Tangled up in BLEU: Reevaluating the Evaluation of Automatic Machine Translation Evaluation Metrics

Nitika Mathur **Timothy Baldwin** **Trevor Cohn**
School of Computing and Information Systems
The University of Melbourne
Victoria 3010, Australia
nmathur@student.unimelb.edu.au {tbaldwin,tcohn}@unimelb.edu.au

Abstract

Automatic metrics are fundamental for the development and evaluation of machine translation systems. Judging whether, and to what extent, automatic metrics concur with the gold standard of human evaluation is not a straightforward problem. We show that current methods for judging metrics are highly sensitive to the translations used for assessment, particularly the presence of outliers, which often leads to falsely confident conclusions about a metric's efficacy. Finally, we turn to pairwise system ranking, developing a method for thresholding performance improvement under an automatic metric against human judgements, which allows quantification of type I versus type II errors incurred, i.e., insignificant human differences in system quality that are accepted, and significant human differences that are rejected. Together, these findings suggest improvements to the protocols for metric evaluation and system performance evaluation in machine translation.

1 Introduction

Automatic metrics are an indispensable part of machine translation (MT) evaluation, serving as a proxy to human evaluation which is considerably more expensive and time-consuming. They provide immediate feedback during MT system development and serve as the primary metric to report the quality of MT systems. Accordingly, the reliability of metrics is critical to progress in MT research.

A particularly worrying finding was made in the most recent Conference on Machine Translation (WMT), as part of their annual competition findings to benchmark progress in translation and translation evaluation. WMT has established a method based on Pearson's correlation coefficient for measuring how well automatic metrics match with human judgements of translation quality, which is used to rank metrics and to justify their widespread use in lieu of human evaluation. Their findings (Ma et al., 2019) showed that if the correlation is computed for metrics using a large cohort of translation systems, typically very high correlations were found between leading metrics and humans (as high as $r = 0.9$). However, if considering only the few best systems, the correlation reduced markedly. This is in contrast to findings at sentence-level evaluation, where metrics are better at distinguishing between high-quality translations compared to low-quality translations (Fomicheva and Specia, 2019).

When considering only the four best systems, the automatic metrics were shown to exhibit negative correlations in some instances. It would appear that metrics can only be relied upon for making coarse distinctions between poor and good translation outputs, but not for assessing similar quality outputs, i.e., the most common application faced when assessing incremental empirical improvements.

Overall these findings raise important questions as to the reliability of the accepted best-practises for ranking metrics, and more fundamentally, cast doubt over these metrics' utility for tuning high-quality systems, and making architecture choices or publication decisions for empirical research.

In this paper, we take a closer look into this problem, using the metrics data from recent years of WMT to answer the following questions:

1. Are the above problems identified with Pearson's correlation evident in other settings besides small collections of strong MT systems? To test this we consider a range of system quality levels, including random samples of systems, and show that the problem is widely apparent.

2. What is the effect of outlier systems in the reported correlations? Systems that are considerably worse than all others can have a dispro-

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 4984–4997
July 5 - 10, 2020. ©2020 Association for Computational Linguistics

portionate effect on the computed correlation, despite offering very little insight into the evaluation problem. We identify a robust method for identifying outliers, and demonstrate their effect on correlation, which for some metrics can result in radically different conclusions about their utility.

3. Given these questions about metrics' utility, can they be relied upon for comparing two systems? More concretely, we seek to quantify the extent of improvement required under an automatic metric such that the ranking reliably reflects human assessment. In doing so, we consider both type I and II errors, which correspond to accepting negative or insignificant differences as judged by humans, versus rejecting human significant differences; both types of errors have the potential to stunt progress in the field.

Overall we find that current metric evaluation methodology can lend false confidence to the utility of a metric, and that leading metrics require either untenably large improvements to serve a gatekeeping role, or overly permissive usage to ensure good ideas are not rejected out of hand. Perhaps unsurprisingly, we conclude that metrics are inadequate as a substitute for human evaluations in MT research. [1]

2 Related work

Since 2007, the Conference on Machine Translation (WMT) has organized an annual shared task on automatic metrics, where metrics are evaluated based on correlation with human judgements over a range of MT systems that were submitted to the translation task. Methods for both human evaluation and meta evaluation of metrics have evolved over the years.

In early iterations, the official evaluation measure was the Spearman's rank correlation of metric scores with human scores (Callison-Burch and Osborne, 2006). However, many MT system pairs have very small score differences, and evaluating with Spearman's correlation harshly penalises metrics that have a different ordering for these systems. This was replaced by the Pearson correlation in 2014 (Bojar et al., 2014). To test whether the difference in the performance of two metrics is statistically significant, the William's test for dependent correlations is used (Graham and Baldwin, 2014), which takes into account the correlation between the two metrics. Metrics that are not outperformed by any other metric are declared as the winners for that language pair.

Pearson's r is highly sensitive to outliers (Osborne and Overbay, 2004): even a single outlier can have a drastic impact on the value of the correlation coefficient; and in the extreme case, outliers can give the illusion of a strong correlation when there is none, or mask the presence of a true relationship. More generally, very different underlying relationships between the two variables can have the same value of the correlation coefficient (Anscombe, 1973).[2]

The correlation of metrics with human scores is highly dependent on the underlying systems used. BLEU (Papineni et al., 2002a) has remained mostly unchanged since it was proposed in 2002, but its correlation with human scores has changed each year over ten years of evaluation (2006 to 2016) on the English–German and German–English language pairs at WMT (Reiter, 2018). The low correlation for most of 2006–2012 is possibly due to the presence of strong rule-based systems that tend to receive low BLEU scores (Callison-Burch and Osborne, 2006). By 2016, however, there were only a few submissions of rule-based systems, and these were mostly outperformed by statistical systems according to human judgements (Bojar et al., 2016). The majority of the systems in the last three years have been neural models, for which most metrics have a high correlation with human judgements.

BLEU has been surpassed by various other metrics at every iteration of the WMT metrics shared task. Despite this, and extensive analytical evidence of the limitations of BLEU in particular and automatic metrics in general (Stent et al., 2005; Callison-Burch and Osborne, 2006; Smith et al., 2016), the metric remains the de facto standard of evaluating research hypotheses.

