.004 , 0.007** ] |
| • SPEDE07F | 0.902 | [ −0.156 , 0.176 ] | 0.880 | [ **0.003 , 0.006** ] |
| • SPEDE07PP | 0.879 | [ −0.161 , 0.202 ] | 0.876 | [ **0.007 , 0.007** ] |
| • SPEDE07P | 0.870 | [ −0.188 , 0.196 ] | 0.869 | [ **0.006 , 0.009** ] |
| • XENERRCATS | 0.884 | [ −0.174 , 0.193 ] | 0.862 | [ **0.011 , 0.015** ] |
| • AMBER | 0.859 | [ −0.084 , 0.398 ] | 0.849 | [ **0.008 , 0.011** ] |
| • WORDBLOCKERRCATS | 0.868 | [ −0.183 , 0.220 ] | 0.839 | [ **0.057 , 0.065** ] |
| • SIMPBLEU | 0.770 | [ −0.210 , 0.318 ] | 0.778 | [ **0.033 , 0.036** ] |
| • BLEU | 0.741 | - | 0.744 | [ **0.008 , 0.016** ] |
| • BLOCKERRCATS | 0.779 | [ −0.257 , 0.293 ] | 0.731 | - |
| | | 12 Systems | | 10k Systems |

**Table 4:** Correlations and confidence intervals of pseudo document-level metrics (averaged segment-level metrics) with human assessment evaluated on original 12 MT systems and 10k hybrid super-sample (WMT-12 Spanish-to-English). Metrics with a different rank order in the original sample and hybrid super-sample evaluations are marked with • and confidence intervals that do not include zero are in bold.

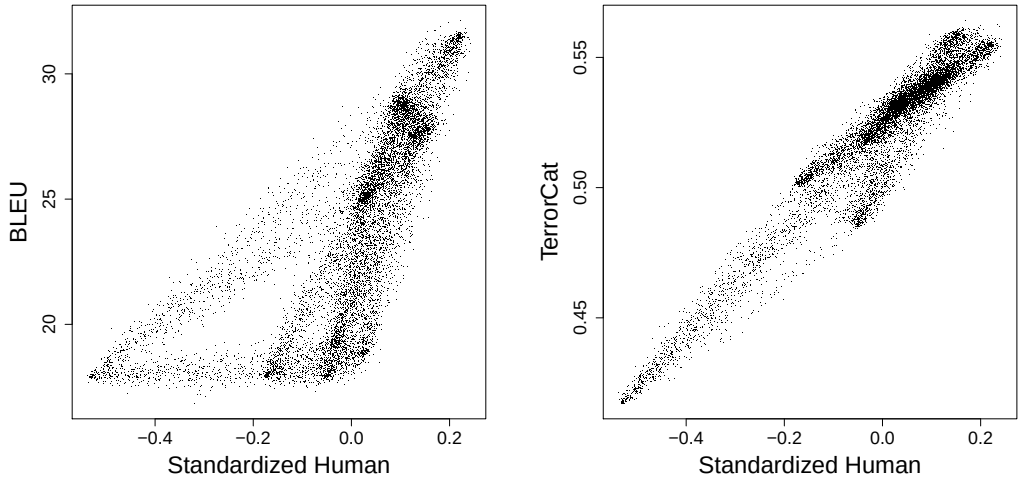


**Figure 4:** Human, TERRORCAT and BLEU scores for 10k super-sampled hybrid MT systems for WMT-12 Spanish-to-English.

more challenging, but the method of human evaluation we employ facilitates the straight-forward computation of human scores for vast numbers of systems directly from the original human evaluation of only $n$ systems. Graham et al. (2013) provide a human evaluation of MT that elicits adequacy assessments of translations, independent of other translations on a fine-grained 100-point rating scale. After score standardization to iron-out differences in individual human assessor scoring strategies, the overall human score for a MT system is simply computed as the mean of the ratings attributed to its translations, and this facilitates the straight-forward computation of a human score for any hybrid system from the original human evaluation of $n$ systems.

