Bidimensional Leaderboards: Generate and Evaluate Language Hand in Hand

Jungo Kasai[*] **Keisuke Sakaguchi** **Ronan Le Bras** **Lavinia Dunagan**
Jacob Morrison **Alexander R. Fabbri** **Yejin Choi** **Noah A. Smith**
Paul G. Allen School of Computer Science & Engineering, University of Washington
Allen Institute for AI   Department of Linguistics, University of Washington
Salesforce Research
`billboard.nlp@gmail.com`

Abstract

Natural language processing researchers have identified limitations of evaluation methodology for generation tasks, with new questions raised about the validity of automatic metrics and of crowdworker judgments. Meanwhile, efforts to improve *generation models* tend to depend on simple n-gram overlap metrics (e.g., BLEU, ROUGE). We argue that new advances on models and metrics should each more directly benefit and inform the other. We therefore propose a generalization of leaderboards, **bidimensional leaderboards** (BILLBOARDs), that simultaneously tracks progress in language generation models and metrics for their evaluation. Unlike conventional *unidimensional* leaderboards that sort submitted systems by predetermined metrics, a BILLBOARD accepts both generators and evaluation metrics as competing entries. A BILLBOARD automatically creates an ensemble metric that selects and linearly combines a few metrics based on a global analysis across generators. Further, metrics are ranked based on their correlation with human judgments. We release four BILLBOARDs for machine translation, summarization, and image captioning.[1] We demonstrate that a linear ensemble of a few diverse metrics sometimes substantially outperforms existing metrics in isolation. Our mixed-effects model analysis shows that most automatic metrics, especially the reference-based ones, overrate machine over human generation, demonstrating the importance of updating metrics as generation models become stronger (and perhaps more similar to humans) in the future.

1 Introduction

Recent modeling advances have led to improved natural language generation in applications such as machine translation and summarization (Ng et al., 2019; Raffel et al., 2020; Brown et al., 2020, *inter alia*). This progress is typically measured with automatic scores, such as BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004), executed by modeling researchers themselves. These metrics allow for fast, inexpensive development cycles. They were adopted based on reported correlations with human judgments at the time the metrics were introduced, but it has since been established that the correspondence can collapse when models of different types are compared (Callison-Burch et al., 2006) or models become increasingly powerful (Ma et al., 2019; Edunov et al., 2020).

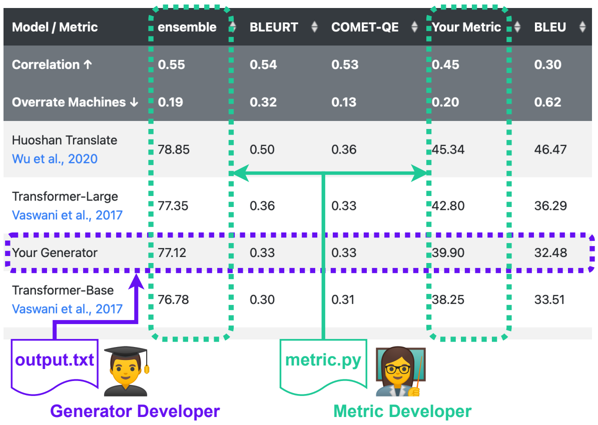

Figure 1: Bidimensional leaderboard (BILLBOARD). When a generator developer submits output text (`output.txt`), BILLBOARD computes all metric scores. When a metric developer submits an executable program (e.g., `metric.py`), BILLBOARD computes correlation with the human judgments, updates the ensemble metric (§2.2), and measures how much the metric overrates machines (§2.3).

Meanwhile, many evaluation metrics that improve correlation with human judgments have been proposed (Clark et al., 2019; Zhang et al., 2020b; Sellam et al., 2020; Hessel et al., 2021, *inter alia*), but this progress has yet to be broadly adopted by the community of researchers focused on advancing models. Indeed, consistent with prior meta-evaluations (Marie et al., 2021), we found that 68%

[*]Work was done during an internship at AI2.
[1]`https://nlp.cs.washington.edu/billboard/`.

Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 3540 - 3557
July 10-15, 2022 ©2022 Association for Computational Linguistics

of the machine translation papers from NAACL and ACL 2021 evaluated their models solely by BLEU, and only 5% measured the performance using recent metrics with contextual representations such as COMET (Rei et al., 2020). Similarly, automatic evaluation in 66% of the summarization papers was done only in terms of ROUGE.[2] We believe this separation between generation modeling and automatic evaluation represents a missed opportunity for each subcommunity to more rapidly benefit from the advances of the other.

We therefore propose an abstraction of conventional leaderboards, **bidimensional leaderboards** (BILLBOARDS), that simultaneously facilitates progress in natural language generation and its evaluation (Fig. 1). A BILLBOARD accepts two types of submissions related to a given task and dataset: **generators** and **metrics**. Unlike conventional leaderboards, model ranking is not tied to a predetermined set of metrics; the generators are ranked based on the metric that currently correlates best with human judgments. Metric submissions are ranked by their correlations to human judgments, and each is stored as an executable program, which will then be used to evaluate future generation submissions. Our BILLBOARD includes a sparse regression that selects and linearly combines three existing metrics, revealing complementary strengths. All leaderboard scores are readily reproducible, allowing research on generation models and automatic metrics to benefit from each other.

We release four BILLBOARD interfaces (`https://nlp.cs.washington.edu/billboard/`) spanning three generation tasks: the WMT20 EN-DE and WMT20 ZH-EN machine translation tasks (Barrault et al., 2020), the CNNDM summarization task (Hermann et al., 2015), and the MSCOCO image captioning task (Lin et al., 2014).

Key Findings Using the collective analyses of BILLBOARDS, our main findings are as follows.

- A simple linear combination of a few (diverse) metrics can sometimes improve correlation. This finding quantifies complementary effects of different metrics and encourages metric developers to seek out aspects of generated text quality not yet measured by existing metrics.

- Using linear mixed-effects models, we find that most automatic metrics, especially conventional, reference-based ones such as BLEU and ROUGE, *overrate* machines over humans in all tasks. This result provides further support for the claim that the metrics should be continually evaluated and updated as our generation models become stronger (and perhaps, closer to humans).