[1] Code, data and additional analysis available at https://github.com/nitikam/tangled

[2] `https://janhove.github.io/teaching/2016/11/21/what-correlations-look-like` contains examples that clearly illustrate the extent of this phenomenon

3 Data

3.1 Direct Assessment (DA)

Following Ma et al. (2019), we use direct assessment (DA) scores (Graham et al., 2017) collected as part of the human evaluation at WMT 2019. Annotators are asked to rate the adequacy of a set of translations compared to the corresponding source/reference sentence on a slider which maps to a continuous scale between 0 and 100. Bad quality annotations are filtered out based on quality control items included in the annotation task. Each annotator's scores are standardised to account for different scales. The score of an MT system is computed as the mean of the standardised score of all its translations. In WMT 19, typically around 1500–2500 annotations were collected per system for language pairs where annotator availability was not a problem. To assess whether the difference in scores between two systems is not just chance, the Wilcoxon rank-sum test is used to test for statistical significance.

3.2 Metrics

Automatic metrics compute the quality of an MT output (or set of translations) by comparing it with a reference translation by a human translator. For the WMT 19 metrics task, participants were also invited to submit metrics that rely on the source instead of the reference (QE . In this paper, we focus on the following metrics that were included in evaluation at the metrics task at WMT 2019:

Baseline metrics

- BLEU (Papineni et al., 2002b) is the precision of n-grams of the MT output compared to the reference, weighted by a brevity penalty to punish overly short translations. BLEU has high variance across different hyper-parameters and pre-processing strategies, in response to which sacreBLEU (Post, 2018) was introduced to create a standard implementation for all researchers to use; we use this version in our analysis.
- TER (Snover et al., 2006) measures the number of edits (insertions, deletions, shifts and substitutions) required to transform the MT output to the reference.
- CHRF (Popović, 2015) uses character n-grams instead of word n-grams to compare the MT output with the reference. This helps with matching morphological variants of words.

Best metrics across language pairs

- YISI-1 (Lo, 2019) computes the semantic similarity of phrases in the MT output with the reference, using contextual word embeddings (BERT: Devlin et al. (2019)).
- ESIM (Chen et al., 2017; Mathur et al., 2019) is a trained neural model that first computes sentence representations from BERT embeddings, then computes the similarity between the two strings. [3]

Source-based metric

- YISI-2 (Lo, 2019) is the same as YISI-1, except that it uses cross-lingual embeddings to compute the similarity of the MT output with the source.

The baseline metrics, particularly BLEU, were designed to use multiple references. However, in practice, they have only have been used with a single reference in recent years.

4 Re-examining conclusions of Metrics Task 2019

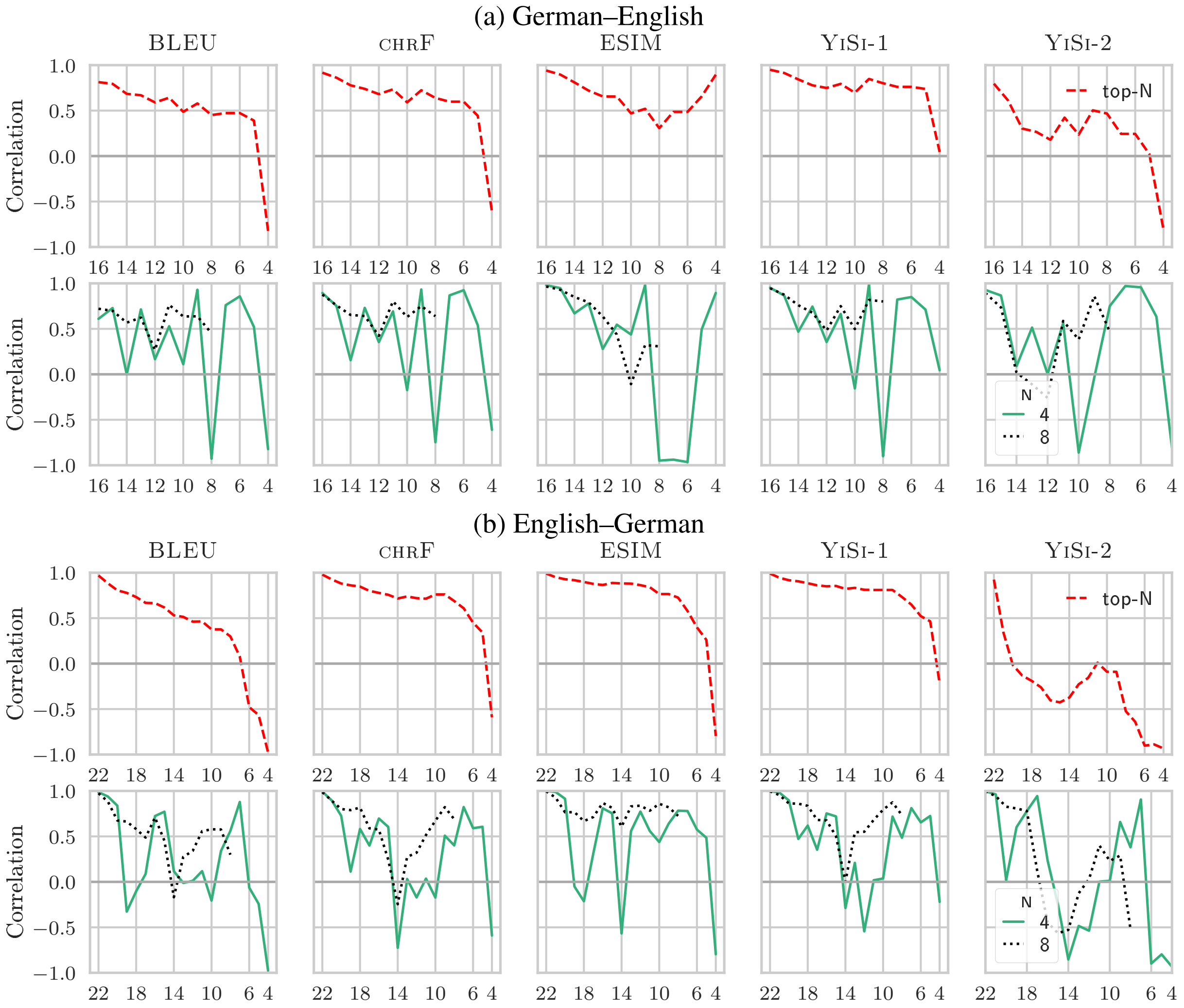
4.1 Are metrics unreliable when evaluating high-quality MT systems?