To demonstrate, we replicate a previous year's WMT metrics shared task, constructing a hybrid super-sample of 10,000 MT systems each with a corresponding metric and human score. Since we do not have access to all document-level metrics that participated in the original shared task, we use segment-level metric scores as *pseudo document-level metrics* by taking the average of segment-level scores of the segments that comprise the test set document. This allows retrospective computation of automatic metric scores for the large set of 10k hybrid MT systems. For the purpose of comparison, in addition to averaged segment-level metrics, we also include document-level BLEU and an analysis of the correlation it achieves in the context of hybrid super-sampling. Human evaluation scores were computed using the mean of a minimum of 1,500 crowd-sourced human ratings per system, where strict quality-controlling of crowd-sourced workers was applied.

Table 4 shows correlations achieved by metrics when evaluated on the original 12 and 10k systems, as well as confidence intervals of the difference in correlation achieved by each metric with that of the next best performing metric in each case.[4] As expected, confidence intervals for differences in correlation are substantially reduced for the larger sample of metrics. Importantly, the change in rank order of metrics when evaluated with reference to a sample of 10k MT systems, as opposed to 12, highlights the risk of concluding an increase in performance from evaluations that include only a small sample of systems.

Figure 4 plots super-sampled human and automatic metric scores for BLEU providing insight into how BLEU scores correspond with human assessment. Worryingly for the range of systems with scores below 20 BLEU points, the plot shows an almost horizontal band of systems spread across a wide range of quality according to human assessors despite extremely similar BLEU scores. The top-performing automatic metric, TERRORCAT, on the other hand, impressively sustains its high correlation with human assessment when evaluated on as many as 10k MT systems, evidence that this metric is indeed highly consistent with human assessment of Spanish-to-English.

Due to space limitations, it is not possible to include pairwise confidence intervals for all pairs of metrics, and instead we include in Figure 5 a heatmap of significant differences in performance, where a significant win is inferred for the metric in a given row over the metric in a given column if the confidence interval of the difference in correlation for that pair did not include zero. Results show the super-sampled evaluation facilitates not only the identification of an outright best-performing metric, TERRORCAT, it also yields an almost total-order ranking of metrics, as significant differences are possible to identify for all but one pairs of competing metrics. Finally, we repeated the metric evaluation with ten distinct super-samples of 10k MT systems with all replications resulting in precisely the same ranking of metrics as shown in Table 4.

## 4 Conclusion

Analysis of evaluation methodologies applied to automatic MT metrics was provided and the risk of over-estimation of significant differences in metric performance identified. Confidence intervals for differences in dependent correlations were recommended as appropriate for evaluation of MT metrics. Hybrid super-sampling was proposed, evaluating metrics with reference to a substantially larger sample of MT systems, achieving genuinely highly conclusive metric rankings.

[4] It should be noted, since participating teams did not intend segment-level metric scores to be averaged as we have done here, correlations are for demonstrative purposes and do not reflect performance of participating teams.

