Re-Examining System-Level Correlations of Automatic Summarization Evaluation Metrics

Anonymous ACL submission

Abstract

How reliably an automatic summarization evaluation metric replicates human judgments of summary quality is quantified by system-level correlations. We identify two ways in which the definition of the system-level correlation is inconsistent with how metrics are used to evaluate systems in practice and propose changes to rectify this disconnect. First, we calculate the system score for an automatic metric using the full test set instead of the subset of summaries annotated by humans, which is currently standard practice. We demonstrate how this small change leads to more precise estimates of system-level correlations. Second, we propose to calculate correlations only on pairs of systems which are separated by differences in automatic scores that are commonly used to argue one system is of higher quality. This allows us to demonstrate that our best estimate of the correlation of ROUGE to human judgments is near 0 in realistic scenarios. Finally, the results from both analyses point to the need for future research to focus on developing more consistent and reliable human evaluations of summaries.[1]

1 Introduction

Automatic evaluation metrics are the most common method that researchers use to quickly and cheaply approximate how humans would annotate the quality of a summarization system (Lin, 2004; Louis and Nenkova, 2013; Zhao et al., 2019; Zhang et al., 2020; Deutsch et al., 2021a, among others). The quality of a metric — how similarly it replicates human annotations of systems — is quantified by calculating the correlation between the metric's scores and human judgments on a set of systems, known as the system-level correlation (Louis and Nenkova, 2013; Deutsch et al., 2021b).

Accurately estimating system-level correlations is critically important. Summarization researchers use automatic metrics during system development to make decisions about which ideas work and which do not, and systems from different research groups are ranked by automatic metrics to define which system is the "state-of-the-art." If we do not have precise estimates of metric quality, it is not clear how much trust the community should put in such evaluation methodologies.

At present, there are disconnects between how automatic metrics are evaluated and how they are used to evaluate systems. First, the metrics' scores which are used in practice are not the ones which are evaluated in system-level correlations: Researchers compare systems based on metric scores calculated on the entire test set but calculate scores for system-level correlations when evaluating metrics on a much smaller subset of annotated summaries. Second, metrics are evaluated in a setting that is much easier than how they are actually used. Metric correlations are calculated using systems that vary greatly in quality, whereas researchers compare new systems to recent work, which are likely to be very close in quality. Discriminating between two systems of similar quality is much harder than doing so between low and high quality systems.

In this work, we re-examine how system-level correlations are calculated and propose two independent changes to make the evaluation of metrics better aligned to how they are actually used to evaluate systems.

First, we propose to modify the system-level correlation definition to use the entire test set to calculate the system scores for automatic metrics instead of only the subset of summaries annotated by humans (§3). With this change, the scores which are used to compare systems are directly evaluated, and we further demonstrate how the precision of our estimate of system-level correlations improves as a result. Calculating system scores over a larger number of instances reduces the variance of the

[1] Our code will be released after publication.

scores, which results in confidence intervals (CIs) for the correlations that are 16-51% more narrow on average (§3.2).

Second, we redefine a high quality metric to be one for which a small difference in score reliably indicates a difference in quality (§4). Then, instead of calculating the correlation with all available system pairs, we only evaluate with pairs of systems whose automatic metric scores differ by some threshold. This allows us to show that a ROUGE-1 score difference of less than 0.5 between systems has almost no correlation to how humans would rank the same two systems according to our best estimates (§4.2). For two other metrics, BERTScore (Zhang et al., 2020) and QAEval (Deutsch et al., 2021a), we show their correlations calculated on system pairs of similar quality are much worse than under the standard correlation definition. These results cast doubt on how reliable automatic evaluation metrics are for measuring summarization system quality in realistic scenarios.

Although our two proposed changes are independent from each other, our experiments using both modifications point to the same direction for future work: In order to have more accurate estimates of metric correlations and to have more trustworthy system evaluations, future research needs to focus on developing more consistent and reliable protocols for human evaluations of summaries.

2 Background

Automatic evaluation metrics are most commonly used to argue that one summarization system is better than another, typically by showing that the value of a metric improves with the "better" system. How similarly automatic metrics replicate human judgments of system quality is quantified by system-level correlations as follows.

The summaries from N systems on M_a input documents are annotated by human judges $\mathcal{Z}$ and scored with an automatic metric $\mathcal{X}$. Then, the system-level correlation between $\mathcal{X}$ and $\mathcal{Z}$ is calculated as

$$r_{\text{SYS}} = \textsc{Corr}\left(\left\{\left(\frac{1}{M_a}\sum_{j}^{M_a} x_i^j, \frac{1}{M_a}\sum_{j}^{M_a} z_i^j\right)\right\}_{i=1}^{N}\right)$$

where x_i^j and z_i^j are the scores of $\mathcal{X}$ and $\mathcal{Z}$ for the summary produced by the i-th system on the j-th input document and $\textsc{Corr}$ is some correlation function. See Fig. 1 for an illustration of this calculation.

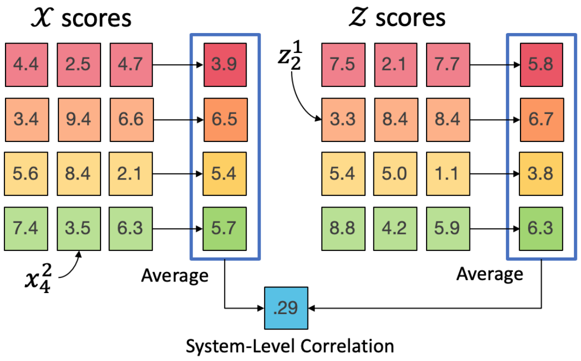

Figure 1: The system-level correlation is calculated between the average $\mathcal{X}$ and $\mathcal{Z}$ scores on a set of summarization systems. x_i^j and z_i^j are the scores for the summary produced by system i (represented by rows) on input document j (represented by columns).

