Beyond Correlation: Making Sense of the Score Differences of New MT Evaluation Metrics

Chi-kiu Lo 羅致翹 chikiu.lo@nrc-cnrc.gc.ca
Rebecca Knowles rebecca.knowles@nrc-cnrc.gc.ca
Cyril Goutte cyril.goutte@nrc-cnrc.gc.ca
National Research Council Canada (NRC-CNRC)

Abstract

While many new automatic metrics for machine translation evaluation have been proposed in recent years, BLEU scores are still used as the primary metric in the vast majority of MT research papers. There are many reasons that researchers may be reluctant to switch to new metrics, from external pressures (reviewers, prior work) to the ease of use of metric toolkits. Another reason is a lack of intuition about the meaning of novel metric scores. In this work, we examine "rules of thumb" about metric score differences and how they do (and do not) correspond to human judgments of statistically significant differences between systems. In particular, we show that common rules of thumb about BLEU score differences do not in fact guarantee that human annotators will find significant differences between systems. We also show ways in which these rules of thumb fail to generalize across translation directions or domains.

1 Introduction

Despite mounting evidence over the course of many years (Akiba et al., 2003; Callison-Burch et al., 2006; Chiang et al., 2008; Tan et al., 2015; Mathur et al., 2020a, i.a.) demonstrating that BLEU (Papineni et al., 2002) has fundamental flaws in accurately reflecting translation quality, it has remained the de facto standard automatic MT evaluation metric for both scientific research and practical deployment (Marie et al., 2021). Numerous research efforts (Callison-Burch et al., 2007; Przybocki et al., 2009; Bojar et al., 2017; Freitag et al., 2021b, 2022, i.a.) have focused on the correlations between human judgments of translation quality and automatic metric scores; year after year, these have shown new metrics correlating better with human judgments than BLEU does. There are certainly some other obstacles beyond correlation with human judgment on translation quality that hinder the adoption of newer and better human-correlating automatic MT evaluation metrics in practice.

Przybocki et al. (2009) outlined four objectives in the search for new and improved automatic MT evaluation metrics: 1) "high correlation with human assessments of translation quality"; 2) "applicable to multiple target languages"; 3) "ability to differentiate between systems of varying quality" and finally, 4) "intuitive interpretation", i.e. whether the scores are meaningful and easy to understand on their own, with values and differences that are interpretable and clear in practice. As the first three objectives can be addressed by the correlation analysis of MT metrics with human judgment on translation quality but not the last one, we believe that gaining an intuitive understanding of the properties and behavior of the metrics is one

Proceedings of Machine Translation Summit XIX, Vol. 1: Research Track, pages 186–199
September 4–8, 2023, Macau SAR, China.

of the remaining challenges that MT researchers are facing when they are considering adopting a new metric. One way to do this is by designing metrics to be easily interpretable; another way is to examine whether we can build up reliable and useful intuitions about existing metrics. In order for metrics to be widely adopted, a combination of these—making new metrics that are more interpretable and simple to understand or debug,[1] as well as forming intuitions about them—may be necessary. Our focus in this work is on the latter, examining existing metrics to understand the meaning of the score differences they present.

In this work, we focus in particular on whether it is possible to get a sense of what kinds of metric score differences may correspond to significant improvements as judged by human annotators.[2] We examine whether this is consistent across target languages and across translation domains, within a specific metric. We do not suggest that this means that MT researchers can forego running significance tests or doing the appropriate human evaluation; as Marie (2022) notes, "A rule of thumb may yield correct results but can't be scientifically credible." However, having these rules of thumb and intuitive senses of metric score meanings may indeed be necessary to encourage broader adoption, so we present this work solely focusing on whether it is possible to build such rules of thumb about some of the modern metrics.

2 Related work

Mathur et al. (2020a) demonstrated that even statistically significant BLEU score differences of 0-3 BLEU points do not reliably correspond to human judgments of significant differences between systems. With a focus on pairwise ranking of systems, Kocmi et al. (2021) argued for evaluating metrics primarily based on whether the metric's pairwise rankings of two systems agrees with human pairwise rankings. They found that among the system pairs that were deemed statistically significant by humans, but where BLEU produced a flipped ranking compared to humans, the median BLEU difference is 1.3 BLEU. They found this result concerning as "BLEU differences higher than 1 or 2 BLEU are commonly and historically considered to be reliable by the field" (Kocmi et al., 2021) and their result showed otherwise. They further encouraged the use of paired statistical significance tests for more reliable conclusions on MT quality improvement. Subsequently, Marie (2022) examined the Conference on Machine Translation (WMT) 2021 and 2022 data to see what thresholds of metric score difference magnitudes corresponded reliably to statistically significant differences in metric scores (at p-values < 0.05, < 0.01, and < 0.001). They found that to claim a significant improvement in metric scores with p-value < 0.001, statistical significance testing should be done for differences lower that 2 BLEU. However, they only focused on significance in metric scores improvement but did not consider whether such thresholds correspond to significance in human judgments.