In general, the correlation of reference-based metrics with human scores is greater than $r = 0.8$ for all language pairs. However, the correlation is dependent on the systems that are being evaluated, and as the quality of MT increases, we want to be sure that the metrics evaluating these systems stay reliable.

To estimate the validity of the metrics for high-quality MT systems, Ma et al. (2019) sorted the systems based on their Direct Assessment scores, and plotted the correlation of the top N systems, with N ranging from all systems to the best four systems. They found that for seven out of 18 language pairs, the correlation between metric and human scores decreases as we decrease N, and tends towards zero or even negative when $N = 4$.

There are four language pairs (German–English, English–German, English–Russian, and English–Chinese) where the quality of the best MT systems is close to human performance (Barrault et al., 2019). If metrics are unreliable for strong MT systems, we would expect to see a sharp degradation in correlation for these language pairs. But as

[3] ESIM's submission to WMT shared task does not include scores for the language pairs en-cs and en-gu. In this paper, we use scores obtained from the same trained model that was used in the original submission.



(a) German–English

(b) English–German

Figure 1: Pearson correlation coefficient computed over the top-N systems (top row), or over a rolling window of 4 or 8 systems (bottom row). The x axis shows the index of the starting system, and systems are sorted by DA quality score.

we look at the top N systems, the correlation decreases for German–English and English–German, stays the same for English–Russian, and actually increases for English–Chinese. On the other hand, we observe this phenomenon with English–Kazakh, where the top systems are far from the quality of human translation.

Is there another explanation for these results? Pearson's r between metrics and DA scores is unstable for small samples, particularly when the systems are very close in terms of quality. The low correlation over top-N systems (when N is small) could be an artefact of this instability. To understand this effect, we instead visualise the correlation of a rolling window of systems, starting with the worst N systems, and moving forward by one system until we reach the top N systems. The number of systems stays constant for all points in these graphs, which makes for a more valid comparison than the original setting where the sample size varies. If the metrics are indeed less reliable for strong systems, we should see the same pattern as with the top N systems.

For the German–English language pair (Figure 1 b), the correlation of most metrics is very unstable when $N = 4$. Both BLEU and CHRF perfectly correlate with human scores for systems ranked 2–5, which then drops to -1 for the top 4 systems. On the other hand, ESIM exhibits the opposite behaviour, even though it shows an upward trend when looking at the top-N systems.

Even worse, for English–German, YISI-2 obtains a perfect correlation at some values of N, when in fact its correlation with human scores is negligible once outliers are removed (Section 4.2).

We observe similar behaviour across all language pairs: the correlation is more stable as N increases, but there is no consistent trend in the correlation that depends on the quality of the systems in the sample.

If we are to trust Pearson's r at small sample sizes, then the reliability of metrics doesn't really depend on the quality of the MT systems. Given that the sample size is small to begin with (typically 10–15 MT systems per language pair), we believe that we do not have enough data to use this method to assess whether metric reliability decreases with the quality of MT systems.

A possible explanation for the low correlation of subsets of MT systems is that it depends on how close these systems are in terms of quality. In the extreme case, the difference between the DA scores of all the systems in the subset can be statistically insignificant, so metric correlation over these systems can be attributed to chance.

4.2 How do outliers affect the correlation of MT evaluation metrics?

An outlier is defined as "an observation (or subset of observations) which appears to be inconsistent with the remainder of the dataset" (Barnett and Lewis, 1974). Pearson's r is particularly sensitive to outliers in the observations. When there are systems that are generally much worse (or much better) than the rest of the systems, metrics are usually able to correctly assign low (or high) scores to these systems. In this case, the Pearson correlation can over-estimate metric reliability, irrespective of the relationship between human and metric scores of other systems.

Based on a visual inspection, we can see there are two outlier systems in the English–German language pair. To illustrate the influence of these systems on Pearson's r, we repeatedly subsample ten systems from the 22 system submissions (see Figure 2). When the most extreme outlier (`en-de-task`) is present in the sample, the correlation of all metrics is greater than 0.97. The selection of systems has a higher influence on the correlation when neither outlier is present, and we can see that YiSi-1 and ESIM usually correlate much higher than BLEU.

One method of dealing with outliers is to calculate the correlation of the rest of the points (called the skipped correlation: Wilcox (2004)). Most of these apply methods to detect multivariate outliers in the joint distribution of the two variables: the metric and human scores in our case. However, multivariate outliers could be system pairs that indicate metric errors, and should not be removed because they provide important data about the metric.

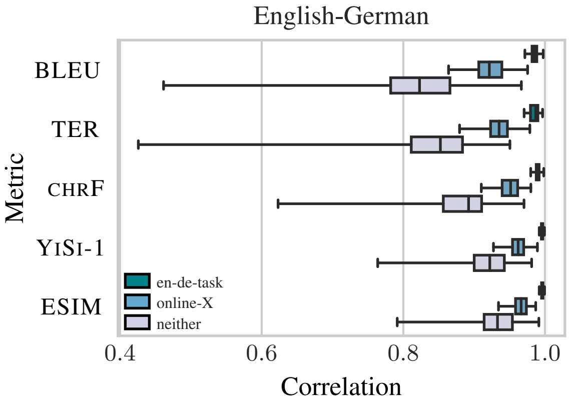


Figure 2: Pearson's r for metrics, when subsampling systems from the English–German language pair. We group the samples in the presence of the two outliers ("`en-de-task`" and "`Online-X`"), and when neither is present.

Thus, we only look towards detecting univariate outliers based on human ratings. One common method is to simply standardise the scores, and remove systems with scores that are too high or too low. However, standardising depends on the mean and standard deviation, which are themselves affected by outliers. Instead, we use the median and the Median Absolute Deviation (MAD) which are more robust (Iglewicz and Hoaglin, 1993; Rousseeuw and Hubert, 2011; Leys et al., 2013).