- When only one reference is available per instance, COMET-QE (a strong *referenceless* metric with crosslingual contextual representations; Rei et al., 2020) achieves higher correlation with human judgments than all reference-based metrics. This raises a concern about the current standard evaluation practice in machine translation and summarization that uses reference-based metrics with a single reference per instance.

- Our findings confirm many others who report that recent metrics achieve substantially higher correlation with human judgments than popular metrics like BLEU and ROUGE in BILLBOARDS. We believe these older metrics continue to be used mainly because modeling researchers value consistency and accessibility of evaluation practice over long periods of time. BILLBOARDS provide a way to maintain long-term comparability of system output while also drawing better conclusions about system quality, using advances in evaluation. All generators continue to be evaluated with new metrics on BILLBOARDS.

2 Bidimensional Leaderboards

We propose BILLBOARDS to simultaneously drive progress in natural language generation and its evaluation, which are often disconnected in current research. We first describe the general framework (§2.1) and the automatic analyses they provide (§2.2-2.3). We then discuss our design choices (§2.4) and the rubric-based, human judgment data necessary to initialize BILLBOARDS (§2.5).

2.1 BILLBOARD Framework

The leaderboard paradigm has driven research on state-of-the-art model performance on many tasks in various fields (e.g., ImageNet, Russakovsky et al., 2015; SQuAD, Rajpurkar et al., 2016). As applications and tasks become more diverse, however, the conventional leaderboard paradigm presents a serious challenge: the assumption becomes too strong that predetermined, automatic metrics can

[2] We examined all papers whose title contains "machine translation" and "summarization." See Appendix A for details.

reliably score the system performance *over time*. In particular, scores from automatic metrics often diverge from human judgments in language generation tasks, especially when models become increasingly powerful (Ma et al., 2019).

Much recent work proposed new evaluation metrics that improve correlations with human judgments in certain generation tasks (Clark et al., 2019; Zhang et al., 2020b; Sellam et al., 2020; Hessel et al., 2021, *inter alia*), but most developers of generation models are not benefiting from them (See Appendix A for our analysis of papers from NAACL/ACL 2021). From the perspective of generation model developers, it is not clear which of these many metrics in the literature is most reliable in which generation task or dataset, resulting in community-wide overuse of long-standing metrics like BLEU and ROUGE. Developers of evaluation metrics, on the other hand, are missing the opportunity to apply their metrics to new generation models and compare them with the existing ones. We propose BILLBOARDS that bridge this gap between generation modeling and evaluation development.

Generators, Metrics, and Scores A BILLBOARD for a language generation task consists of sets of generators and evaluation metrics: $\mathcal{G} = \{G_i\}_{i=1}^{I}, \mathcal{M} = \{M_j\}_{j=1}^{J}$. Each generator G_i takes as input X_k (e.g., source text in machine translation) and generates text: $Y_{i,k} = G_i(X_k)$. A metric M_j assigns a score to each generated text given the generation input and the corresponding set of references $\mathcal{R}_k$: $s_{i,j,k} = M_j(Y_{i,k}, \mathcal{R}_k, X_k)$. The last two arguments to the function are optional; some metrics do not require references (i.e., *referenceless* or *quality estimation* metrics) or the generation input (e.g., BLEU). We then compute the aggregate score $s_{i,j}$ by averaging $s_{i,j,k}$ over K test examples.

Rankings In contrast to standard leaderboards, BILLBOARDS have a dynamic set of evaluation metrics, and generators are not ranked by a predefined metric. We first rank the metrics by measuring their correlations to human judgments as commonly done in the generation evaluation literature (Zhang et al., 2020b; Sellam et al., 2020). Let $h_{i,k}$ be a human score for $Y_{i,k}$ (i.e., output from generator G_i on input X_k). We compute the instance-level Pearson correlation for every metric M_j between $h_{i,k}$ and $s_{i,j,k}$ (M_j score for $Y_{i,k}$). All metrics are ranked by their correlations. We then use the top metric M_{j^*} to rank the generators in the descending order of s_{i,j^*}. We defer our discussions on alternative design choices (§2.4) and human evaluations (§2.5). We note, however, that the overall framework of BILLBOARDS still holds regardless of these decisions.

2.2 Ensemble of Metrics

So far, we have assumed that metrics are used individually in isolation, but BILLBOARDS provide a unique opportunity to examine metrics collectively. Different metrics can capture different aspects of generation quality; even if a metric is not sufficiently informative in isolation, it might reflect an important aspect of text quality that the existing metrics overlook. Here we consider a straightforward and interpretable ensemble of metrics using a regression model with ℓ_1 regularization (Tibshirani, 1994). Let the ensemble's score be

$$\hat{h}_{i,k} = \sum_{j=1}^{J} w_j \cdot s_{i,j,k},$$

where w_j is a scalar coefficient associated with the jth metric and the intercept term is suppressed. We optimize the vector of coefficients $\mathbf{w}$ with the pairs of output text and a human score $\{Y_{i,k}, h_{i,k}\}_{k=1}^{K}$ from the test data:

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \sum_{k=1}^{K} \left(h_{i,k} - \hat{h}_{i,k} \right)^2 + \lambda \|\mathbf{w}\|_1$$

The ℓ_1 regularization produces sparse coefficients and improves interpretability by removing highly correlated metrics. Moreover, it avoids the need for practitioners to run many metrics to obtain an ensemble score when used outside our BILLBOARDS. Our goal for the ensemble is to provide a useful signal to the research community, rather than to achieve the best possible correlation with human judges at a given time; we tune λ to get three nonzero coefficients. Every metric is standardized by its mean and standard deviation on the test data.

Similar to the individual metrics, we rank this ensemble metric by its correlation to the human judgments. To make fair comparisons, we simulate situations where the ensemble is applied to a newly submitted generator that has no human evaluations. Specifically, we perform cross validation that holds out the human judgments for each generator G_i and runs regression on the rest; we then apply these I regression models to the corresponding held-out data and calculate the overall correlation. We will see that the ensemble metric outperforms all individual metrics in some cases, suggesting that different metrics can capture different aspects.