Nevertheless, there remain some common "gut feelings" among researchers and reviewers about what constitutes "significant" improvement on the basis of metric score differences alone, without running human evaluation or significance tests. As Marie et al. (2021) note, the majority of MT papers since 2018 do not use significance tests and instead rely on score differences. One number commonly tossed about informally is that a score difference of around 2 BLEU points can typically be expected to be significant. But where does this assumption come from, and does it hold? One possible source for this is Koehn (2004), which found, specific to the particular test scenario that "Even for small test sets of size 300 sentences (about 9000 words), we can reliably draw the right conclusion, if the true BLEU score difference is at least 2-3%."[3] In that setting,

[1] Another reason, beyond the scope of this work, that researchers may be hesitant to adopt new, complex metrics, is the possibility that they may have unexpected failure modes (see, e.g., Yan et al., 2023).

[2] Concurrent work, Deutsch et al. (2023), provides another way to examine score differences and their relation to human annotation.

[3] Note that this refers to a score difference of 2 or 3 BLEU points, not relative improvement.

the "right conclusion" is the one that matches the conclusion drawn from a very large test set (30,000 sentences) about which of two systems is better, based on automatic metric scores. The goal of that work is to identify how small of a dataset can still be reliably used (along with bootstrap resampling for statistical significance) to draw conclusions about the automatically measured differences between two systems. However, sometimes this kind of BLEU difference is used informally as a proxy for whether a *human* annotator will find the difference notable, something that does not follow from that particular paper. Marie (2022) found that for systems from WMT21 and 22, almost all system pairs with a BLEU difference greater than 2.0 were significantly different with p-value < 0.001, though this significance judgment relates only to the metric scores and not to any human annotation. In this work, we focus on a question more closely related to this and to Mathur et al. (2020a), rather than Marie (2022): whether there exist rules of thumb about metric score differences and their correspondence to significant differences in *human judgments*. Regardless of the exact source of these rules of thumb (which may never be known) or the exact BLEU score difference (or exact relative improvement) of a particular rule of thumb, some researchers feel that they have a sense of metric score differences, and we examine how that may correspond (or not) to judgments of MT quality across a range of metrics in this work.

Similar to Mathur et al. (2020a), we are interested in the relationship between metric score differences and significant differences in human scores. That work is interested both in Type I errors (where an insignificant metric difference might correspond to an actually significant difference under human evaluation) and Type II errors (where the metric score difference is significant, but the human evaluation does not find a significant difference). We take a related but slightly different approach to examining this relationship. We examine the "rules of thumb" about which metric score differences are meaningful. Using the large number of system pairs from WMT evaluations, we look at how the metric difference between two systems is related to the probability that the human annotations find the systems to be statistically significantly different. We select a threshold for this probability and examine the metric difference that corresponds. We then examine whether this is consistent across different test sets, domains, and target languages. That is, *are* there consistent rules of thumb about metric score differences? Or is there too much variation?

3 Do BLEU score rules of thumb correspond to human judgments?

In casual discussion and sometimes even formal work or reviewing, there is often a conflation of several (somewhat) orthogonal topics, which may be the source of these intuitions and rules of thumb. Sometimes "significant" is used simply to mean some value of "large", unrelated to precise definitions of statistical significance testing. Marie et al. (2021) note the use of this convention and suggest that it indicates some level of consensus among researchers on BLEU differences, albeit a consensus that is not necessarily well-founded; they address a number of other pitfalls in MT evaluation as well. In particular, in their meta-evaluation of 769 MT papers, they note that the majority of recent papers do not perform statistical significance tests, relying instead just on the "amplitude of the differences between metric scores to state whether they are significant or not"; in fact they note that even a BLEU score difference of around 1 may be used by most MT papers as "*significant* evidence of the superiority of an MT system and as an improvement in translation quality" (Marie et al., 2021). These are assumptions that sufficiently large metric score differences guarantee significant differences in *metric* scores; when combined with the assumption that metric scores and human scores are well-correlated, this often leads to the assumption that a certain metric score difference guarantees a statistically significant difference in *human scores*. We examine this relationship between statistically significant differences in human scores and the magnitude of metric differences in this work.

First, we investigate whether the more generous rule of thumb surrounding the significance of 2 BLEU improvement has a basis in fact. While Marie (2022) has shown that (at least for WMT21 and 22) such a BLEU difference tends to be a significant difference ($p < 0.001$) in metrics score, does that mean that human annotators will judge the pair of systems to be meaningfully different? That is to say, we assess the probability that an MT system pair would be judged by humans as having a statistically significant difference in quality, if BLEU showed a difference of 2 or more points for that pair.