For MT systems with human scores s, we use the following steps to detect outlier systems:

1. Compute MAD, which is the median of all absolute deviations from the median

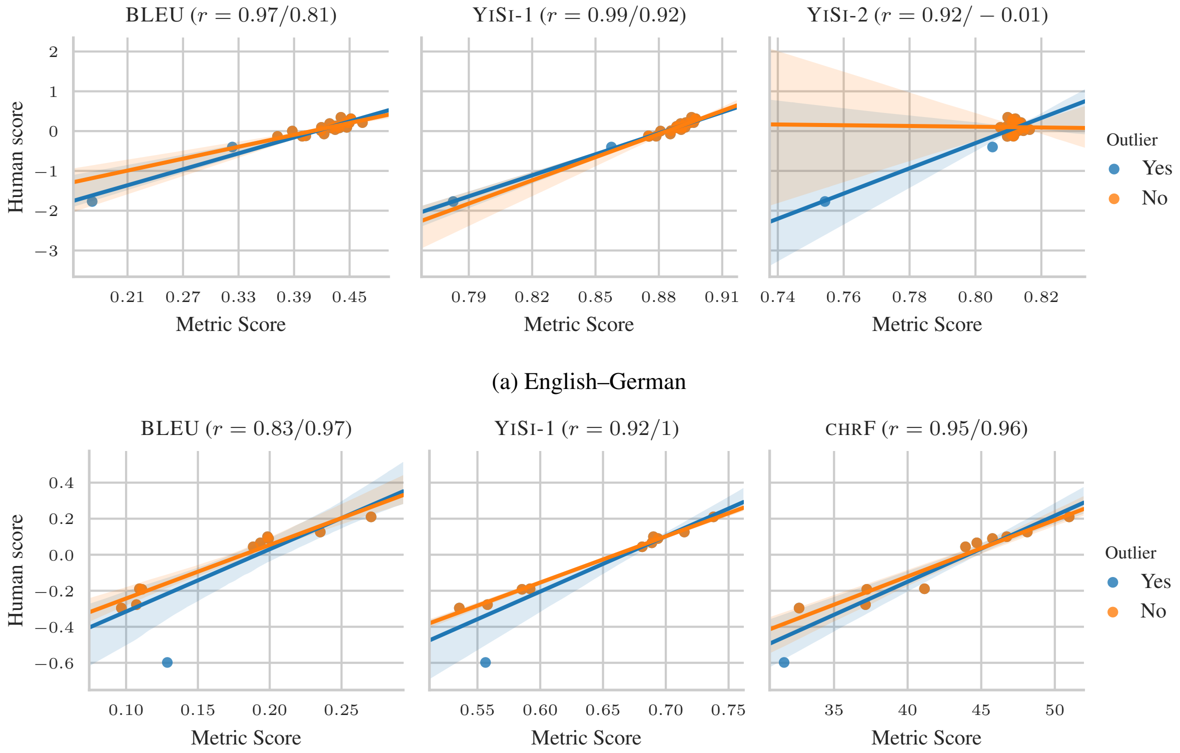
$$\text{MAD} = 1.483 \times \text{median}(|s - \text{median}(s)|)$$

2. compute robust scores:

$$z = (s - \text{median}(s))/\text{MAD}$$

3. discard systems where the magnitude of z exceeds a cutoff (we use 2.5)

Tables 1 and 2 show Pearson's r with and without outliers for the language pairs that contain outliers. Some interesting observations, are as follows:



(a) English–German

(b) Gujarati–English

Figure 3: Scatter plots (and Pearson's r) for metrics with and without outliers

	de–en		gu–en		kk–en		lt–en		ru–en		zh–en	
	All	−out	All	−out	All	−out	All	−out	All	−out	All	−out
#sys	16	15	11	10	11	9	11	10	14	13	15	13
BLEU	0.81	0.79	0.83	0.97	0.95	0.91	0.96	0.97	0.87	0.81	0.90	0.81
TER	0.87	0.81	0.89	0.95	0.80	0.57	0.96	0.98	0.92	0.90	0.84	0.72
chrF	0.92	0.86	0.95	0.96	0.98	0.77	0.94	0.93	0.94	0.88	0.96	0.84
ESIM	0.94	0.90	0.88	0.99	0.99	0.95	0.99	0.99	0.97	0.95	0.99	0.96
YiSi-1	0.95	0.91	0.92	1.00	0.99	0.92	0.98	0.98	0.98	0.95	0.98	0.90
YiSi-2	0.80	0.61	−0.57	0.82	−0.32	0.66	0.44	0.35	−0.34	0.71	0.94	0.62

Table 1: Correlation of metrics with and without outliers ("All" and "−out", resp.) for the to-English language pairs that contain outlier systems

	de–cs		en–de		en–fi		en–kk		en–ru		fr–de	
	All	−out	All	−out	All	−out	All	−out	All	−out	All	−out
#sys	11	10	22	20	12	11	11	9	12	11	10	7
BLEU	0.87	0.74	0.97	0.81	0.97	0.94	0.85	0.58	0.98	0.95	0.87	0.85
TER	0.89	0.79	0.97	0.84	0.98	0.96	0.94	0.55	0.99	0.98	0.89	0.67
chrF	0.97	0.97	0.98	0.88	0.99	0.97	0.97	0.90	0.94	0.97	0.86	0.80
ESIM	0.98	0.99	0.99	0.93	0.96	0.93	0.98	0.90	0.99	0.99	0.94	0.83
YiSi-1	0.97	0.98	0.99	0.92	0.97	0.94	0.99	0.89	0.99	0.98	0.91	0.85
YiSi-2	0.61	0.12	0.92	−0.01	0.70	0.48	0.34	0.69	−0.77	0.13	−0.53	0.07

Table 2: Correlation of metrics with and without outliers ("All" and "−out", resp.) for the language pairs into languages other than English that contain outlier systems.

- for language pairs like Lithuanian–English and English–Finnish, the correlation between the reference based metrics and DA is high irrespective of the presence of the outlier;
- the correlation of BLEU with DA drops sharply from 0.85 to 0.58 for English–Kazakh when outliers are removed;
- for English–German, the correlation of BLEU and TER appears to be almost as high as that of YiSi-1 and ESIM. However, when we remove the two outliers, there is a much wider gap between the metrics.
- if metrics wrongly assign a higher score to an outlier (e.g. most metrics in Gujarat–English), removing these systems increases correlation, and reporting only the skipped correlation is not ideal.

To illustrate the severity of the problem, we show examples from the metrics task data where outliers present the illusion of high correlation when the metric scores are actually independent of the human scores without the outlier. For English–German, the source-based metric YiSi-2 correctly assigns a low score to the outlier `en-de-task`. When this system is removed, the correlation is near zero. At the other extreme, YiSi-2 incorrectly assigns a very high score to a low-quality outlier in the English–Russian language pair, resulting in a strongly negative correlation. When we remove this system, we find there is no association between metric and human scores.

The results for all metrics that participated in the WMT 19 metrics task are presented in Tables 3, 4 and 5 in the appendix.