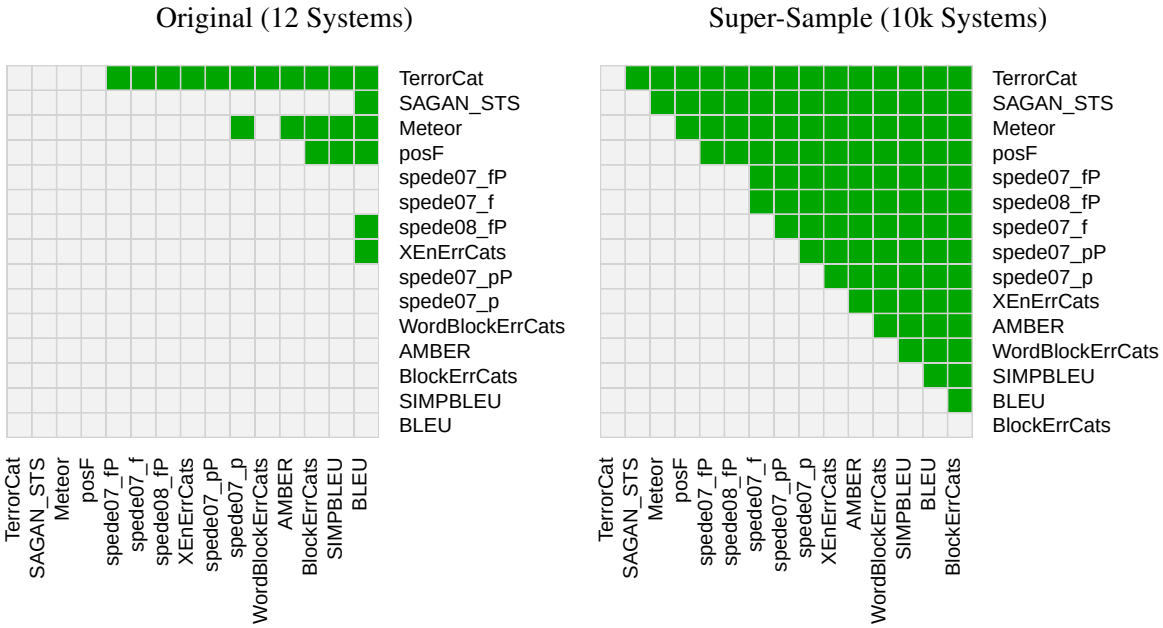


**Figure 5:** Pairwise conclusions for pseudo document-level metrics (averaged segment-level metrics) from WMT-12 Spanish-to-English metrics shared task, where a green cell indicates a significant win for the metric in a given row over the metric in the corresponding column.

## Acknowledgments

We wish to thank the anonymous reviewers and Ondřej Bojar for valued feedback and WMT organisers for provision of data sets. This project has received funding from the European Union Horizon 2020 research and innovation programme under grant agreement 645452 (QT21) and the ADAPT Centre for Digital Content Technology (`www.adaptcentre.ie` ) at Dublin City University funded under the SFI Research Centres Programme (Grant 13/RC/2106) co-funded under the European Regional Development Fund.

## References

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Barry Haddow, Matthias Huck, Chris Hokamp, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Carolina Scarton, Lucia Specia, and Marco Turchi. 2015. Findings of the 2015 Workshop on Statistical Machine Translation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 1–46, Lisbon, Portugal, September. Association for Computational Linguistics.

Yvette Graham and Timothy Baldwin. 2014. Testing for significance of increased correlation with human judgment. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 172–176, Doha, Qatar. Association for Computational Linguistics.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2013. Continuous measurement scales in human evaluation of machine translation. In *Proceedings of the 7th Linguistic Annotation Workshop & Interoperability with Discourse*, pages 33–41, Sofia, Bulgaria. Association for Computational Linguistics.

Matouš Macháček and Ondřej Bojar. 2013. Results of the WMT13 metrics shared task. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 45–51, Sofia, Bulgaria, August. Association for Computational Linguistics.

Matouš Macháček and Ondřej Bojar. 2014. Results of the WMT14 metrics shared task. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 293–301, Baltimore, Maryland, USA, June. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2001. BLEU: A method for automatic evaluation of machine translation. Technical Report RC22176 (W0109-022), IBM Research, Thomas J. Watson Research Center.

Miloš Stanojević, Amir Kamran, Philipp Koehn, and Ondřej Bojar. 2015. Results of the WMT15 metrics shared task. In *Proceedings of the Tenth Workshop on*

*Statistical Machine Translation*, pages 256–273, Lisbon, Portugal, September. Association for Computational Linguistics.

Xiaofeng Wu, Hui Yu, and Qun Liu. 2014. RED, the DCU-CASICT submission of metrics tasks. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 420–425, Baltimore, Maryland, USA, June. Association for Computational Linguistics.

Guang Yong Zou. 2007. Toward using confidence intervals to compare correlations. *Psychological Methods*, 12(4):399 – 413.