3.1 Data

We use the human direct assessment (DA) and direct assessment with scalar quality metric (DA+SQM, which we refer to in figures as SQM for conciseness) scores collected at the WMT News/General shared tasks from 2019 to 2022 (Barrault et al., 2019, 2020; Akhbardeh et al., 2021; Kocmi et al., 2022) and organized in the MT Metrics Eval package.[4] The MT Metrics Eval package includes all scores from baseline and participating MT evaluation metrics in the Metrics shared task (Ma et al., 2019; Mathur et al., 2020b; Freitag et al., 2021b, 2022), covering all segments of all MT systems in WMT News/General shared tasks. It also contains complete information about which segments of each MT system were annotated by human evaluators on translation quality, allowing us to run paired t-test for each system pair on their sentence-level human DA/SQM (normalized) scores.

3.2 DA/SQM

In DA (Graham et al., 2017) at WMT, human annotators are asked to rate translations compared to the corresponding source/reference sentence on a slider of continuous scale between 0 and 100. The difference between DA and the DA+SQM performed at WMT22 (Kocmi et al., 2022) is that, for the latter, the slider is marked with seven tick marks where four of them are labeled with quality guidelines. The sentence-level human scores are standardized using z-scores.

3.3 Automatic MT evaluation metrics

The automatic MT evaluation metrics chosen for this study are the baselines and the high-performing participants in the WMT19-22 Metrics shared tasks. **BLEU** (Papineni et al., 2002) is the (clipped) precision of word n-grams between the MT output and its reference weighted by a brevity penalty. **spBLEU** (Team et al., 2022) is BLEU computed with subword tokenization done by standardized Sentencepiece Models (Kudo and Richardson, 2018). **chrF** (Popović, 2015) uses character n-gram to compare the MT output with the reference and it is a balance of precision and recall. **BERTScore** (Zhang et al., 2020) uses cosine similarity of contextual embeddings from pretrained transformers to compute F-score of sentence level similarity. **BLEURT-20** (Sellam et al., 2020) is fine-tuning RemBERT to predict DA score for a MT-reference pair. **COMET-20** (Rei et al., 2020) is fine-tuning XLM-R to predict DA score for a MT-source-reference tuple. **YiSi-1** (Lo, 2019) measures the semantic similarity between the MT output and reference by the IDF-weighted cosine similarity of contextual embeddings extracted from pretrained language models, e.g. RoBERTa, CamemBERT, XLM-R, etc., depending on the target language in evaluation. **COMET-22** (Rei et al., 2022) is an ensemble of two models: COMET-20 and a multitask model jointly predicting sentence-level MQM and word-level translation quality annotation. **metricX XXL** is the MQM prediction from a massive multi-task metric fine-tuned 30B mT5 using a variety of human feedback data such as, DA, MQM, QE, NLI and Summarization Eval. **UniTE** (Wan et al., 2022) is a learnt metric that unified

[4] https://github.com/google-research/mt-metrics-eval

commit: bdda529ce4fae9cec8156ea8a0abd94fe1b85988

ΔBLEU	2.0	5.0	10.0
$Pr(p < 0.05 \vert \Delta\text{BLEU})$	0.56	0.70	0.91

Table 1: Probability of significant human score difference at $p < 0.05$ given Δ BLEU of 2.0, 5.0 and 10.0 respectively.

the reference-based, reference-free and MT-source-reference way of evaluation trained on data with synthetic translation quality label.

3.4 Statistical significance test on human scores and isotonic regression

To ensure enough statistical power in the paired t-test on the sentence-level human DA/SQM (normalized) scores, we first filter out system pairs that have fewer than 250 sentences in common annotated by the human evaluators. Since we are running the significance test on the normalized human scores with the sign of the human score differences known, we run the one-sided t-test with the equal variance assumption.

After collecting the metric score difference and p-value of the t-test on the human scores for each system pair, we fit the data to an isotonic regression (Robertson et al., 1988) that predicts whether the human score difference will be significant given the metric's score difference. Isotonic regression produces a non-decreasing function where the classifier output is interpretable as a confidence level.[5] We set p<0.05 as the significance level of human scores. Thus, the output of our isotonic regression function can be viewed as $Pr(p < 0.05|\Delta M)$ where p is the p-value of the t-test on the human scores for each system pair and ΔM is the metric score difference.

3.5 Results

Figure 1 and 2 show the (log) p-value of one-sided paired t-test on human DA/SQM z-scores[6] for each metric score difference of each system pair in WMT19-22, across all translation directions and domains. Note that each system is only compared against other systems within its same language pair and direction (and for which there is an overlap of at least 250 common human-annotated segments for the pair of systems).

For all the metrics, we can choose metric score difference cutoffs (i.e., a point along the x-axis) to give a particular level of confidence that this metric difference genuinely reflects significant human score differences. Drawing a line up from the metric difference to the red line enables us to say that the metric difference at that x-value corresponds to a confidence level at corresponding y-value on the red line (for example, as seen in Table 1, a BLEU score difference of 2.0 corresponds to a 56% chance of the corresponding human evaluation finding a significant difference between the two systems). However, in the sub-figure of BLEU, we can see that data points are more spread out to the top-center of the graph. This indicates even where the BLEU differences are high human evaluators are not always finding the two systems to be significantly different; these are areas where the conclusion drawn from BLEU would be incorrect from the perspective of human evaluation. More data points spread out to the top-center also means having to make a tradeoff in the rule of thumb: either a very high score difference for high confidence of human judgment significance, or a smaller score difference but a lower confidence that the difference will be judged to be significant by human annotators.