5 Beyond correlation: metric decisions for system pairs

In practice, researchers use metric scores to compare pairs of MT systems, for instance when claiming a new state of the art, evaluating different model architectures, or even in deciding whether to publish. Basing these judgements on metric score alone runs the risk of making wrong decisions with respect to the true gold standard of human judgements. That is, while a change may result in a significant improvement in BLEU, this may not be judged to be an improvement by human assessors.

Thus, we examine whether metrics agree with DA on all the MT systems pairs across all languages used in WMT 19.

Following Graham et al. (2014), we use statisti-

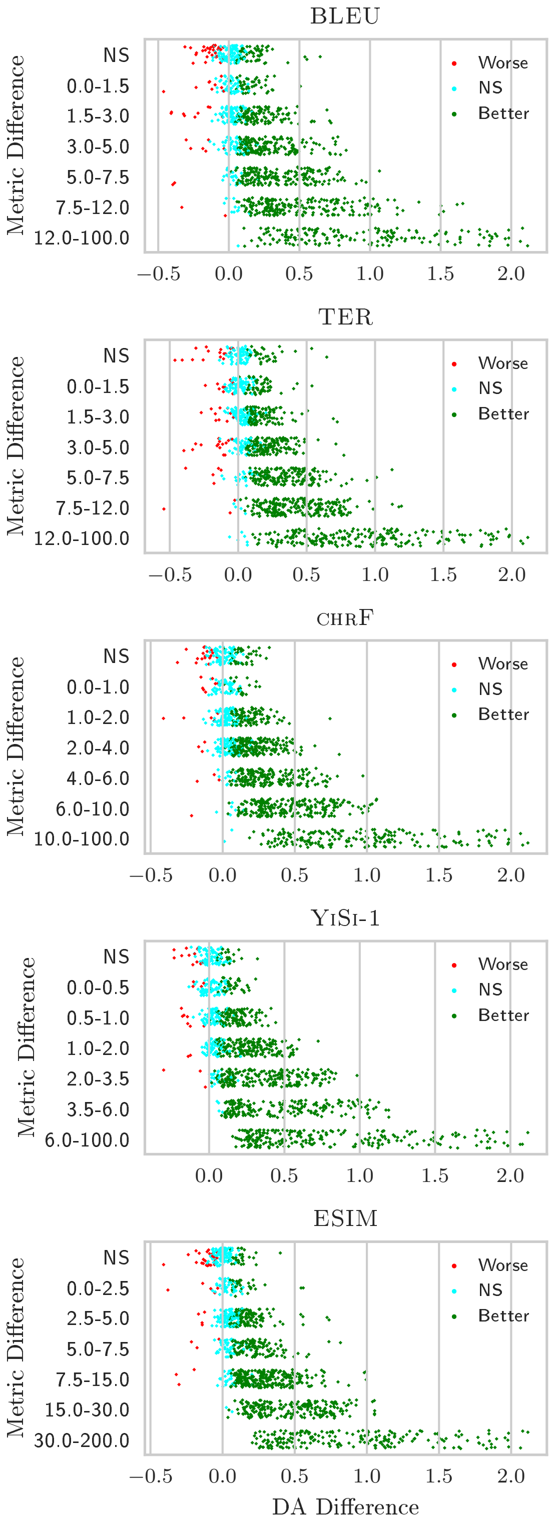


Figure 4: Pairwise differences in human DA evaluation (x-axis) compared to difference in metric evaluation (binned on y-axis; NS means insignificant metric difference). The colours indicate pairs judged by humans to be insignificantly different (cyan/light gray), significantly worse (red/dark gray on the left) and significantly better (green/dark gray on the right).

cal significance tests to detect if the difference in scores (human or metric) between two systems (S1 and S2) can just be attributed to chance.

For human scores, we apply the Wilcoxon rank-sum test which is used by WMT when ranking systems. We use the bootstrap method (Koehn, 2004) to test for statistical significance of the difference in BLEU between two systems. YISI-1 and ESIM compute the system score as the average of sentence scores, so we use the paired t-test to compute significance. Although CHRF is technically the macro-average of n-gram statistics over the entire test set, we treat this as a micro-average when computing significance such that we can use the more powerful paired t-test over sentence scores.

Figure 4 visualises the agreement between metric score differences and differences in human DA scores. Ideally, only differences judged as truly significant would give rise to significant and large magnitude differences under the metrics; and when metrics judge differences to be insignificant, ideally very few instances would be truly significant. However, this is not the case: there are substantial numbers of insignificant differences even for very high metric differences (cyan, for higher range bins); moreover, the "NS" category — denoting an insignificant difference in metric score — includes many human significant pairs (red and green, top bin).

Considering BLEU (top plot in Figure 2), for insignificant BLEU differences, humans judge one system to be better than the other for half of these system pairs. This corresponds to a Type I error. It is of concern that BLEU cannot detect these differences. Worse, the difference in human scores has a very wide range. Conversely, when the BLEU score is significant but in the range 0–3, more than half of these systems are judged to be insignificantly different in quality (corresponding to a Type II error). For higher BLEU deltas, these errors diminish, however, even for a BLEU difference between 3 and 5 points, about a quarter of these system pairs are of similar quality. This paints a dour picture for the utility of BLEU as a tool for gatekeeping (i.e., to define a 'minimum publishable unit' in deciding paper acceptance on empirical grounds, through bounding the risk of Type II errors), as the unit would need to be whoppingly large to ensure only meaningful improvements are accepted. Were we seek to minimise Type I errors in the interests of nurturing good ideas, the threshold would need to be so low as to be meaningless, effectively below the level required for acceptance of the bootstrap significance test.

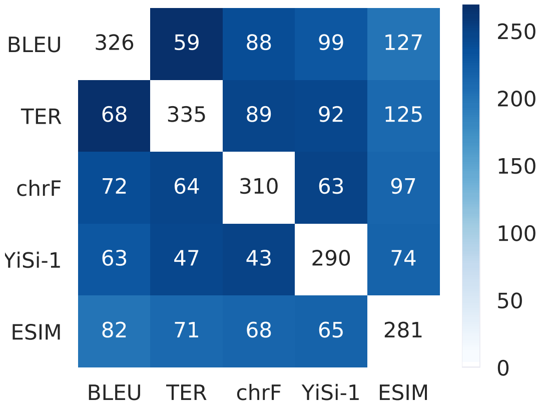


Figure 5: The agreement between metric errors over all 1362 system comparisons. The values in the diagonal indicate the total number of Type 1 and Type 2 errors for the metric. The off-diagonal cells show the total number of errors made by the row-metric where the column-metric is correct.