Reproduciblity The ensemble metric is updated every time a new metric is submitted (Fig. 1). For reproducibility, we keep track of every past ensemble metric with a signature that indicates its coefficients, λ, and input metrics in the backend. Similar to SACREBLEU (Post, 2018), model developers can report the signature for easy replication of their scores from the ensemble metric.[3] Further, all generation outputs are saved on the leaderboards, so model developers can download outputs from all past models and compare in any way.

2.3 Mixed-Effects Model Analysis

Recent work (Kasai et al., 2022) observed that automatic metrics tend to *overrate* machine-generated text over human one on the MSCOCO image captioning task (Chen et al., 2015). This problem is particularly severe in conventional metrics that are based on n-gram overlap such as BLEU and CIDEr (Vedantam et al., 2015). This raises a significant concern about the continuous use of these conventional metrics in generation tasks as models become increasingly powerful (and more similar to humans); those metrics unintentionally discourage researchers from developing human-like, strong generation models. To quantify this undesirable property, we propose a linear mixed-effects model that compares the two groups of machine- and human-generated text. The underlying model assumes that $s_{i,j,k}$, the score from metric M_j for generator G_i and test example k, can be expressed as (the intercept term is suppressed for brevity):

$$s_{i,j,k} = \beta_0^j \mathbb{1}\{G_i \text{ is machine}\} + \beta_1^j h_{i,k} + \gamma_k + \epsilon_{i,j,k}$$

where γ_k is the random effect for example k, and $\epsilon_{i,j,k}$ is Gaussian noise. Intuitively, β_0^j measures how much metric M_j *overrates* machine generation over human one, compared against the human judgment $h_{i,k}$. $\beta_0^j = 0$ means being neutral, and indeed we will find that β_0^j is significantly positive in most cases (§4). We standardize all metric scores over the test samples to compare the size of β_0^j. We apply the *lme4* package (Bates et al., 2015).

2.4 Design Choices and Discussion

In our current setup, we make several design choices for metrics and their rankings:

- **M.1** Metrics are expected to positively correlate with the generation output quality.
- **M.2** By default, metrics are ranked based on their instance-level Pearson correlations with human judgments. We also compute and present their system-level Kendall rank correlations.
- **M.3** When available, reference-based metrics use multiple references per instance.

M.1 implies that we need to take the negative of metric scores that are intended to negatively correlate (e.g., TER, Snover et al., 2006). This normalization is also done in WMT metric competitions (Callison-Burch et al., 2007, 2008, *inter alia*).

While instance-level correlations are commonly used to evaluate and compare automatic metrics for various language generation tasks (Sellam et al., 2020; Fabbri et al., 2021; Hessel et al., 2021, *inter alia*), there are several alternatives to M.2. For example, Pearson, Spearman's rank, or Kendall rank correlations can be used on a system (i.e., generator) level (Callison-Burch et al., 2007; Macháček and Bojar, 2014; Mathur et al., 2020b). However, such system-level correlations would substantially reduce data points to compare automatic scores, resulting in many ties in the ranking. Spearman's and Kendall rank correlations become brittle when multiple generators are similar in overall output quality; penalizing a metric for swapping two similar generators is misleading (Macháček and Bojar, 2014). Moreover, if a metric can perform well on an instance level, it can be used to augment human judgments by, for example, flagging likely wrong ratings (Mathur et al., 2020b). Thus, we encourage researchers to develop metrics that correlate well with human judgments on an instance level. Prior work also points out other problems in ranking metrics like *outlier effects* where outlier systems have a disproportionately large effect on the overall correlation (Mathur et al., 2020a,b). We therefore assume M.2 in the current version of BILLBOARDS, but this can be modified in a future version.

M.3 is supported by our experimental results in §4 that multiple references substantially improve reference-based metrics, and a single reference is often insufficient to outperform strong referenceless metrics. Some metrics have specifications for multiple references (e.g., BLEU, CIDEr). In the other cases, we evaluate outputs against every reference and take the maximum score, following prior work on image captioning evaluation (Zhang et al.,

[3] E.g., ensemble.wmt20-zh-en+refs.AB+version.1.