In this work, we use Kendall's τ (the "b" variant[2]) as the correlation function because we are most concerned with a metric's ability to correctly determine whether one system is better than another since that is how metrics are used in practice. Kendall's τ is computed based on the number of system pairs out of $\binom{N}{2}$ which are ranked the same by $\mathcal{X}$ and $\mathcal{Z}$. It is defined as

$$\tau = \frac{P - Q}{\sqrt{(P + Q + T) \cdot (P + Q + U)}} \tag{1}$$

where P and Q are the number of pairs ranked the same or different by $\mathcal{X}$ and $\mathcal{Z}$, respectively, and T and U are the number of ties only in $\mathcal{X}$ or $\mathcal{Z}$, respectively.

Because the computation of r_{SYS} involves randomness — its value depends on which M_a input documents (and even which N systems) were used — it is only an approximate of the true correlation between $\mathcal{X}$ and $\mathcal{Z}$. As such, Deutsch et al. (2021b) proposed various methods for calculating confidence intervals for r_{SYS}. For instance, their BOOT-INPUTS method uses bootstrapping to repeatedly resample the M_a input documents used to calculate r_{SYS}, thereby calculating a confidence interval for the true r_{SYS} value for $\mathcal{X}$ and $\mathcal{Z}$.

Datasets The datasets that are used in this paper's analyses are SummEval (Fabbri et al., 2021) and REALSumm (Bhandari et al., 2020), two

[2] https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.kendalltau.html

recently collected datasets with human annotations for summary quality collected from the CNN/DailyMail dataset (Nallapati et al., 2016). SummEval has $M_a = 100$ summaries annotated with a summary relevance score for $N = 16$ systems. REALSumm has $M_a = 100$ summaries annotated with a Lightweight Pyramid score (Shapira et al., 2019) for $N = 25$ systems. We correlate the scores of the automatic metrics to these annotations. The CNN/DailyMail test split has $11,490$ instances.

Automatic Metrics Our experiments will analyze three different reference-based automatic evaluation metrics which were chosen because they were demonstrated to have the best correlations with human judgments on the SummEval and REALSumm datasets (Deutsch et al., 2021b). ROUGE-n (Lin, 2004) evaluates a generated summary by calculating an F_1 score on the number of n-grams it has in common with a human-written reference summary. BERTScore (Zhang et al., 2020) aligns the generated and reference summaries' tokens based on their BERT embeddings (Devlin et al., 2019) and calculates a score based on the similarity of the aligned tokens' embeddings. QA-Eval (Deutsch et al., 2021a) compares the two summaries by automatically generating questions from the reference and calculating what proportion of those questions are answered correctly by the generated summary.

3 Evaluating with All Available Instances

Although the above definition of the system-level correlation has been used by recent meta-evaluation studies of metrics (Bhandari et al., 2020; Fabbri et al., 2021; Deutsch et al., 2021b), there is a disconnect between how the automatic metrics are evaluated and how they are used in practice.

Researchers who develop summarization systems evaluate those systems with automatic metrics on all M_t test instances, not just the subset of M_a instances which were annotated by humans. Evaluating a system on a larger number of summaries may end up changing the system's score, which could potentially alter the overall ranking of a set of systems. Therefore, the rankings that are used by practitioners to determine system quality are not the ones which are being evaluated in the standard definition of system-level correlation.[3]

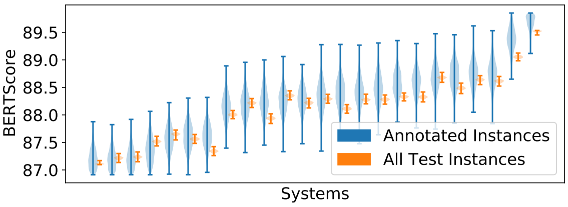

Figure 2: The bootstrapped 95% confidence intervals for the BERTScore of each system in the REALSumm dataset using M_a annotated instances in blue and M_t instances in orange. Evaluating systems with M_t instances leads to far better estimate of their true scores.

To that end, we propose to modify the correlation definition to use all M_t instances to calculate the system scores for the automatic metrics. That is (differences in bold):

$$r_{\text{SYS}} = \textsc{Corr}\left(\left\{\left(\frac{1}{\mathbf{M_t}}\sum_{j}^{\mathbf{M_t}} x_i^j, \frac{1}{M_a}\sum_{j}^{M_a} z_i^j\right)\right\}_{i=1}^{N}\right)$$

In practice with modern, large-scale datasets, this minor change could mean estimating system quality based on ≈10k inputs instead of around 100. This new definition now properly evaluates the way metrics are actually used by researchers.

We expect that scoring systems with M_t inputs instead of M_a should lead to a better estimate of the true automatic metric score, which would in turn result in a lower-variance estimate of the correlation between $\mathcal{X}$ and $\mathcal{Z}$ in the form of smaller confidence intervals for r_{SYS}. In the next sections, we carry out analyses to demonstrate that this is true.

3.1 Reducing Automatic Metric Variance

First, we empirically show that scoring systems with M_t instances instead of M_a does indeed reduce the variance of the estimate of the automatic metric scores and subsequently increases the stabilities of the system rankings.