The systems evaluated consist of a mix of systems submitted by researchers (mostly neural models) and anonymous online systems (where the MT system type is unknown). Even when we restrict the set of systems to only neural models submitted by researchers, the patterns of Type 1 and Type 2 errors remain the same (figure omitted for space reasons).

TER makes similar errors: TER scores can wrongly show that a system is much better than another when humans have judged them similar, or even worse, drawn the opposite conclusion.

CHRF, YISI-1 and ESIM have fewer errors compared to BLEU and TER. When these metrics mistakenly fail to detect a difference between systems, the human score difference is considerably lower than for BLEU. Accordingly, they should be used in place of BLEU. However the above argument is likely to still hold true as to their utility for gatekeeping or nurturing progress, in that the thresholds would still be particularly punitive or permissive, for the two roles, respectively.

Finally, Figure 5 looks at agreement between metric decisions when comparing MT systems. As expected, when BLEU or TER disagree with CHRF, ESIM, or YISI-1, the former are more likely to be wrong. BLEU and TER have an 80% overlap in errors. The decisions of ESIM, a trained neural model, diverge a little more from the other metrics. Overall, despite the variety of approaches towards the task, all five metrics have common biases: over half of all erroneous decisions made by a particular metric are made in common with all other metrics.

6 Conclusion

In this paper, we revisited the findings of the metrics task at WMT 2019, which flagged potential problems in the current best practises for assessment of evaluation metrics.

Pearson's correlation coefficient is known to be unstable for small sample sizes, particularly when the systems in consideration are very close in quality. This goes some way to explaining the findings whereby strong correlations between metric scores and human judgements evaporate when considering small numbers of strong systems. We show that the same can be true for any small set of similar quality systems, not just the top systems. This effect can partly be attributed to noise due to the small sample size, rather than true shortcomings in the metrics themselves. We need better methods to empirically test whether our metrics are less reliable when evaluating high quality MT systems.

A more serious problem, however, is outlier systems, i.e. those systems whose quality is much higher or lower than the rest of the systems. We found that such systems can have a disproportionate effect on the computed correlation of metrics. The resulting high values of correlation can then lead to to false confidence in the reliability of metrics. Once the outliers are removed, the gap between correlation of BLEU and other metrics (e.g. CHRF, YISI-1 and ESIM) becomes wider. In the worst case scenario, outliers introduce a high correlation when there is no association between metric and human scores for the rest of the systems. Thus, future evaluations should also measure correlations after removing outlier systems.

Finally, the same value of correlation coefficient can describe different patterns of errors. Any single number is not adequate to describe the data, and visualising metric scores against human scores is the best way to gain insights into metric reliability. This could be done with scatter plots (e.g. Figure 3a) for each language pair, or Figure 5, which compresses this information into one graph.

Metrics are commonly used to compare two systems, and accordingly we have also investigated the real meaning encoded by a difference in metric score, in terms of what this indicates about human judgements of the two systems. Most published work report BLEU differences of 1-2 points, however at this level we show this magnitude of difference only corresponds to true improvements in quality as judged by humans about half the time. Although our analysis assumes the Direct Assessment human evaluation method to be a gold standard despite its shortcomings, our analysis does suggest that the current rule of thumb for publishing empirical improvements based on small BLEU differences has little meaning.

Overall, this paper adds to the case for retiring BLEU as the de facto standard metric, and instead using other metrics such as CHRF, YISI-1, or ESIM in its place. They are more powerful in assessing empirical improvements. However, human evaluation must always be the gold standard, and for continuing improvement in translation, to establish significant improvements over prior work, all automatic metrics make for inadequate substitutes.

To summarise, our key recommendations are:

- When evaluating metrics, use the technique outlined in Section 4.2 to remove outliers before computing Pearson's r.
- When evaluating MT systems, stop using BLEU or TER for evaluation of MT, and instead use CHRF, YISI-1, or ESIM;
- Stop using small changes in evaluation metrics as the sole basis to draw important empirical conclusions, and make sure these are supported by manual evaluation.

Acknowledgements

We are grateful to the anonymous reviewers for their comments and valuable suggestions. This work was supported in part by the Australian Research Council.

References

Francis J Anscombe. 1973. Graphs in statistical analysis. *The American Statistician*, 27(1):17–21.

Vic Barnett and Toby Lewis. 1974. *Outliers in Statistical Data*. Wiley.

Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019.

Findings of the 2019 conference on machine translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy. Association for Computational Linguistics.

Ondřej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, Radu Soricut, Lucia Specia, and Aleš Tamchyna. 2014. Findings of the 2014 workshop on statistical machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 12–58, Baltimore, Maryland, USA. Association for Computational Linguistics.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin Verspoor, and Marcos Zampieri. 2016. Findings of the 2016 conference on machine translation. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 131–198, Berlin, Germany. Association for Computational Linguistics.

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Re-evaluating the role of Bleu in machine translation research. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, Trento, Italy. Association for Computational Linguistics.

Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Si Wei, Hui Jiang, and Diana Inkpen. 2017. Enhanced LSTM for natural language inference. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1657–1668, Vancouver, Canada. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, USA.

Marina Fomicheva and Lucia Specia. 2019. Taking MT evaluation metrics to extremes: Beyond correlation with human judgments. *Computational Linguistics*, 45(3):515–558.

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

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2017. Can machine translation systems be evaluated by the crowd alone? *Natural Language Engineering*, 23(1):3–30.

Yvette Graham, Nitika Mathur, and Timothy Baldwin. 2014. Randomized significance tests in machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 266–274, Baltimore, Maryland, USA. Association for Computational Linguistics.

Boris Iglewicz and David Caster Hoaglin. 1993. *How to detect and handle outliers*, volume 16. Asq Press.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain.

Christophe Leys, Christophe Ley, Olivier Klein, Philippe Bernard, and Laurent Licata. 2013. Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of Experimental Social Psychology*, 49(4):764–766.

Chi-kiu Lo. 2019. YiSi — a unified semantic MT quality evaluation and estimation metric for languages with different levels of available resources. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 507–513, Florence, Italy.

Qingsong Ma, Johnny Wei, Ondřej Bojar, and Yvette Graham. 2019. Results of the WMT19 metrics shared task: Segment-level and strong MT systems pose big challenges. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 62–90, Florence, Italy.