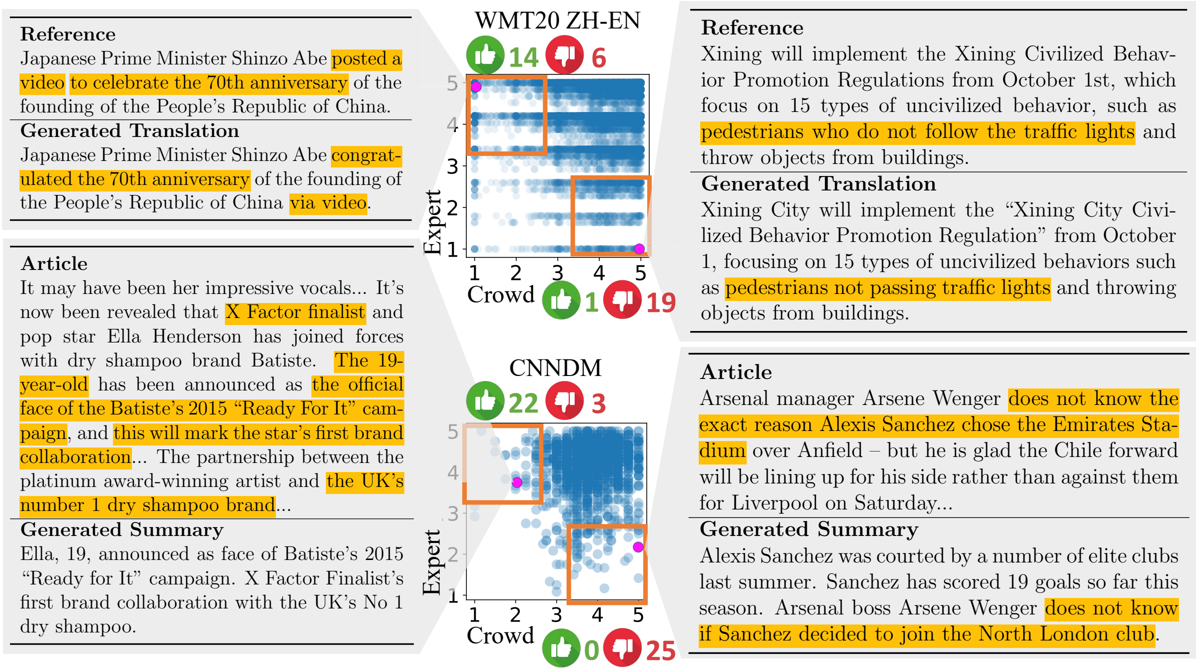

Figure 2: Comparisons and meta-evaluations of crowdworker and rubric-based, expert evaluations for WMT20 ZH-EN and CNNDM summarization. Every dot represents one test instance that is evaluated by the same numbers of experts and crowdworkers (one for WMT20 ZH-EN and three for CNNDM) for fair comparisons. We randomly sampled instances with diverging evaluations in two areas ☐ and conducted binary meta-evaluations (good 👍 or bad quality 👎). **Meta-evaluations agree more with the expert evaluations**: 👍 > 👎 in the upper left squares and 👎 > 👍 in the lower right squares. We suspect that the highlighted text might have caused the disagreement.

2020b; Hessel et al., 2021).[4]

2.5 Human Evaluation

Human evaluations are required to initialize BILLBOARDS; they are used to rank metrics, train the metric ensembling model, and assess how much each metric overrates machines. Recent work, however, points out problems when evaluations are done by crowdworkers even when extensive quality controls are performed (Gillick and Liu, 2010; Toral et al., 2018; Freitag et al., 2021; Clark et al., 2021). Freitag et al. (2021) show that rubric-based machine translation evaluations by professional translators led to substantially different generator rankings from the crowdsource evaluations in WMT 2020 (Barrault et al., 2020), where WMT participants or Amazon Mechanical Turkers directly assess each translation's adequacy by a single score (*direct assessment*). These crowdworker evaluations depend highly on individual annotators' discretion and understanding of the annotation scheme (Freitag et al., 2021; Clark et al., 2021), making it difficult to decompose, interpret, and validate (Kasai et al., 2022). Moreover, these direct assessment scores make it difficult to interpret evaluation results for downstream applications where some aspects are particularly important (e.g., accessibility for people with visual impairments in image captioning, Gleason et al., 2020; gender bias in machine translation, Stanovsky et al., 2019).

Motivated by this line of work, we perform meta-evaluations to compare crowdsourced and rubric-based expert evaluations. Fig. 2 plots overall scores for test examples from WMT20 ZH-EN (Barrault et al., 2020; Freitag et al., 2021) and CNNDM summarization (Fabbri et al., 2021). Each instance is evaluated by averaging the same number of crowdworkers and expert scores for fair comparisons. We see that substantially many instances fall into disagreement: crowdworkers give much higher scores than experts (lower right square) or the reverse (upper left square). We sample and shuffle 20/25 examples from either type and ask a meta-evaluator to make a binary decision (good 👍 or bad quality 👎).[5] Meta-evaluations agree more with the expert evaluations (e.g., 22 and 0 👍 in the upper left

[4] Intuitively, the maximum score measures the distance to the closest out of equally valid generations.

[5] The meta-evaluations were done by a bilingual speaker (WMT20 ZH-EN) and the first author of this paper (CNNDM).

and lower right squares for CNNDM, respectively). In the examples on the left, crowdworkers fail to properly assess a valid translation with different structure than the reference (*posted a video to celebrate* vs. *congratulated via video*) or a summary that combines information from different parts of the article. The examples on the right illustrate that crowdworkers can be fooled by inaccurate yet fluent generations (*does not know the reason* vs. *does not know if Sanchez decided*). Given this result, we decide to initialize our BILLBOARDS with rubric-based expert evaluations for all generation tasks. We still encourage future work to explore ways to improve crowdsourced evaluations for scalability.

3 Experiments

Having established the framework, we set up BILLBOARDS for three natural language generation tasks: machine translation, summarization, and image captioning. To maximize the performance of reference-based metrics, we use as many references as possible for each task. See §4 for an analysis on the effect of varying numbers of references.

3.1 Tasks

Machine Translation We experiment with two language pairs from the WMT 2020 news translation task (Barrault et al., 2020): Chinese→English (**WMT20 ZH-EN**) and English→German (**WMT20 EN-DE**). We use outputs from all submitted translation systems.[6] These two language pairs have expert, rubric-based scores (MQM) from Freitag et al. (2021) for a subset of 10 submitted systems, including the top-performing systems and human translations. Each output sentence is evaluated by three professional translators. Following Freitag et al. (2021), the three scores are averaged to get an instance-level score.

We use all human translations available as a reference set for reference-based metrics. Concretely, every test instance in WMT20 ZH-EN has two translations provided by different human translation services: Human-A and Human-B (Barrault et al., 2020). In addition to Human-A and Human-B, WMT20 EN-DE provides a translation that is created by linguists who are asked to paraphrase Human-A and Human-B as much as possible (Human-P, Freitag et al., 2020). These paraphrased translations are shown to increase correlations with human judgments by mitigating the *translationese effect* and diversifying the reference when the generation quality is measured by reference-based metrics (Freitag et al., 2020).

Along with all submitted generators in WMT20 ZH-EN and WMT20 EN-DE, we train three transformer baselines with the `fairseq` library (Ott et al., 2019) and place them in our BILLBOARDS: **transformer-base**, **transformer-large**, and **transformer-large-ensemble** with similar hyperparameters (e.g., 6-layer encoder and decoder) to the ones trained on the WMT16 EN-DE data in Vaswani et al. (2017).[7] These baselines allow researchers to compare their translation models without resource-intensive techniques such as backtranslation (Sennrich et al., 2016a), model ensembling, and deep encoders (Kasai et al., 2021a). These techniques are all used in top-performing systems of WMT20 (Wu et al., 2020a; Kiyono et al., 2020) but might be infeasible in many research settings. See Appendix B for a list of all hyperparameters for the baselines.