Ideally, the $\mathcal{X}$ score for a system would be its "oracle" $\mathcal{X}$ score, equal to the expected value of $\mathcal{X}$ for a document sampled from the latent distribution over documents defined by the dataset. Since this cannot be calculated, it is approximated by averaging the $\mathcal{X}$ score on a sample (i.e., either the M_a or M_t input documents). Because $M_t \gg M_a$, we

[3] We suspect this methodology is an artifact of how system-level correlations were first calculated for summarization in the DUC shared tasks when the dataset sizes were small enough that $M_a = M_t$ (Dang and Owczarzak, 2008, among others).

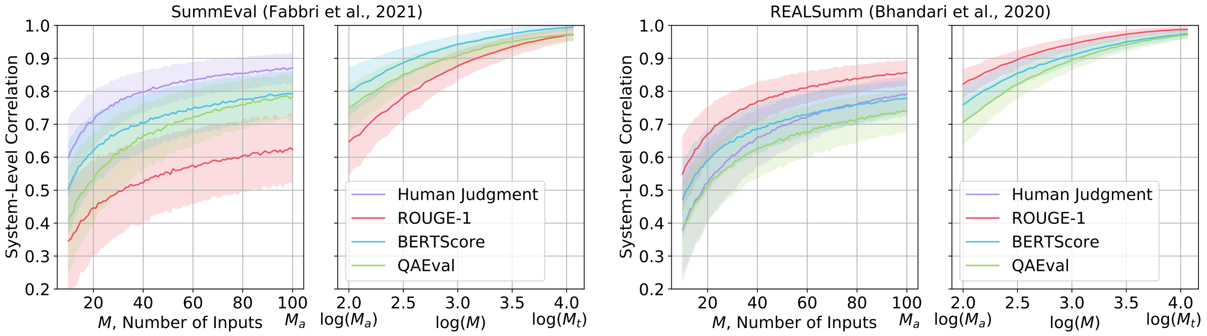

Figure 3: Bootstrapped estimates of the stabilities of the system rankings for automatic metrics and human annotations on SummEval (left) and REALSumm (right). The τ value quantifies how similar two system rankings would be if they were computed with two random sets of M input documents. When all M_t test instances are used, the automatic metrics' rankings become near constant. The error regions represent ± 1 standard deviation.

expect that the variance of this estimate using M_t inputs should be lower than when using M_a.

To quantify this, we calculated the variance of estimating the oracle $\mathcal{X}$ score using both M_a and M_t input documents via bootstrapping. We randomly sampled M input documents with replacement, recomputed the system scores, and calculated the variance of those scores over 1k iterations. For all three metrics on both datasets, we found around a 99% reduction in the variance when M_t inputs were used instead of M_a, clearly demonstrating that evaluating systems with M_t inputs results in a better estimate of the system scores. In Fig. 2, this is visualized for BERTScore on the REALSumm dataset.

However, because we are interested in evaluating the metrics' rankings, we also quantify how much of an effect this reduction in variance has on the stability of the system rankings induced by $\mathcal{X}$. Similarly to the system scores, there is an oracle ranking of systems for $\mathcal{X}$, equal to the ordering of systems by their respective oracle $\mathcal{X}$ scores. As the variance of the system score estimates decreases, the computed ranking of systems should begin to converge to the oracle $\mathcal{X}$ ranking. We aim to understand to what extent this happens if M_t instances are used for evaluation instead of M_a.

To quantify this notion, we calculate the Kendall's τ between two system rankings for $\mathcal{X}$ that were based on two sets of M input documents, each sampled with replacement from the set of available documents. This simulates how much the system rankings would change if the evaluation procedure was run twice, each time with M random input documents. This quantity is calculated 1k times for various values of M and plotted in Fig. 3.

As M approaches M_t, the automatic metrics' τ values approach 1, which is significantly higher than the respective values at M_a, typically around 0.6-0.8. A value near 1 means that the rankings calculated using M_t inputs are almost constant, implying the rankings have converged to the oracle ranking. Therefore, the reduction in variance from evaluating on M_t instances does indeed greatly stabilize the system rankings.

Fig. 3 also contains the same analysis performed for the human judgments $\mathcal{Z}$ in both datasets, although it is limited to a maximum of M_a input documents. We see that on both datasets the judgments' rankings are still quite variable, reaching a maximum of around 0.8-0.85 τ.

3.2 Confidence Interval Analysis

Next, we show that the improved estimate of system scores leads to a more precise estimate of r_{SYS} by demonstrating the widths of the confidence intervals for r_{SYS} decrease.

The confidence intervals for r_{SYS} calculated using bootstrapping methods proposed by Deutsch et al. (2021b) are rather wide. For instance, the 95% CI for ROUGE-2 on SummEval is $[-.09, .84]$, demonstrating a rather high level of uncertainty in its value. This is problematic because it means we do not have a good picture of how reliable automatic evaluation metrics are. Reducing the width of the CIs will help us better understand the true metric quality.