More importantly, table 1 shows the probabilities of significant human score difference at p<0.05 given BLEU differences of 2.0, 5.0 and 10.0 respectively. For 2 BLEU difference, the probability that human evaluators find the MT output significant different is as low as 56%, i.e.

[5] `https://scikit-learn.org/stable/modules/isotonic.html`

[6] Points with lower y-axis values have smaller p-values and are "more" statistically significant.

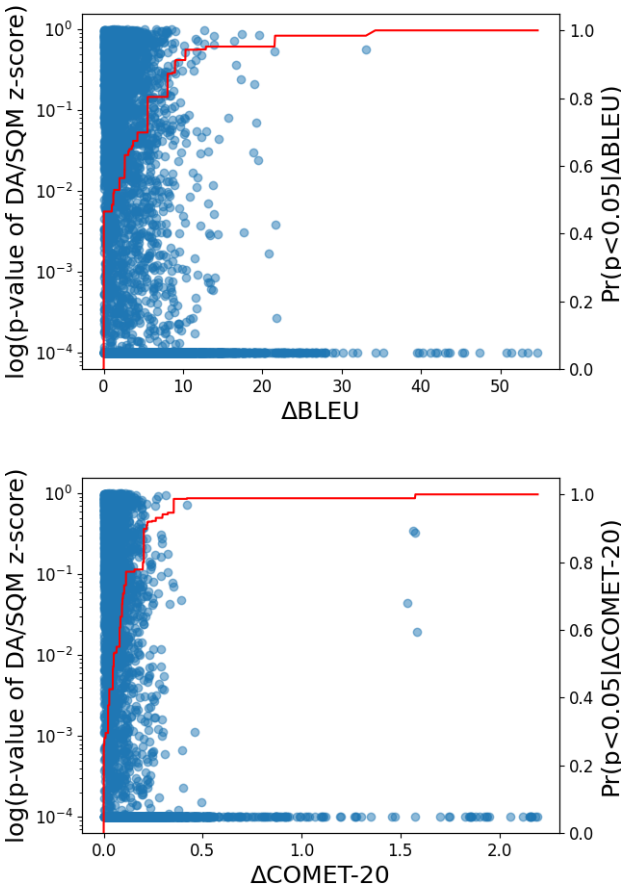

Figure 1: Log p-value of one-sided paired t-test on human DA/SQM z-scores for each metric (top: BLEU, bottom: COMET-20) score difference of each system pair in WMT19-22, all translation directions/domains. Red line is the isotonic regression fit to all data points, representing $Pr(p < 0.05|\Delta M)$. Note: for readability, p-values of $p \leq 0.0001$ are rounded up to 0.0001.

nearly one in every two times when we observe 2 BLEU improvement, it does not correspond to a significant human difference. A wider BLEU improvement margin (5 or 10 points) is needed for higher confidence that translation quality improvement will be judged to be significant by human annotators. This indicates that these rule of thumb intuitions about what kind of BLEU score differences are meaningful (or statistically significant) appear to be overstated and inaccurate, at least when it comes to significant differences in *human* judgment, which is generally considered to be the gold standard and what metric scores are seeking to replicate.

Finally, table 2 shows the cutoff of metrics' score differences for human notable difference at 50%, 80% and 95% confidence level. This table serves as a lookup between BLEU differences and differences in some of the modern metrics. For example, we see that a BLEU score difference of 1.2 corresponds to 50% confidence that human annotators will agree with the metric's ranking of the two systems and do so with a significant difference. Meanwhile, a COMET-20 score difference of 0.05 would have the same 50% chance of human-judged significant difference.