Summarization We use the CNN/DailyMail corpus (**CNNDM**, Hermann et al., 2015; Nallapati et al., 2016). We use the standard train/dev./test split and 24 models from Fabbri et al. (2021). 100 test articles are annotated with 10 summaries written by humans (Kryscinski et al., 2019). For those 100 articles, rubric-based, expert evaluations for 18 generators, including human-written highlights, are provided by Fabbri et al. (2021).[8] Each output summary is evaluated by three experts along four dimensions: *coherence* (collective quality of all summary sentences), *consistency* (factual alignment with the article, penalizing for hallucinations), *fluency* (quality of the individual sentences), and *relevance* (selection of important content). An instance-level score is computed by averaging scores over all these categories and the three experts. Note that this aggregation method can be modified, depending on the downstream task of interest (Kasai et al., 2022). All 10 human-written summaries are used

[6] https://www.statmt.org/wmt20/translation-task.html.

[7] Data and models are available at https://github.com/jungokasai/billboard/tree/master/baselines.

[8] Some of the outputs are lowercased and/or tokenized. In these cases, we apply the NLTK detokenizer (Bird et al., 2009) and/or the Stanford CoreNLP truecaser (Manning et al., 2014). We encourage, however, future model developers to provide clean, untokenized output to improve the reproducibility and transparency of evaluation results (Post, 2018; Kasai et al., 2022).

Dataset	$\|\mathcal{G}\|$	$\|\mathcal{M}\|$	Top Gen.	Single Metrics Top Metric	Single Metrics Corr.	Ensemble of Metrics Linear Combination	Ensemble of Metrics Corr.
WMT20 ZH-EN	19	15	Huoshan	COMET	0.55	1.72·COMET-QE+1.48·COMET+1.21·BLEURT	**0.61**
WMT20 EN-DE	17	11	Tohoku	COMET	0.49	1.19·COMET+0.36·COMET-QE+0.02·Prism-ref	**0.51**
CNNDM	26	15	Lead-3	COMET	**0.41**	2.85·COMET+0.26·COMET-QE+0.01·BERTScore	0.29
MSCOCO	4	15	VinVL-large	RefCLIP-S	**0.45**	2.08·RefCLIP-S+1.51·RefOnlyC+0.82·CIDEr	**0.45**

Table 1: Summary of BILLBOARDs as of Jan. 10, 2022. Huoshan: Wu et al. (2020a); Tohoku: Kiyono et al. (2020); VinVL-large: Zhang et al. (2021); COMET, COMET-QE: Rei et al. (2020); BLEURT: Sellam et al. (2020); Prism-ref: Thompson and Post (2020); BERTScore: Zhang et al. (2020b); RefCLIP-S: Hessel et al. (2021); RefOnlyC: Kasai et al. (2022). COMET-QE is a *referenceless* metric. BLEURT is specifically trained to evaluate into-English translations. RefCLIP-S uses image features unlike most metrics for image captioning. RefOnlyC removes image features from RefCLIP-S and only uses reference text features from CLIP (Radford et al., 2021).

as the reference set for reference-based metrics.[9]

Image Captioning We use the **MSCOCO** dataset (Lin et al., 2014) that consists of everyday-scene photos sampled from Flickr. Every image is annotated with five captions written by crowdworkers (Chen et al., 2015). We apply the standard *Karpathy split* (Karpathy and Fei-Fei, 2015). For each of 500 test images, rubric-based evaluations (THUMB 1.0) are available for five systems, including one caption from a crowdworker (Kasai et al., 2022). Similar to machine translation and summarization, we use all five crowdworker captions as a reference set for reference-based metrics.

3.2 Mixed-Effects Models

Our mixed-effects model analyzes how much every automatic metric overrates machines over humans (§2.3). This means that we need to free up one human generation per instance to measure its scores in the reference-based metrics. For machine translation, we score Human-B using the reference set of Human-A (WMT20 ZH-EN) or Human-A and Human-P (WMT20 EN-DE). For CNNDM, we use concatenated highlights as human-generated summaries and use the 10 human-written summaries from Kryscinski et al. (2019) as the reference. We follow Kasai et al. (2022) for MSCOCO and score their randomly-selected *Human* caption using the other four as the reference. As the distinction between the *reference* and *human generation* (e.g., Human-A vs. Human B on WMT20 ZH-EN) is arbitrary, we found that swapping the roles would still lead to similar results (See Appendix E).

[9] Prior work used a concatenation of author-written highlights as a reference, but here we do not add it to the reference set. This is because these highlights are sometimes noisy (e.g., containing URLs) or lack coherence (Fabbri et al., 2021).

4 Results and Analysis

Here we discuss the current results and make several key observations about the state of language generation evaluation. Table 1 summarizes the four BILLBOARDs. It is particularly noteworthy that COMET, a metric designed for machine translation, achieves the best correlation on the CNNDM summarization task as well. COMET evaluates the similarity between the crosslingual representations from XLM-RoBERTa (Conneau et al., 2020) for input text and its translation candidate. But these crosslingual representations can, of course, be used *monolingually* for English summarization. This illustrates an additional benefit of BILLBOARDs that centralize different generation tasks and find surprising task transferability of learning-based metrics. See Appendices B and C for lists of all participating generators and metrics.

Ensemble Metric The rightmost section of Table 1 shows the chosen metrics and their coefficients in the ensemble (§2.2). On the machine translation tasks, the ensemble metric outperforms the top individual metric.[10] In particular, we see a substantial gain of 0.06 points in WMT20 ZH-EN. The *referenceless* metric of COMET-QE is selected both for WMT20 ZH-EN and WMT20 EN-DE, suggesting complementary effects of diverse metrics. To further test this hypothesis, we perform ablations that drop one out of the three metrics at a time (Table 2). We see that only dropping COMET-QE would result in a decrease in the correlation score. This implies that the reference-

[10] We found a major reason for the anomaly in CNNDM; an outlier generator that does not use the standard CNNDM training data (the GPT-2 zero-shot model; Ziegler et al., 2019) has a disproportionately large effect on the regression models. The ensemble metric outperformed the top individual metric of COMET when the zero-shot model was removed.

less metric provides important information that the others do not.