We suspect that the large width of the confidence interval is due to the variance of the system rankings of the automatic metrics and human judgments. The more unstable the rankings are with respect to

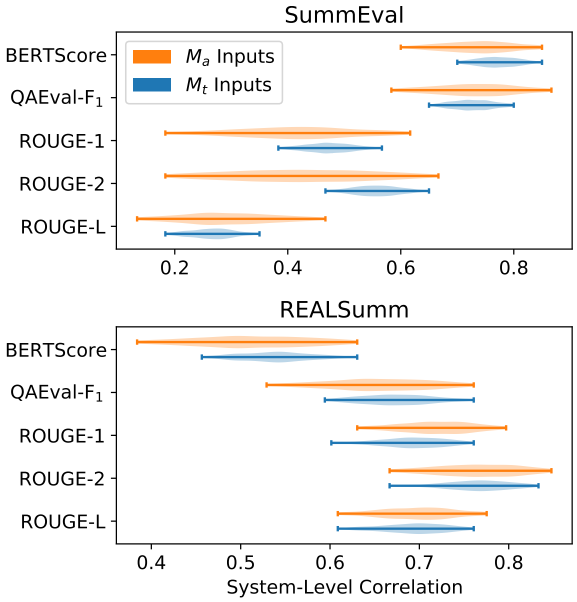

Figure 4: 95% confidence intervals for r_{SYS} calculated with the BOOT-INPUTS resampling method when the system rankings for the automatic metrics are calculated using only the judged data (orange) versus the entire test set (blue). Scoring systems with more summaries leads to better (more narrow) estimates of r_{SYS}.

the M inputs, the larger the variance of the estimate of r_{SYS} should be since very different system rankings would be compared on each bootstrapping iteration. Deutsch et al. (2021b) used M_a input documents to calculate their CIs. Therefore, we expect the improved stability of the automatic metric system rankings from evaluating on M_t instances should result in a more narrow confidence interval for r_{SYS} since some noise has been removed from this computation.

To demonstrate this, we calculated 95% CIs for r_{SYS} using the BOOT-INPUT method on SummEval and REALSumm using both M_a and M_t input documents, shown in Fig. 4. We find that the widths of the CIs shrank on average by 51% on SummEval and 16% on REALSumm. The largest decrease in width is in the ROUGE family of metrics on SummEval, likely because that metric and dataset combination saw the biggest improvement in ranking stability (see Fig. 3). Thus, the improved estimate of the system scores did result in more precise estimates of r_{SYS}. We repeated this analysis using the other bootstrapping methods proposed by Deutsch et al. (2021b), and the results are discussed in Appendix A.

3.3 Conclusions & Recommendations

By estimating system quality using automatic metrics on all available instances instead of only those which were annotated, we showed that the variances of the system scores and subsequent rankings reduce significantly, resulting in better estimates of r_{SYS}. Because this methodology additionally directly evaluates the system scores used by researchers, we recommend future work do the same.

In order to continue to improve the estimate of r_{SYS}, as much variance as possible needs to be removed from the system rankings. Evaluating systems using M_t instances removed a large amount of variance from the automatic metric rankings, but as demonstrated in Fig. 3, the human annotations still have a large amount of variance.

The human rankings' variances can either be reduced by annotating more summaries per system or making the annotations more consistent. Since the human rankings' stabilities in Fig. 3 are mostly beginning to plateau — especially for SummEval — it may be prohibitively expensive to collect a sufficient number of annotations to better stabilize the rankings (Wei and Jia, 2021). Therefore, we expect the more feasible solution is to improve the consistency of the human annotations, for example by better training the annotators or improving the annotation interface.

4 Evaluating with Realistic System Pairs

Next, we argue that the set of systems used to evaluate metrics is not reflective of how metrics are used in practice and propose a new system-level correlation variant to address this problem.

4.1 Evaluating with All System Pairs

The N systems which are used for calculating system-level correlations are typically those which participated in a shared task, as in DUC/TAC (Dang and Owczarzak, 2008, among others), or those which have been published in the previous 3-4 years (Bhandari et al., 2020; Fabbri et al., 2021). As such, they are typically rather diverse in terms of their qualities, both as rated by human annotators and automatic metrics.

The system scores of all of the systems in the REALSumm dataset as evaluated by humans and automatic metrics are shown in Fig. 5. Clearly, the scores are rather diverse. For example, the systems cluster into low, medium, and high quality groups (with an additional outlier) as evaluated by ROUGE.

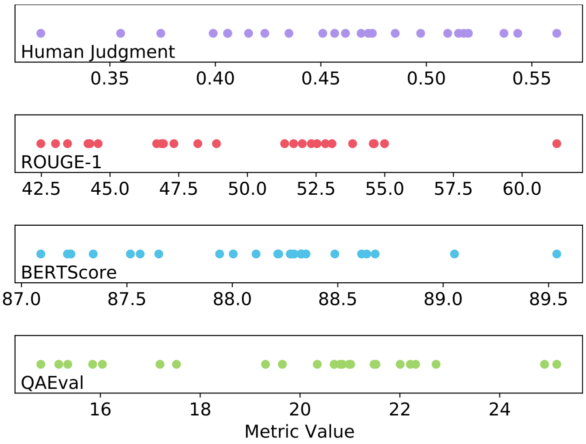

Figure 5: The systems (each represented by a point) on the two datasets (shown here for REALSumm) are rather diverse in quality as measured by both human judgments and automatic metrics.

A difference of around 5 ROUGE points between them is a rather large gap for ROUGE scores.

The standard definition for a high quality evaluation metric is one which correctly ranks a set of systems with respect to human judgments. As such, the implementation of the system-level correlation calculated with Kendall's τ will rank all N systems according to the human annotations and an automatic metric, then count how many pairs were ranked the same out of all $\binom{N}{2}$ pairs (see §2). As a consequence, even pairs of systems which are separated by a large margin according to the automatic metric — likely systems with a clear difference in quality — are included in the evaluation. Therefore, automatic metrics are rewarded for correctly ranking such "easy" system pairs.