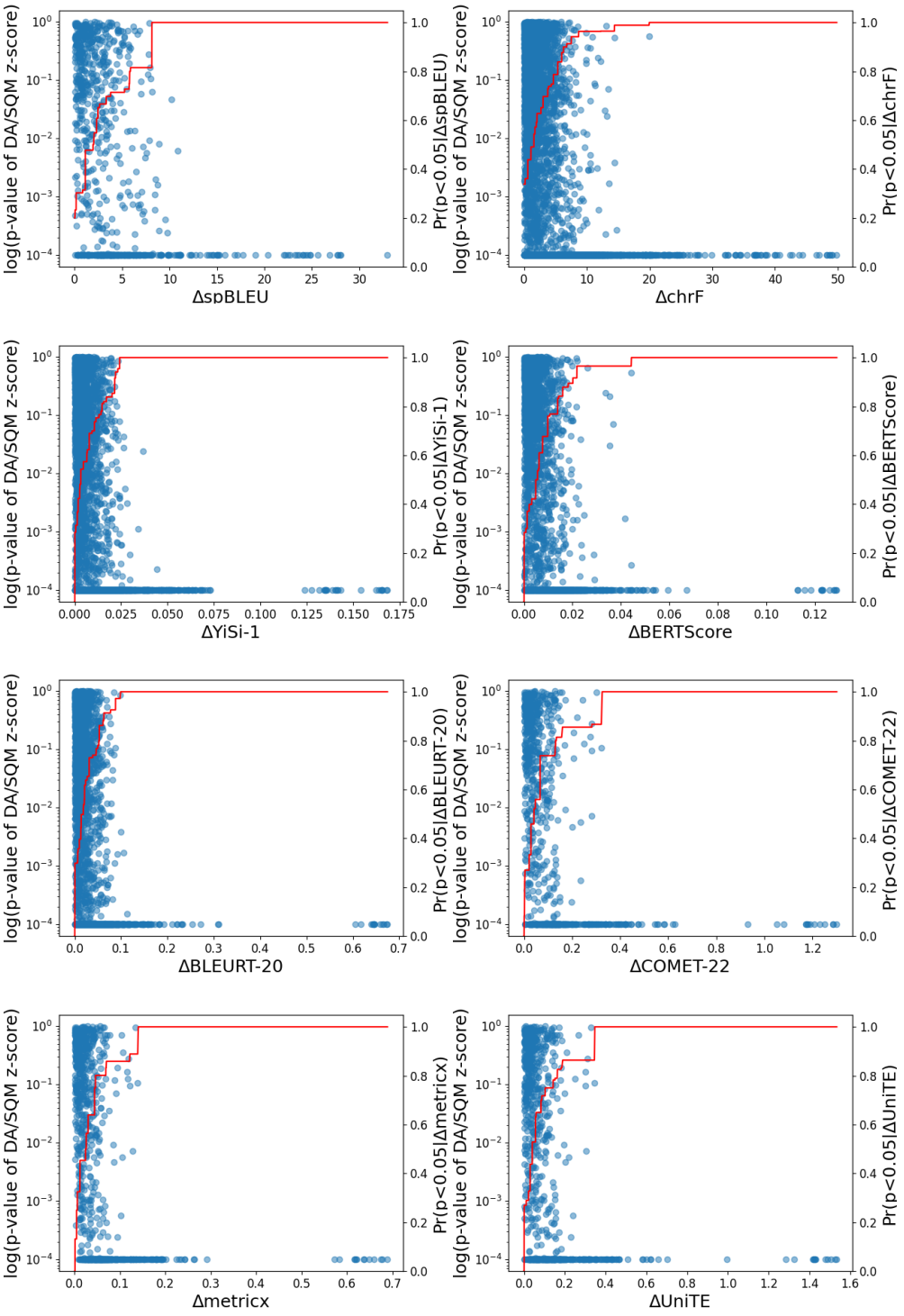

Figure 2: Log p-value of one-sided paired t-test on human DA/SQM z-scores for each metric score difference (top-to-bottom, left-to-right: spBLEU, chrF, YiSi-1, BERTScore, BLEURT-20, COMET-22, metricx, UniTE) of each system pair in WMT19-22 all translation directions and domains. The red line is the isotonic regression fit to all the data points, representing $Pr(p < 0.05|\Delta M)$. Note: for readability, p-values of $p \leq 0.0001$ are rounded up to 0.0001.

$Pr(p < 0.05 \mid \Delta M)$	0.5	0.8	0.95
surface matching			
ΔBLEU	1.2	5.5	12.9
ΔspBLEU	1.9	5.8	8.1
ΔchrF	1.6	5.4	8.7
neural (before 2022)			
ΔBERTScore	0.005	0.014	0.022
ΔBLEURT-20	0.018	0.052	0.088
ΔCOMET-20	0.05	0.20	0.35
ΔYiSi-1	0.003	0.015	0.023
neural (in 2022)			
ΔCOMET-22	0.04	0.13	0.33
Δmetricx	0.02	0.05	0.14
ΔUniTE	0.04	0.16	0.35

Table 2: Cutoff of metrics' score differences for significant human difference at 50%, 80% and 95% confidence level.

4 Discussion

Another unaddressed problem of the rules of thumb is that the MT community may sometimes treat them as though they are language and domain independent, applying rules of thumb across different target languages and domains without considering their differences. This is the case despite the fact that it is widely known that language typology affects BLEU scores (for example, highly inflected languages may see their BLEU scores penalized due to single-character differences in affixes). In addition, recent WMT Metrics shared tasks (Freitag et al., 2021b, 2022) has moved on to using multidimensional quality metric (MQM) (Lommel et al., 2014) as the human annotation method for translation quality for more consistent and reliable annotations (Freitag et al., 2021a). We now investigate into the consistency of the cutoff of metrics' score differences at 80% confidence level for different target languages, evaluation domains and human annotation methods.