Removed Metric	–	COMET	COMET-QE	BLEURT
Correlation	0.61	0.61	0.57	0.61

Table 2: Ensemble ablation studies on WMT20 ZH-EN. Only removing COMET-QE leads to a correlation drop. See Appendix D for the other datasets.

Mixed-Effects Models Seen in Table 3 are the results from our analysis that measures how much metrics *overrate* machines over humans (§2.3). We see that the fixed-effect coefficient β_0 is significantly positive in most cases. Referenceless metrics tend to have smaller coefficients. This can be due to the more diverse nature of human text than machine-generated text; reference-based metrics give a low score to human text that differs from the references even if it is of high quality. The conventional n-gram overlap-based metrics (BLEU, ROUGE, and CIDEr) have particularly large coefficients. These results suggest that the evaluation practice should be regularly updated as our generation models become stronger (and perhaps, more similar to human generation) in the future. Note that unlike the other tasks, "human-generated text" for CNNDM summarization is an automatic concatenation of author highlights, which contains substantial noise (Fabbri et al., 2021). This might explain the neutral and negative coefficients.

ZH-EN	COMET	COMET-QE	BLEURT	BLEU
	$0.27_{\pm 0.02}$	$0.13_{\pm 0.01}$	$0.32_{\pm 0.02}$	$0.62_{\pm 0.02}$
EN-DE	COMET	COMET-QE	Prism-ref	BLEU
	$0.08_{\pm 0.03}$	$-0.17_{\pm 0.02}$	$0.44_{\pm 0.02}$	$0.33_{\pm 0.03}$
CNNDM	COMET	COMET-QE	BERTScore	ROUGE-L
	$-0.17_{\pm 0.12}$	$0.02_{\pm 0.11}$	$-0.04_{\pm 0.12}$	$0.33_{\pm 0.13}$
COCO	RefCLIP-S	RefOnlyC	CIDEr	CLIP-S
	$0.09_{\pm 0.06}$	$0.24_{\pm 0.06}$	$0.43_{\pm 0.06}$	$-0.04_{\pm 0.05}$

Table 3: β_0 fixed-effect coefficients from the linear mixed-effects models, quantifying how much automatic metrics **overrate** machines over humans, relative to human raters. $\beta_0 = 0$ is neutral, and statistical significance is indicated by red (positive) or blue text (negative). The subscripts indicate 90% confidence intervals. Three metrics that correlate best with the human judgments are shown as well as one popular metric. COMET-QE and CLIP-S are *referenceless*. See §E for the other metrics.

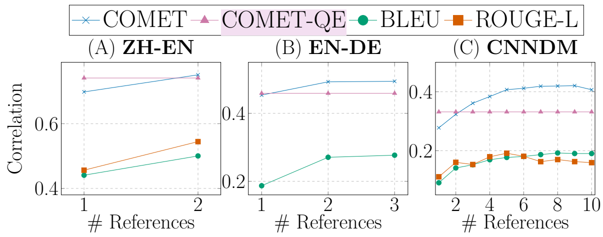

Figure 3: Correlations with varying numbers of references. In all cases, one reference is not sufficient to outperform the referenceless COMET-QE metric. The default ROUGE assumes English input.

Effects of the Number of References Fig. 3 plots correlations over varying numbers of references. COMET was the top-performing *reference-based* metric regardless of the number of references, but we observe that it underperforms the refererenceless metric when only one reference is given. Model performance in machine translation and summarization is commonly measured by applying reference-based metrics against one reference per instance in the research community. Our finding thus raises a further concern about the current evaluation practice. Finally, we see that popular choices of BLEU and ROUGE metrics have much lower correlations than the recent metrics over various numbers of references, in line with the recent studies (Mathur et al., 2020a, *inter alia*).

5 Related and Future Work

Related Benchmarks WMT organizes the metric competition track in parallel with the translation task every year (Mathur et al., 2020b; Barrault et al., 2020, *inter alia*). Participants submit automatic scores for the translation outputs from the parallel translation task. Unfortunately, most of these new metrics are not used by subsequent machine translation work, perhaps because they are tested solely against the concurrent translation submissions and it is up to model developers to execute or even implement new metrics. The GEM workshop (Gehrmann et al., 2021) conducts extensive analysis of models and evaluation methods over a wide set of generation tasks. BILLBOARDS ease the burden through *standard* leaderboard experience where generator developers only need to upload generation outputs for the test split. BILLBOARDS also offer automatic ensembling of metrics and quantify the diversity that a new metric adds. The human-in-the-loop GENIE leaderboard (Khashabi et al., 2021) centralizes crowdsourced evaluations for generation tasks. The current BILLBOARD setup is based on rubric-based, expert evaluation data from previous work, but future work can explore ways to improve crowdsourced evaluations and use them to update BILLBOARDS.

From Bidimensional to Multidimensional BILLBOARDS lend themselves to a natural extension: *multidimensional leaderboards*. In particular, generation models have more aspects than generation quality, such as training and inference efficiency, sample efficiency, and robustness. These aspects are often ignored in the current leaderboard paradigm but are important to better serving practitioners' needs (Schwartz et al., 2019; Ethayarajh and Jurafsky, 2020; Mishra and Arunkumar, 2021). There are ongoing modeling and benchmarking efforts especially for efficient machine translation (Heafield et al., 2020; Peng et al., 2021; Kasai et al., 2021b, *inter alia*). We leave this extension to future work and specifically target the gap between generation modeling and evaluation.

6 Conclusion

We introduced BILLBOARDS, a simple yet powerful generalization of leaderboards that bridges the gap between generation modeling and evaluation research. We established and released four BILLBOARDS on machine translation, summarization, and image captioning tasks. We demonstrated that their built-in analysis of metric ensembling and mixed-effects modeling revealed key insights into the current state of natural language generation and its evaluation methods. BILLBOARDS allow for a standard leaderboard experience both on the modeling and evaluation sides. We invite submissions from researchers through our website.