4.2 Evaluating with Realistic Pairs

This standard evaluation setting does not reflect how summarization metrics are actually used by researchers. New systems are typically only slightly better than previous work. Based on a survey of summarization papers in *ACL conferences over the past few years, we found that the average improvement over baseline/state-of-the-art models that was reported on the CNN/Dailymail dataset was on average 0.5 ROUGE-1. It is rarely the case that the improvement in automatic metrics is very large. Therefore, evaluating metrics using pairs of systems which *are* separated by a large margin does not reflect the reality that metrics are very frequently used to compare those separated by a small margin. Including "easy" system pairs in the system-level correlation likely overestimates the quality of the metrics in settings which occur in practice.

To that extent, we redefine a high quality evaluation metric to be one for which a small difference in scores reliably indicates a difference in quality. We quantify this by proposing a variant of the system-level τ which is calculated between system pairs which are separated by a pre-defined automatic metric score margin. Instead of using all $\binom{N}{2}$ system pairs, only pairs whose difference in scores falls within the margin are used to calculate the system-level correlation. We denote this correlation variant as $r_{\text{SYS}}\Delta(\ell, u)$ where ℓ and u are the lower- and upper-bounds of the allowable differences in automatic metrics' scores. This would enable, for example, evaluating how well ROUGE correlates to human annotations on system pairs that are separated by 0.0-0.5 ROUGE points, thereby directly evaluating the scenario in which ROUGE is used to make decisions about system quality.

In Fig. 6 we report the $r_{\text{SYS}}\Delta(\ell, u)$ correlations for $\ell = 0.0$ and various values of u on both the SummEval and REALSumm datasets (more combinations of ℓ and u are included in Appendix B). That is, we evaluate r_{SYS} only on system pairs which are separated by at most an automatic score of u. The values of u were selected by picking the minimum u which would result in evaluating on $10\%, 20\%, \ldots, 100\%$ of the $\binom{N}{2}$ possible system pairs closest in score to be consistent across all three metrics.

The correlations for each metric on the system pairs closest in score are far lower than the correlations evaluated on all of the system pairs. For instance, the correlation of BERTScore on SummEval with the closest 20% of system pairs ($u \approx 0.2$) is only 0.42 compared to 0.77 under the standard definition of r_{SYS}. Thus, it is clear that the metrics are much less reliable approximations of human judgments when the system scores are close than was previously known. Evaluating on all possible system pairs leads to an overly optimistic view of automatic metric quality.

The $r_{\text{SYS}}\Delta(\ell, u)$ correlation of ROUGE for $\ell = 0.0$ and $u = 0.5$ — a typical improvement reported by researchers — is 0.08 and 0.0 on the SummEval and REALSumm datasets. Therefore, these results suggest the most popular summarization evaluation metric agrees with human judgments of system quality in realistic scenarios only slightly better than or equal to random chance.

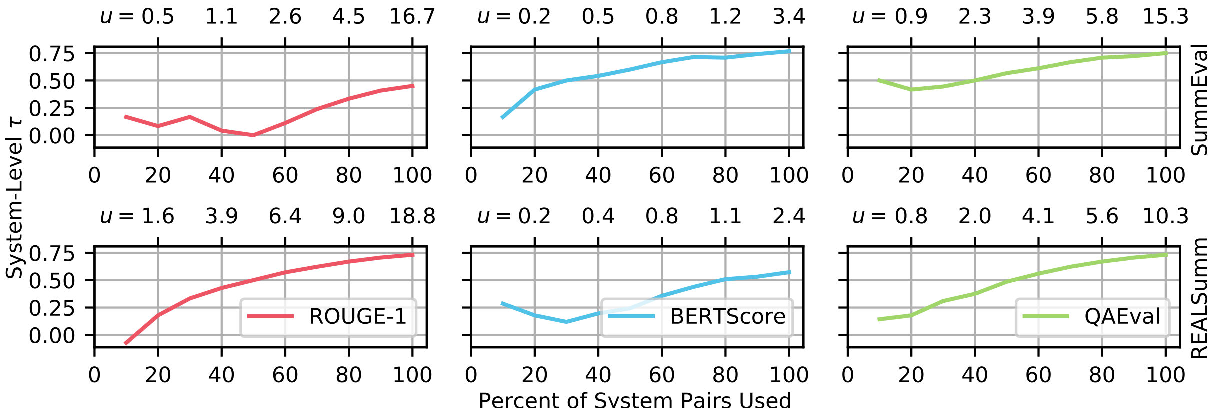

Figure 6: The $r_{\text{SYS}}\Delta(\ell, u)$ correlations on the SummEval (top) and REALSumm (bottom) datasets for $\ell = 0$ and various values of u (additional combinations of ℓ and u can be found in Appendix B). The u values were chosen to select the $10\%, 20\%, \ldots, 100\%$ of the pairs of systems closest in score. Each u is displayed on the top of each plot. For instance, 20% of the $\binom{N}{2}$ system pairs on SummEval are separated by < 0.5 ROUGE-1, and the system-level correlation on those pairs is around 0.08. As more systems are used in the correlation calculation, the allowable gap in scores between system pairs increases, and are therefore likely easier to rank, resulting in higher correlations.

This result also offers an explanation for why a naive metric such as ROUGE achieves moderately strong correlations under the standard definition of the system-level correlation (0.45 and 0.73 on SummEval and REALSumm) despite well known flaws and criticisms (Passonneau et al., 2005; Conroy and Dang, 2008; Deutsch and Roth, 2020, among others): It has benefited from an easy evaluation protocol. Despite its simplicity, it is not too surprising that a large gap of 5-10 ROUGE points actually does correctly rank system pairs. Most of its positive correlation comes from such easy examples.