Nitika Mathur, Timothy Baldwin, and Trevor Cohn. 2019. Putting evaluation in context: Contextual embeddings improve machine translation evaluation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2799–2808, Florence, Italy.

Jason W Osborne and Amy Overbay. 2004. The power of outliers (and why researchers should always check for them). *Practical Assessment, Research & Evaluation*, 9(6):1–12.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002a. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, USA.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002b. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of* *the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, USA.

Maja Popović. 2015. chrF: character n-gram f-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.

Ehud Reiter. 2018. A structured review of the validity of BLEU. *Computational Linguistics*, 44(3):393–401.

Peter J Rousseeuw and Mia Hubert. 2011. Robust statistics for outlier detection. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(1):73–79.

Aaron Smith, Christian Hardmeier, and Joerg Tiedemann. 2016. Climbing mont BLEU: The strange world of reachable high-BLEU translations. In *Proceedings of the 19th Annual Conference of the European Association for Machine Translation*, pages 269–281.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the Association for Machine Transaltion in the Americas*, pages 223–231.

Amanda Stent, Matthew Marge, and Mohit Singhai. 2005. Evaluating evaluation methods for generation in the presence of variation. In *Computational Linguistics and Intelligent Text Processing*, pages 341–351, Berlin, Heidelberg. Springer Berlin Heidelberg.

Rand Wilcox. 2004. Inferences based on a skipped correlation coefficient. *Journal of Applied Statistics*, 31(2):131–143.

A The effect of removing outlier systems on the results of the WMT 19 metrics task

	de–cs		de–fr	fr–de	
	All	−out	All	All	−out
n	11	10	11	10	7
BEER	**0.978**	0.976	**0.941**	0.848	**0.794**
BLEU	**0.941**	0.922	0.891	0.864	**0.821**
CDER	0.864	0.734	**0.949**	0.852	**0.794**
CharacTER	0.965	0.959	0.928	0.849	**0.848**
chrF	**0.974**	**0.970**	0.931	0.864	**0.796**
chrF+	0.972	0.967	0.936	0.848	**0.785**
EED	**0.982**	**0.984**	**0.940**	0.851	0.792
ESIM	**0.980**	**0.986**	**0.950**	**0.942**	**0.825**
hLEPORa_baseline	0.941	0.903	0.814	–	–
hLEPORb_baseline	**0.959**	**0.951**	0.814	–	–
NIST	**0.954**	0.944	**0.916**	0.862	**0.800**
PER	0.875	0.757	0.857	**0.899**	0.427
sacreBLE-BLEU	0.869	0.742	0.891	0.869	**0.846**
sacreBLE-chrF	**0.975**	**0.980**	**0.952**	0.882	**0.815**
TER	0.890	0.787	**0.956**	**0.895**	**0.673**
WER	0.872	0.749	**0.956**	**0.894**	**0.657**
YiSi-0	**0.978**	0.972	**0.952**	0.820	**0.836**
YiSi-1	0.973	**0.980**	**0.969**	**0.908**	**0.846**
YiSi-1_srl	–	–	–	**0.912**	0.814
Source-based metrics:					
ibm1-morpheme	0.355	0.009	0.509	0.625	**0.357**
ibm1-pos4gram	–	–	0.085	0.478	0.719
YiSi-2	0.606	0.122	0.721	0.530	**0.066**

Table 3: Pearson correlation of metrics for the language pairs that do not involve English. For language pairs that contain outlier systems, we also show correlation after removing outlier systems. Correlations of metrics not significantly outperformed by any other for that language pair are highlighted in bold.

	de–en		**fi–en**	**gu–en**		**kk–en**		**lt–en**		**ru–en**		**zh–en**	
	All	−out	All	All	−out	All	−out	All	−out	All	−out	All	−out
n	16	15	12	11	10	11	9	11	10	14	13	15	13
BEER	0.906	0.852	**0.993**	0.952	0.982	0.986	**0.930**	0.947	0.948	0.915	0.819	0.942	0.806
BERTR	**0.926**	**0.897**	0.984	0.938	**0.995**	0.990	0.829	0.948	0.959	**0.971**	**0.933**	0.974	0.911
BLEU	0.849	0.770	0.982	0.834	0.975	0.946	**0.912**	0.961	**0.980**	0.879	0.830	0.899	0.807
CDER	0.890	0.827	**0.988**	0.876	0.975	0.967	0.843	**0.975**	**0.981**	0.892	0.875	0.917	0.847
CHARACTER	0.898	0.852	**0.990**	0.922	0.978	0.953	0.833	0.955	0.963	0.923	0.828	0.943	0.845
CHRF	**0.917**	0.862	**0.992**	0.955	0.962	0.978	0.775	0.940	0.933	0.945	0.876	0.956	0.841
CHRF+	**0.916**	0.860	**0.992**	0.947	0.961	0.976	0.769	0.940	0.934	0.945	0.878	0.956	0.851
EED	0.903	0.853	**0.994**	0.976	0.988	0.980	0.779	0.929	0.930	0.950	0.872	0.949	0.840
ESIM	**0.941**	**0.896**	0.971	0.885	0.986	0.986	**0.945**	**0.989**	**0.990**	**0.968**	**0.946**	**0.988**	**0.961**
HLEPORA_BASELINE	–		–	–		0.975	0.855	–		–		0.947	0.879
HLEPORB_BASELINE	–		–	–		0.975	0.855	0.906	0.930	–		0.947	0.879
METEOR++_2.0(SYNTAX)	0.887	0.844	**0.995**	0.909	0.939	0.974	0.859	0.928	0.935	**0.950**	0.878	0.948	0.836
METEOR++_2.0(SYNTAX+COPY)	0.896	0.850	**0.995**	0.900	0.930	0.971	0.871	0.927	0.931	**0.952**	**0.890**	0.952	0.841
NIST	0.813	0.705	0.986	0.930	0.985	0.942	0.837	0.944	**0.963**	0.925	**0.878**	0.921	0.722
PER	0.883	0.808	**0.991**	0.910	0.948	0.737	0.533	0.947	0.933	0.922	0.880	0.952	0.884
PREP	0.575	0.452	0.614	0.773	0.967	0.776	**0.817**	0.494	0.397	0.782	0.685	0.592	0.111
SACREBLE-BLEU	0.813	0.794	0.985	0.834	0.975	0.946	**0.912**	0.955	0.967	0.873	0.813	0.903	0.807
SACREBLE-CHRF	0.910	0.852	**0.990**	0.952	0.937	0.969	0.750	0.935	0.923	0.919	0.874	0.955	0.846
TER	0.874	0.812	**0.984**	0.890	0.947	0.799	0.566	0.960	0.975	0.917	0.896	0.840	0.717
WER	0.863	0.803	0.983	0.861	0.926	0.793	0.579	0.961	**0.981**	0.911	0.885	0.820	0.716
WMDO	0.872	**0.857**	**0.987**	0.983	0.981	**0.998**	**0.953**	0.900	0.923	0.942	0.844	0.943	0.851
YISI-0	0.902	0.847	**0.993**	**0.993**	0.990	0.991	0.876	0.927	0.933	**0.958**	**0.889**	0.937	0.782
YISI-1	**0.949**	**0.914**	**0.989**	0.924	**0.997**	0.994	**0.920**	0.981	0.978	**0.979**	**0.947**	**0.979**	0.899
YISI-1_SRL	**0.950**	**0.916**	**0.989**	0.918	**0.998**	0.994	**0.917**	**0.983**	**0.981**	**0.978**	0.943	0.977	0.897
Source-based metrics:													
IBM1-MORPHEME	0.345	0.223	0.740	–		–		0.487	0.638	–		–	
IBM1-POS4GRAM	0.339	0.137	–	–		–		–		–		–	
LASIM	0.247	0.334	–	–		–		–		0.310	0.260	–	
LP	0.474	0.279	–	–		–		–		0.488	0.168	–	
UNI	0.846	**0.809**	0.930	–		–		–		0.805	0.666	–	
UNI+	0.850	**0.805**	0.924	–		–		–		0.808	0.669	–	
YISI-2	0.796	0.612	0.642	0.566	0.820	0.324	0.662	0.442	0.346	0.339	0.708	0.940	0.622
YISI-2_SRL	0.804	0.630	–	–		–		–		–		0.947	0.675