4.1 Consistency across target languages

We divide the target languages into several groups: we examine all target languages together, English (the most common target language), and three groups of other target languages. These remaining groups are split into languages that use alphabetical/abugida writing systems (which we call group I: Bengali, Czech, German, French, Hausa, Croatian, Icelandic, Kazakh, Lithuanian, Polish, Russian and Ukrainian), those that use logographic writing systems (which we call group II: Chinese and Japanese), and then separately Inuktitut (which uses an abugida but is also the most morphologically complex of the target languages at WMT, in addition to being low-resource as compared to many of the other language pairs, and being covered by a smaller set of the metrics). For simplicity and space-related reasons, we select a single threshold: 80% confidence that the score difference will correspond to a significant ($p < 0.05$) human score difference. The resulting thresholds are shown in Table 3.

Beginning with BLEU, we observe a fairly stark difference between the groups of languages, with English requiring an 8.0 BLEU difference and group I languages requiring a 3.6 BLEU difference for this confidence level, with Inuktitut falling between the two. This pattern is repeated across the other metrics, though it varies by metric whether the group II languages are more similar to English or to the group I languages (in some of the pre-2022 neural metrics, the group II languages require an even smaller metric score difference than English).

target lang.	all	English	I	II	Inuktitut
surface matching					
ΔBLEU	5.5	8.0	3.6	8.0	4.5
ΔspBLEU	5.8	8.1	2.4	6.2	—
ΔchrF	5.4	6.2	3.0	3.8	6.2
neural (before 2022)					
ΔBERTScore	0.014	0.016	0.011	0.009	—
ΔBLEURT-20	0.052	0.063	0.033	0.018	—
ΔCOMET-20	0.20	0.20	0.10	0.08	0.05
ΔYiSi-1	0.015	0.022	0.005	0.010	0.023
neural (in 2022)					
ΔCOMET-22	0.13	0.33	0.07	0.08	—
Δmetricx	0.05	0.15	0.03	0.05	—
ΔUniTE	0.16	0.35	0.06	0.09	—

Table 3: Comparison of thresholds of ΔM when $Pr(p < 0.05|\Delta M) = 0.8$ for different target languages. Language group I contains system pairs translating into Bengali, Czech, German, French, Hausa, Croatian, Icelandic, Kazakh, Lithuanian, Polish, Russian and Ukrainian. Language group II contains system pairs translating into Chinese and Japanese.

In addition to highlighting the difference between languages, this also highlights another challenge: that variations on metrics have different thresholds. This should come as little surprise; even simple differences in preprocessing are known to produce differences in the same metric scores (Post, 2018). For example, we observe some inconsistency in the thresholds for BLEU and spBLEU.

BERTScore has the most consistent threshold where human annotators agree that the translation quality improvements are significant. This perhaps is because it is an untrained metric based on one multilingual pretrained transformer model so that it avoids having inconsistent implications like YiSi-1, a metric with language specific models or BLEURT-20, COMET-22 and UniTE, metrics that may be overfit to predict human scores for higher correlation.

4.2 Consistency across domains

We perform a similar comparison of thresholds for 80% confidence in human evaluation statistical significance (at $p < 0.05$) in Table 4 across domains. This analysis is restricted to 2022, where the evaluation was multi-domain. Here we combine all target languages. We again observe inconsistency across metrics, though some metrics show smaller relative threshold differences. For example, it requires double the BLEU score difference margin to be confident that translation quality of systems in the ecommerce and conversational domains significantly improved according to human evaluators, as compared to the news and social domains. For this analysis, COMET-22 has the most consistent cutoff across different domains.

4.3 Do human annotation methods matter?

Similar to the previous analyses, we perform a comparison of thresholds for 80% confidence in human evaluation statistical significance (at $p < 0.05$) in Table 5 for different human annotation protocols. Some metrics, like BLEU and chrF, show much higher score differences required for 80% confidence under MQM evaluation, while others like BLEURT-20, COMET-20, and UniTE show the opposite. More study would be required to understand these differences across human evaluation protocols and determine how to compare across different annotation methods.

domain	news	social	ecommerce	conversational
surface matching				
ΔBLEU	10.0	12.0	23.0	22.0
ΔspBLEU	10.0	13.0	14.0	10.0
ΔchrF	8.5	12.0	6.5	19.0
neural (before 2022)				
ΔBERTScore	0.013	0.016	0.009	0.025
ΔBLEURT-20	0.058	0.100	0.073	0.070
ΔCOMET-20	0.20	0.40	0.26	0.25
ΔYiSi-1	0.035	0.003	0.002	0.025
neural (in 2022)				
ΔCOMET-22	0.14	0.15	0.18	0.11
Δmetricx	0.10	0.18	0.13	0.10
ΔUniTE	0.17	0.45	0.31	0.28

Table 4: Comparison of thresholds of ΔM when $Pr(p < 0.05|\Delta M) = 0.8$ across domains.

annotation	DA/SQM	MQM
surface matching		
ΔBLEU	5.5	12.9
ΔspBLEU	5.8	5.7
ΔchrF	5.4	10.0
neural (before 2022)		
ΔBERTScore	0.014	0.012
ΔBLEURT-20	0.052	0.028
ΔCOMET-20	0.20	0.13
ΔYiSi-1	0.015	0.011
neural (in 2022)		
ΔCOMET-22	0.13	0.11
Δmetricx	0.05	0.03
ΔUniTE	0.16	0.06

Table 5: Comparison of thresholds of ΔM when $Pr(p < 0.05|\Delta M) = 0.8$ for different human annotation methods.