4.3 Conclusions & Recommendations

One interpretation of the results in Fig. 6 is that the correlations in realistic settings are trending very low, meaning automatic metrics are not nearly sensitive enough to distinguish between systems with only minor differences in quality. This is problematic because this is the scenario in which metrics are most frequently used, and therefore they are not very reliable methods of evaluating summarization systems. However, it is not all bad news. Because the standard system-level τ values are moderately positive, consistent improvements in automatic metrics over time will likely result in better quality systems. Similarly to stochastic gradient descent, not every reported improvement is real, but on average over time, the quality does improve.

A more conservative interpretation of these results is that we cannot reach any definitive conclusions about the actual correlation values from this data because: (1) the available number of system pairs to calculate these correlations is rather small and (2) we may not have enough human annotations to accurately distinguish between similarly performing systems (Wei and Jia, 2021), so the ground-truth judgments may not be reliable. Unfortunately, these are our best estimates of the correlations with the available data. Not knowing how much we can trust automatic metrics is not a good outcome.

We recommend that proposals of new evaluation metrics also report correlations on system pairs with various differences in scores in addition to the standard system-level correlation definition. Reporting this information would better inform users of metrics about how likely humans would agree their observed improvement is real based on its value.

Future work should focus on improving the consistency of human annotations, both to better estimate the true system-level correlations and so they may be the definitive method of evaluating systems since automatic metrics may be unreliable. Finally, new data collection efforts for metric evaluation should consider collecting targeted pairwise judgments between systems which are close in quality to better evaluate realistic comparisons instead of direct assessments across a variety of systems of diverse quality.

5 Related Work

The methodology behind meta-evaluating summarization evaluation metrics was established during the DUC/TAC shared tasks (Dang and Owczarzak, 2008, among others). In addition to competitions for developing high-quality summarization systems, there were also shared tasks for creating automatic metrics that correlated well with human judgments. The benchmark datasets created during DUC/TAC were small in size by today's standards because they were manually collected multi-document summarization datasets, which are hard to create at scale. As such, all of the model-generated summaries on the full test set were annotated (so $M_a = M_t$; §3), unlike for current datasets which are too large to fully annotate.

Recently, there has been growing interest in revisiting the meta-evaluation of automatic evaluation metrics for summarization, in part due to the large differences between currently popular summarization datasets and those used in DUC/TAC. We view our work as continuing this direction of research.

Peyrard (2019) argues that current evaluation metrics do not work as well when they are used to evaluate high-performing systems compared to those which were evaluated in DUC/TAC.

Both Fabbri et al. (2021) and Bhandari et al. (2020) re-evaluated how well existing evaluation metrics work on the popular CNN/DailyMail dataset (Nallapati et al., 2016) by collecting judgments of summary quality using recent state-of-the-art systems. These datasets were used in our analyses. While the goal of these works was to identify which metrics correlated best with human judgments, our goal is to point out the ways in which the current methodology of meta-evaluating metrics is inconsistent with how they are used.

Then, the work of Deutsch et al. (2021b) proposed statistical methods for estimating and comparing correlation values. In contrast to our work, they provide statistical tools for analyzing correlations, whereas we propose new definitions of correlations.

Finally, Wei and Jia (2021) provided a theoretical analysis of the bias and variance of automatic and human evaluations of machine translations and summaries. Among their conclusions, they argue for evaluating metrics with pairwise accuracy (Kendall's τ) and that it may be prohibitively expensive to collect enough human annotations to distinguish between two systems with very similar quality. Our work further argues that metrics should be evaluated with a variant of Kendall's τ calculated using realistic system pairs (§4). Unfortunately, their results suggest that collecting enough human annotations to accurately measure how well automatic metrics perform in this setting may be very difficult.

6 Conclusion

In this work, we proposed two independent changes to how the system-level correlation of metrics is calculated to better align with how they are used to evaluate systems. Our analyses showed that these modifications led to lower-variance estimates of correlations and that commonly reported improvements in metric scores may not reliably predict how humans would annotate system quality. The results from both analyses point to the need for future work to develop more consistent and reliable methods of manually annotating summary quality.

References

Manik Bhandari, Pranav Narayan Gour, Atabak Ashfaq, Pengfei Liu, and Graham Neubig. 2020. Re-evaluating Evaluation in Text Summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9347–9359, Online. Association for Computational Linguistics.

John M. Conroy and Hoa Trang Dang. 2008. Mind the Gap: Dangers of Divorcing Evaluations of Summary Content from Linguistic Quality. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 145–152, Manchester, UK. Coling 2008 Organizing Committee.

Hoa Trang Dang and Karolina Owczarzak. 2008. Overview of the TAC 2008 Update Summarization Task. In *Proc. of the Text Analysis Conference (TAC)*.

Daniel Deutsch, Tania Bedrax-Weiss, and Dan Roth. 2021a. Towards Question-Answering as an Automatic Metric for Evaluating the Content Quality of a Summary. *Transactions of the Association for Computational Linguistics*, 9.

Daniel Deutsch, Rotem Dror, and Dan Roth. 2021b. A Statistical Analysis of Summarization Evaluation Metrics using Resampling Methods. *Transactions of the Association for Computational Linguistics*, 9.

Daniel Deutsch and Dan Roth. 2020. Understanding the Extent to which Summarization Evaluation Metrics Measure the Information Quality of Summaries. *ArXiv*, abs/2010.12495.