Table 4: Pearson correlation of metrics for the to-English language pairs. For language pairs that contain outlier systems, we also show correlation after removing outlier systems. Correlations of metrics not significantly outperformed by any other for that language pair are highlighted in bold.

	en–cs	**en–de**		**en–fi**		**en–gu**	**en–kk**		**en–lt**	**en–ru**		**en–zh**
	All	All	−out	All	−out	All	All	−out	All	All	−out	All
n	11	22	20	12	11	11	11	9	12	12	11	12
BEER	**0.990**	0.983	0.869	**0.989**	**0.978**	0.829	0.971	0.826	**0.982**	0.977	0.947	0.803
BLEU	0.897	0.921	0.419	**0.969**	0.943	0.737	0.852	0.576	**0.989**	0.986	0.967	0.901
CDER	0.985	0.973	0.849	**0.978**	**0.957**	0.840	0.927	0.668	**0.985**	**0.993**	**0.981**	0.905
CHARACTER	**0.994**	**0.986**	**0.886**	0.968	0.939	**0.910**	0.936	**0.895**	0.954	**0.985**	**0.982**	0.862
CHRF	0.990	0.979	0.881	**0.986**	**0.972**	**0.841**	**0.972**	**0.900**	**0.981**	0.943	0.968	0.880
CHRF+	**0.991**	0.981	0.883	**0.986**	**0.970**	**0.848**	**0.974**	**0.907**	**0.982**	0.950	0.973	0.879
EED	**0.993**	**0.985**	**0.894**	**0.987**	**0.978**	**0.897**	**0.979**	**0.883**	0.975	0.967	**0.984**	0.856
ESIM	–	**0.991**	**0.928**	0.957	0.926	–	**0.980**	**0.900**	**0.989**	**0.989**	**0.986**	0.931
HLEPORA_BASELINE	–	–		–		0.841	0.968	**0.852**	–	–		–
HLEPORB_BASELINE	–	–		–		0.841	0.968	0.852	0.980	–		–
NIST	0.896	0.321	0.246	0.971	0.936	0.786	0.930	0.611	**0.993**	**0.988**	**0.973**	0.884
PER	0.976	0.970	0.815	**0.982**	**0.961**	0.839	0.921	0.545	0.985	0.981	0.955	0.895
SACREBLE-BLEU	**0.994**	0.969	0.806	**0.966**	0.939	0.736	0.852	0.576	**0.986**	0.977	0.946	0.801
SACREBLE-CHRF	0.983	0.976	0.874	0.980	0.958	0.841	**0.967**	0.840	0.966	**0.985**	**0.988**	0.796
TER	0.980	0.969	0.841	**0.981**	**0.960**	**0.865**	0.940	0.547	**0.994**	**0.995**	**0.985**	0.856
WER	0.982	0.966	0.831	**0.980**	**0.958**	**0.861**	0.939	0.525	**0.991**	**0.994**	**0.983**	0.875
YISI-0	**0.992**	0.985	0.869	**0.987**	**0.977**	0.863	0.974	0.840	0.974	0.953	**0.967**	0.861
YISI-1	0.962	**0.991**	**0.917**	0.971	0.937	**0.909**	**0.985**	**0.892**	0.963	**0.992**	**0.978**	**0.951**
YISI-1_SRL	–	**0.991**	**0.917**	–		–	–		–	–		**0.948**
Source-based metrics:												
IBM1-MORPHEME	0.871	0.870	0.198	0.084	0.254	–	–		0.810	–		–
IBM1-POS4GRAM	–	0.393	0.449	–		–	–		–	–		–
LASIM	–	0.871	0.007	–		–	–		–	0.823	0.336	–
LP	–	0.569	0.558	–		–	–		–	0.661	0.178	–
UNI	0.028	0.841	0.251	0.907	0.808	–	–		–	0.919	0.760	–
UNI+	–	–		–		–	–		–	0.918	0.746	–
USFD	–	0.224	0.301	–		–	–		–	0.857	0.514	–
USFD-TL	–	0.091	0.212	–		–	–		–	0.771	0.177	–
YISI-2	0.324	0.924	0.014	0.696	0.478	0.314	0.339	**0.685**	0.055	0.766	0.134	0.097
YISI-2_SRL	–	0.936	0.155	–		–	–		–	–		0.118

Table 5: Correlation of metrics for the from-English language pairs. For language pairs that contain outlier systems, we also show correlation after removing outlier systems. Values in bold indicate that the metric is not significantly outperformed by any other metric under the Williams Test.