Acknowledgements

We thank Kyunghyun Cho, Elizabeth Clark, Jesse Dodge, Saadia Gabriel, Michal Guerquin, Daniel Khashabi, Swaroop Mishra, Jamie Morgenstern, Phoebe Mulcaire, Hao Peng, Sofia Serrano, the ARK group at UW, the Mosaic team at the Allen Institute for AI, and the anonymous reviewers for their helpful feedback on this work. We also thank Xinyan Yu for WMT Chinese-English meta-evaluations and Muyun Yang and Shujian Huang for sharing with us the CWMT training data. This work was supported in part by the DARPA MCS program through NIWC Pacific (N66001-19-2-4031) and Google Cloud Compute.

References

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. SPICE: semantic propositional image caption evaluation. In *Proc. of ECCV*.

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proc. of CVPR*.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proc. of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*.

Loïc Barrault, Magdalena Biesialska, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Matthias Huck, Eric Joanis, Tom Kocmi, Philipp Koehn, Chi-kiu Lo, Nikola Ljubešić, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Santanu Pal, Matt Post, and Marcos Zampieri. 2020. Findings of the 2020 conference on machine translation (WMT20). In *Proc. of WMT*.

Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*.

Rachel Bawden, Giorgio Maria Di Nunzio, Cristian Grozea, Inigo Jauregi Unanue, Antonio Jimeno Yepes, Nancy Mah, David Martinez, Aurélie Névéol, Mariana Neves, Maite Oronoz, Olatz Perez-de Viñaspre, Massimo Piccardi, Roland Roller, Amy Siu, Philippe Thomas, Federica Vezzani, Maika Vicente Navarro, Dina Wiemann, and Lana Yeganova. 2020. Findings of the WMT 2020 biomedical translation shared task: Basque, Italian and Russian as new additional languages. In *Proc. of WMT*.

Steven Bird, Evan Klein, and Edward Loper. 2009. *Natural Language Processing with Python*. Cambridge University Press.

Florian Böhm, Yang Gao, Christian M. Meyer, Ori Shapira, Ido Dagan, and Iryna Gurevych. 2019. Better rewards yield better summaries: Learning to summarise without references. In *Proc. of EMNLP*.

Léo Bouscarrat, Antoine Bonnefoy, Thomas Peel, and Cécile Pereira. 2019. STRASS: A light and effective method for extractive summarization based on sentence embeddings. In *Proc. of ACL*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda

Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Proc. of NeurIPS*.

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2007. (meta-) evaluation of machine translation. In *Proc. of WMT*.

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2008. Further meta-evaluation of machine translation. In *Proc. of WMT*.

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Re-evaluating the role of Bleu in machine translation research. In *Proc. of EACL*.

Tanfang Chen, Weiwei Wang, Wenyang Wei, Xing Shi, Xiangang Li, Jieping Ye, and Kevin Knight. 2020. DiDi's machine translation system for WMT2020. In *Proc. of WMT*.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C. Lawrence Zitnick. 2015. Microsoft COCO captions: Data collection and evaluation server.

Yen-Chun Chen and Mohit Bansal. 2018. Fast abstractive summarization with reinforce-selected sentence rewriting. In *Proc. of ACL*.

Elizabeth Clark, Tal August, Sofia Serrano, Nikita Haduong, Suchin Gururangan, and Noah A. Smith. 2021. All that's 'human' is not gold: Evaluating human evaluation of generated text. In *Proc. of ACL*.

Elizabeth Clark, Asli Celikyilmaz, and Noah A. Smith. 2019. Sentence mover's similarity: Automatic evaluation for multi-sentence texts. In *Proc. of ACL*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proc. of ACL*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proc. of NAACL*.

Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. 2019. Unified language model pre-training for natural language understanding and generation. In *Proc. of NeurIPS*.

Yue Dong, Yikang Shen, Eric Crawford, Herke van Hoof, and Jackie Chi Kit Cheung. 2018. BanditSum: Extractive summarization as a contextual bandit. In *Proc. of EMNLP*.

Sergey Edunov, Myle Ott, Marc'Aurelio Ranzato, and Michael Auli. 2020. On the evaluation of machine translation systems trained with back-translation. In *Proc. of ACL*.

Kawin Ethayarajh and Dan Jurafsky. 2020. Utility is in the eye of the user: A critique of NLP leaderboards. In *Proc. of EMNLP*.

Matan Eyal, Tal Baumel, and Michael Elhadad. 2019. Question answering as an automatic evaluation metric for news article summarization. In *Proc. of NAACL*.

Alexander R Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. SummEval: Re-evaluating summarization evaluation. *TACL*.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *TACL*.

Markus Freitag, David Grangier, and Isaac Caswell. 2020. BLEU might be guilty but references are not innocent. In *Proc. of EMNLP*.

Sebastian Gehrmann, Tosin P. Adewumi, Karmanya Aggarwal, Pawan Sasanka Ammanamanchi, Aremu Anuoluwapo, Antoine Bosselut, Khyathi Raghavi Chandu, Miruna-Adriana Clinciu, Dipanjan Das, Kaustubh D. Dhole, Wanyu Du, Esin Durmus, Ondrej Dusek, Chris Emezue, Varun Gangal, Cristina Garbacea, Tatsunori Hashimoto, Yufang Hou, Yacine Jernite, Harsh Jhamtani, Yangfeng Ji, Shailza Jolly, Dhruv Kumar, Faisal Ladhak, Aman Madaan, Mounica Maddela, Khyati Mahajan, Saad Mahamood, Bodhisattwa Prasad Majumder, Pedro Henrique Martins, Angelina McMillan-Major, Simon Mille, Emiel van Miltenburg, Moin Nadeem, Shashi Narayan, Vitaly Nikolaev, Rubungo Andre Niyongabo, Salomey Osei, Ankur P. Parikh, Laura Perez-Beltrachini, Niranjan Ramesh Rao, Vikas Raunak, Juan Diego Rodriguez, Sashank Santhanam, João Sedoc, Thibault Sellam, Samira Shaikh, Anastasia Shimorina, Marco Antonio Sobrevilla Cabezudo, Hendrik Strobelt, Nishant Subramani, Wei Xu, Diyi Yang, Akhila Yerukola, and Jiawei Zhou. 2021. The GEM benchmark: Natural language generation, its evaluation and metrics. In *Proc. of GEM*.