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, Minnesota. Association for Computational Linguistics.

Alexander Fabbri, Wojciech Kryscinski, Bryan McCann, R. Socher, and Dragomir Radev. 2021. SummEval: Re-evaluating Summarization Evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Annie Louis and Ani Nenkova. 2013. Automatically Assessing Machine Summary Content Without a Gold Standard. *Computational Linguistics*, 39:267–300.

Ramesh Nallapati, Bowen Zhou, Cícero Nogueira dos Santos, Çaglar Gülçehre, and Bing Xiang. 2016. Abstractive Text Summarization using Sequence-to-sequence RNNs and Beyond. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning, CoNLL 2016, Berlin, Germany, August 11-12, 2016*, pages 280–290. ACL.

Rebecca J Passonneau, Ani Nenkova, Kathleen McKeown, and Sergey Sigelman. 2005. Applying the Pyramid Method in DUC 2005. In *Proceedings of the document understanding conference (DUC 05), Vancouver, BC, Canada*.

Maxime Peyrard. 2019. Studying Summarization Evaluation Metrics in the Appropriate Scoring Range. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5093–5100, Florence, Italy. Association for Computational Linguistics.

Ori Shapira, David Gabay, Yang Gao, Hadar Ronen, Ramakanth Pasunuru, Mohit Bansal, Yael Amsterdamer, and Ido Dagan. 2019. Crowdsourcing Lightweight Pyramids for Manual Summary Evaluation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 682–687. Association for Computational Linguistics.

Johnny Wei and Robin Jia. 2021. The Statistical Advantage of Automatic NLG Metrics at the System Level. 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 6840–6854, Online. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 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.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. 2019. MoverScore: Text Generation Evaluating with Contextualized Embeddings and Earth Mover Distance. In *Proc. of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

A Additional Confidence Interval Results

In addition to the Boot-Inputs CI method proposed by Deutsch et al. (2021b), the authors also proposed Boot-Systems and Boot-Both. Each of the three methods makes assumptions about whether the set of N systems and M input documents are fixed or variable during the bootstrapping calculation. For instance, Boot-Inputs assumes the N systems are always the same and the M input documents are random, then subsequently resamples M input documents on each bootstrapping iteration to calculate the confidence interval. Boot-Systems does the opposite by resampling which N systems are used while holding the original M input documents fixed. Boot-Both assumes both the systems and inputs are variable.

Figures 7 and 8 contain the 95% CIs for ROUGE, BERTScore, and QAEval on the SummEval and REALSumm datasets using the Boot-Systems and Boot-Both methods calculated using all M_t test instances and only the M_a annotated instances (Boot-Inputs included in the main body of the paper, Fig. 4). The widths of the Boot-Both CIs decreased by 14% and 12%, whereas the Boot-Systems CIs only decreased by 1% and 6%.

The Boot-Systems widths likely decreased less because its estimation of r_{SYS} is not dependent on the variance of the system score estimates. Since the set of M input documents is fixed, the system scores do not change at all during bootstrapping, so increasing the number of summaries used to estimate those scores should not have a major effect on the estimation of r_{SYS}.

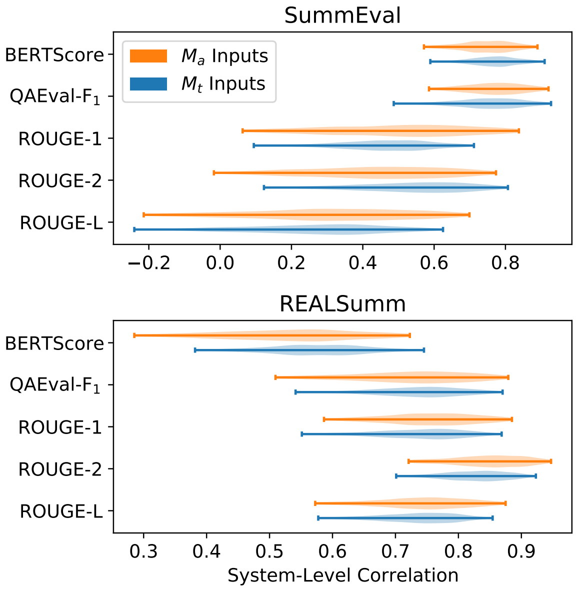

Figure 7: The 95% CIs calculated using the Boot-Systems bootstrapping method with M_a summaries in orange and M_t in blue.

B Additional $r_{\text{SYS}}\Delta(\ell, u)$ Results

Fig. 9 contains the $r_{\text{SYS}}\Delta(\ell, u)$ correlations for ROUGE, BERTScore, and QAEval for various combinations of ℓ and u on both the SummEval and REALSumm datasets. The first rows of each heatmap are plotted in Fig. 6.

We see that as the allowed score gap between system pairs is allowed to increase (i.e., adding "easier" pairs to rank), the correlation increases by a large margin over the correlation on pairs close in score. All of the metrics have nearly perfect correlation when the system pairs are separated by large margins.