5 Conclusions

We presented an empirical study of the relationship between statistically significant differences in human scores and the magnitude of metric differences. We showed that the rules of thumb surrounding the significance of BLEU improvement does not hold according to human judgment on translation quality (regardless of whether the rule of thumb is exactly 1 or 2 or even slightly larger BLEU differences). We provided an intuitive interpretation between BLEU differences and the differences in some of the modern metrics. However, we found that for some metrics, the score differences corresponding to significant improvements as judged by human annotators may not be transferable across target languages or translation domains. We have to emphasize again that we do not suggest that this means that MT researchers can forego running significance tests or doing the appropriate human evaluation. This work only supports an intuitive senses of metric score meanings to encourages broader adoption of new automatic MT evaluation metrics.

References

Akhbardeh, F., Arkhangorodsky, A., Biesialska, M., Bojar, O., Chatterjee, R., Chaudhary, V., Costa-jussa, M. R., España-Bonet, C., Fan, A., Federmann, C., Freitag, M., Graham, Y., Grundkiewicz, R., Haddow, B., Harter, L., Heafield, K., Homan, C., Huck, M., Amponsah-Kaakyire, K., Kasai, J., Khashabi, D., Knight, K., Kocmi, T., Koehn, P., Lourie, N., Monz, C., Morishita, M., Nagata, M., Nagesh, A., Nakazawa, T., Negri, M., Pal, S., Tapo, A. A., Turchi, M., Vydrin, V., and Zampieri, M. (2021). Findings of the 2021 conference on machine translation (WMT21). In *Proceedings of the Sixth Conference on Machine Translation*, pages 1–88, Online. Association for Computational Linguistics.

Akiba, Y., Sumita, E., Nakaiwa, H., Yamamoto, S., and Okuno, H. G. (2003). Experimental comparison of MT evaluation methods: RED vs.BLEU. In *Proceedings of Machine Translation Summit IX: Papers*, New Orleans, USA.

Barrault, L., Biesialska, M., Bojar, O., Costa-jussà, M. R., Federmann, C., Graham, Y., Grundkiewicz, R., Haddow, B., Huck, M., Joanis, E., Kocmi, T., Koehn, P., Lo, C.-k., Ljubešić, N., Monz, C., Morishita, M., Nagata, M., Nakazawa, T., Pal, S., Post, M., and Zampieri, M. (2020). Findings of the 2020 conference on machine translation (WMT20). In *Proceedings of the Fifth Conference on Machine Translation*, pages 1–55, Online. Association for Computational Linguistics.

Barrault, L., Bojar, O., Costa-jussà, M. R., Federmann, C., Fishel, M., Graham, Y., Haddow, B., Huck, M., Koehn, P., Malmasi, S., Monz, C., Müller, M., Pal, S., Post, M., and Zampieri, M. (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.

Bojar, O., Graham, Y., and Kamran, A. (2017). Results of the WMT17 metrics shared task. In *Proceedings of the Second Conference on Machine Translation*, pages 489–513, Copenhagen, Denmark. Association for Computational Linguistics.

Callison-Burch, C., Fordyce, C., Koehn, P., Monz, C., and Schroeder, J. (2007). (meta-) evaluation of machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 136–158, Prague, Czech Republic. Association for Computational Linguistics.

Callison-Burch, C., Osborne, M., and Koehn, P. (2006). Re-evaluating the role of Bleu in machine translation research. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 249–256, Trento, Italy. Association for Computational Linguistics.

Chiang, D., DeNeefe, S., Chan, Y. S., and Ng, H. T. (2008). Decomposability of translation metrics for improved evaluation and efficient algorithms. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 610–619, Honolulu, Hawaii. Association for Computational Linguistics.

Deutsch, D., Foster, G., and Freitag, M. (2023). Ties matter: Modifying kendall's tau for modern metric meta-evaluation.

Freitag, M., Foster, G., Grangier, D., Ratnakar, V., Tan, Q., and Macherey, W. (2021a). Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.

Freitag, M., Rei, R., Mathur, N., Lo, C.-k., Stewart, C., Avramidis, E., Kocmi, T., Foster, G., Lavie, A., and Martins, A. F. T. (2022). Results of WMT22 metrics shared task: Stop using BLEU – neural metrics are better and more robust. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 46–68, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Freitag, M., Rei, R., Mathur, N., Lo, C.-k., Stewart, C., Foster, G., Lavie, A., and Bojar, O. (2021b). Results of the WMT21 metrics shared task: Evaluating metrics with expert-based human evaluations on TED and news domain. In *Proceedings of the Sixth Conference on Machine Translation*, pages 733–774, Online. Association for Computational Linguistics.