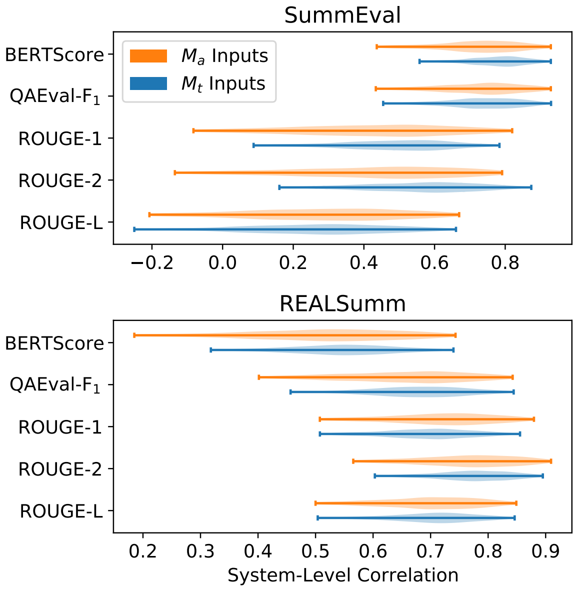

Figure 8: The 95% CIs calculated using the Boot-Both bootstrapping method with M_a summaries in orange and M_t in blue.

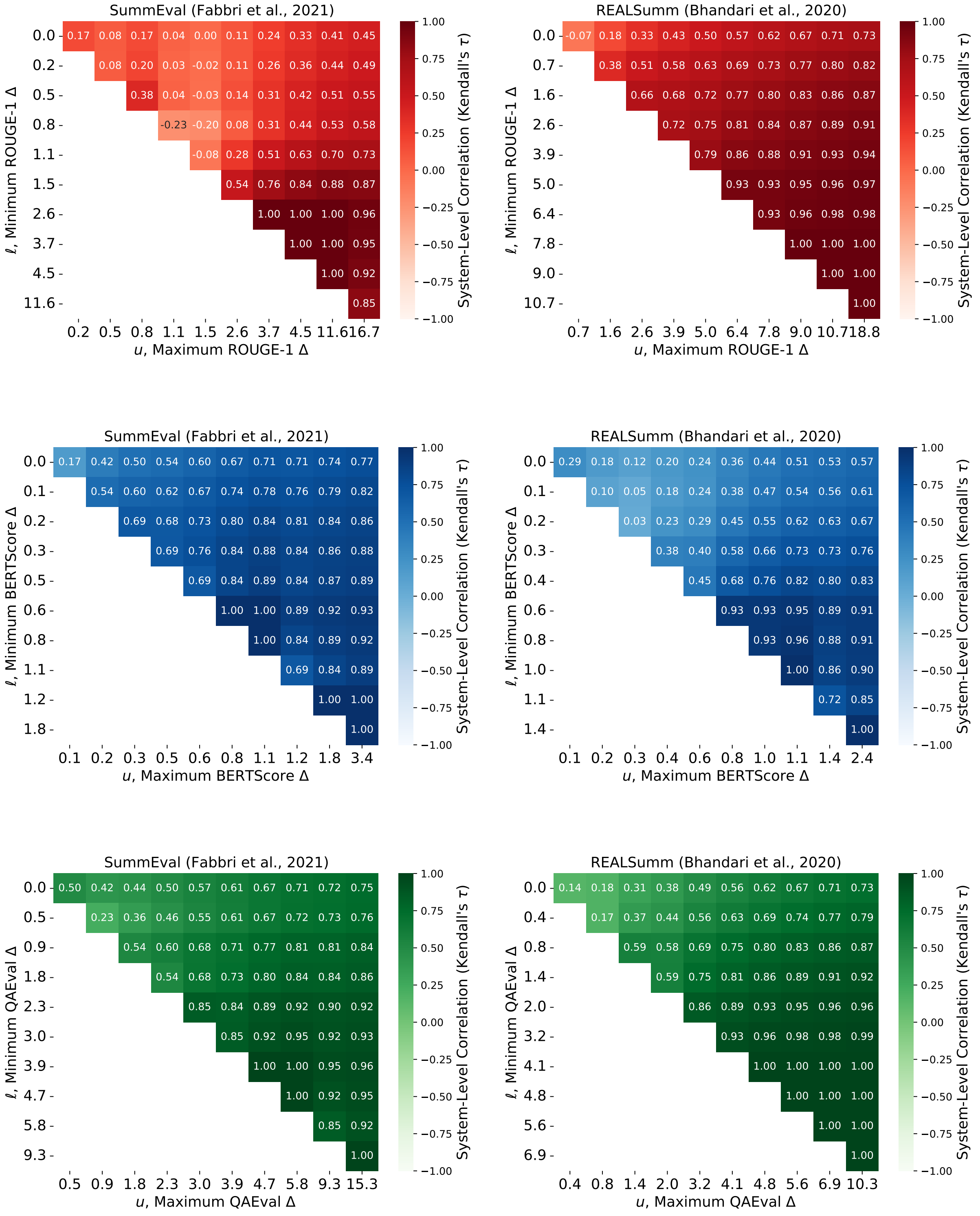

Figure 9: $r_{\text{SYS}}\Delta(\ell, u)$ correlations for various combinations of ℓ and u (see §4.2) for ROUGE (top), BERTScore (middle), and QAEval (bottom) on SummEval (left) and REALSumm (right). The values of ℓ and u were chosen so that each value in the heatmaps evaluates on 10% more system pairs than the value to its left. For instance, the first row evaluates on $10\%, 20\%, \ldots, 100\%$ of the system pairs. The second row evaluates on $10\%, 20\%, \ldots, 90\%$ of the system pairs, never including the 10% of pairs which are closest in score. The first row of each of the heatmaps is plotted in Fig. 6. The correlations on realistic score differences between systems are in the upper left portion of the heatmaps and contain the lowest correlations overall. Evaluating on all pairs is the top-rightmost entry, and the "easiest" pairs (those separated by a large score margin) are in the bottom right.