Graham, Y., Baldwin, T., Moffat, A., and Zobel, J. (2017). Can machine translation systems be evaluated by the crowd alone. *Nat. Lang. Eng.*, 23(1):3–30.

Kocmi, T., Bawden, R., Bojar, O., Dvorkovich, A., Federmann, C., Fishel, M., Gowda, T., Graham, Y., Grundkiewicz, R., Haddow, B., Knowles, R., Koehn, P., Monz, C., Morishita, M., Nagata, M., Nakazawa, T., Novák, M., Popel, M., and Popović, M. (2022). Findings of the 2022 conference on machine translation (WMT22). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 1–45, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Kocmi, T., Federmann, C., Grundkiewicz, R., Junczys-Dowmunt, M., Matsushita, H., and Menezes, A. (2021). To ship or not to ship: An extensive evaluation of automatic metrics for machine translation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 478–494, Online. Association for Computational Linguistics.

Koehn, P. (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. Association for Computational Linguistics.

Kudo, T. and Richardson, J. (2018). SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Lo, C.-k. (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. Association for Computational Linguistics.

Lommel, A., Uszkoreit, H., and Burchardt, A. (2014). Multidimensional quality metrics (mqm): A framework for declaring and describing translation quality metrics. *Tradumàtica*, (12):0455–463.

Ma, Q., Wei, J., Bojar, O., and Graham, Y. (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. Association for Computational Linguistics.

Marie, B. (2022). Yes, we need statistical significance testing. https://pub.towardsai.net/yes-we-need-statistical-significance-testing-927a8d21f9f0.

Marie, B., Fujita, A., and Rubino, R. (2021). Scientific credibility of machine translation research: A meta-evaluation of 769 papers. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7297–7306, Online. Association for Computational Linguistics.

Mathur, N., Baldwin, T., and Cohn, T. (2020a). Tangled up in BLEU: Reevaluating the evaluation of automatic machine translation evaluation metrics. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4984–4997, Online. Association for Computational Linguistics.

Mathur, N., Wei, J., Freitag, M., Ma, Q., and Bojar, O. (2020b). Results of the WMT20 metrics shared task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 688–725, Online. Association for Computational Linguistics.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). 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, Pennsylvania, USA. Association for Computational Linguistics.

Popović, M. (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.

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

Przybocki, M., Peterson, K., Bronsart, S., and Sanders, G. (2009). The nist 2008 metrics for machine translation challenge—overview, methodology, metrics, and results. *Machine Translation*, 23(2/3):71–103.

Rei, R., C. de Souza, J. G., Alves, D., Zerva, C., Farinha, A. C., Glushkova, T., Lavie, A., Coheur, L., and Martins, A. F. T. (2022). COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Rei, R., Stewart, C., Farinha, A. C., and Lavie, A. (2020). COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Robertson, T., Wright, F., and Dykstra, R. (1988). *Order Restricted Statistical Inference*. Probability and Statistics Series. Wiley.

Sellam, T., Das, D., and Parikh, A. (2020). BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Tan, L., Dehdari, J., and van Genabith, J. (2015). An awkward disparity between BLEU / RIBES scores and human judgements in machine translation. In *Proceedings of the 2nd Workshop on Asian Translation (WAT2015)*, pages 74–81, Kyoto, Japan. Workshop on Asian Translation.

Team, N., Costa-jussà, M. R., Cross, J., Çelebi, O., Elbayad, M., Heafield, K., Heffernan, K., Kalbassi, E., Lam, J., Licht, D., Maillard, J., Sun, A., Wang, S., Wenzek, G., Youngblood, A., Akula, B., Barrault, L., Gonzalez, G. M., Hansanti, P., Hoffman, J., Jarrett, S., Sadagopan, K. R., Rowe, D., Spruit, S., Tran, C., Andrews, P., Ayan, N. F., Bhosale, S., Edunov, S., Fan, A., Gao, C., Goswami, V., Guzmán, F., Koehn, P., Mourachko, A., Ropers, C., Saleem, S., Schwenk, H., and Wang, J. (2022). No language left behind: Scaling human-centered machine translation.

Wan, Y., Liu, D., Yang, B., Zhang, H., Chen, B., Wong, D., and Chao, L. (2022). UniTE: Unified translation evaluation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8117–8127, Dublin, Ireland. Association for Computational Linguistics.

Yan, Y., Wang, T., Zhao, C., Huang, S., Chen, J., and Wang, M. (2023). BLEURT has universal translations: An analysis of automatic metrics by minimum risk training. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5428–5443, Toronto, Canada. Association for Computational Linguistics.

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., and Artzi, Y. (2020). Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.