Sebastian Gehrmann, Yuntian Deng, and Alexander Rush. 2018. Bottom-up abstractive summarization. In *Proc. of EMNLP*.

Ulrich Germann. 2020. The University of Edinburgh's submission to the German-to-English and English-to-German tracks in the WMT 2020 news translation and zero-shot translation robustness tasks. In *Proc. of WMT*.

Dan Gillick and Yang Liu. 2010. Non-expert evaluation of summarization systems is risky. In *Proc. of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*.

Cole Gleason, Amy Pavel, Emma McCamey, Christina Low, Patrick Carrington, Kris M. Kitani, and Jeffrey P. Bigham. 2020. Twitter a11y: A browser extension to make twitter images accessible. In *Proc. of CHI*.

Han Guo, Ramakanth Pasunuru, and Mohit Bansal. 2018. Soft layer-specific multi-task summarization with entailment and question generation. In *Proc. of ACL*.

Jeremy Gwinnup and Tim Anderson. 2020. The AFRL WMT20 news translation systems. In *Proc. of WMT*.

Hany Hassan, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mengnan Li, Shujie Liu, Tie-Yan Liu, Renqian Luo, Arul Menezes, Tao Qin, Frank Seide, Xu Tan, Fei Tian, Lijun Wu, Shuangzhi Wu, Yingce Xia, Dongdong Zhang, Zhirui Zhang, and Ming Zhou. 2018. Achieving human parity on automatic Chinese to English news translation.

Kenneth Heafield, Hiroaki Hayashi, Yusuke Oda, Ioannis Konstas, Andrew Finch, Graham Neubig, Xian Li, and Alexandra Birch. 2020. Findings of the fourth workshop on neural generation and translation. In *Proc. of WNGT*.

Karl Moritz Hermann, Tomás Kociský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Proc. of NeurIPS*.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. CLIPScore: A reference-free evaluation metric for image captioning. In *Proc. of EMNLP*.

Wan-Ting Hsu, Chieh-Kai Lin, Ming-Ying Lee, Kerui Min, Jing Tang, and Min Sun. 2018. A unified model for extractive and abstractive summarization using inconsistency loss. In *Proc. of ACL*.

Yichen Jiang and Mohit Bansal. 2018. Closed-book training to improve summarization encoder memory. In *Proc. of EMNLP*.

Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proc. of CVPR*.

Jungo Kasai, Nikolaos Pappas, Hao Peng, James Cross, and Noah A. Smith. 2021a. Deep encoder, shallow decoder: Reevaluating non-autoregressive machine translation. In *Proc. of ICLR*.

Jungo Kasai, Hao Peng, Yizhe Zhang, Dani Yogatama, Gabriel Ilharco, Nikolaos Pappas, Yi Mao, Weizhu Chen, and Noah A. Smith. 2021b. Finetuning pretrained transformers into RNNs. In *Proc. of EMNLP*.

Jungo Kasai, Keisuke Sakaguchi, Lavinia Dunagan, Jacob Morrison, Ronan Le Bras, Yejin Choi, and Noah A. Smith. 2022. Transparent human evaluation for image captioning. In *Proc. of NAACL*.

Daniel Khashabi, Gabriel Stanovsky, Jonathan Bragg, Nicholas Lourie, Jungo Kasai, Yejin Choi, Noah A. Smith, and Daniel S. Weld. 2021. GENIE: A leaderboard for human-in-the-loop evaluation of text generation.

Shun Kiyono, Takumi Ito, Ryuto Konno, Makoto Morishita, and Jun Suzuki. 2020. Tohoku-AIP-NTT at WMT 2020 news translation task. In *Proc. of WMT*.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proc. of ACL Demo and Poster Sessions*.

Wojciech Kryscinski, Nitish Shirish Keskar, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Neural text summarization: A critical evaluation. In *Proc. of EMNLP*.

Wojciech Kryściński, Romain Paulus, Caiming Xiong, and Richard Socher. 2018. Improving abstraction in text summarization. In *Proc. of EMNLP*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proc. of ACL*.

Zuchao Li, Hai Zhao, Rui Wang, Kehai Chen, Masao Utiyama, and Eiichiro Sumita. 2020. SJTU-NICT's supervised and unsupervised neural machine translation systems for the WMT20 news translation task. In *Proc. of WMT*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Proc. of Text Summarization Branches Out*.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: common objects in context. In *Proc. of ECCV*.

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. In *Proc. of EMNLP*.

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 *Proc. of WMT*.

Matouš Macháček and Ondřej Bojar. 2014. Results of the WMT14 metrics shared task. In *Proc. of WMT*.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proc. of ACL System Demonstrations*.

Benjamin Marie, Atsushi Fujita, and Raphael Rubino. 2021. Scientific credibility of machine translation research: A meta-evaluation of 769 papers. In *Proc. of ACL*.

Nitika Mathur, Timothy Baldwin, and Trevor Cohn. 2020a. Tangled up in BLEU: Reevaluating the evaluation of automatic machine translation evaluation metrics. In *Proc. of ACL*.

Nitika Mathur, Johnny Wei, Markus Freitag, Qingsong Ma, and Ondřej Bojar. 2020b. Results of the WMT20 metrics shared task. In *Proc. of WMT*.

Fandong Meng, Jianhao Yan, Yijin Liu, Yuan Gao, Xianfeng Zeng, Qinsong Zeng, Peng Li, Ming Chen, Jie Zhou, Sifan Liu, and Hao Zhou. 2020. WeChat neural machine translation systems for WMT20. In *Proc. of WMT*.

Swaroop Mishra and Anjana Arunkumar. 2021. How robust are model rankings : A leaderboard customization approach for equitable evaluation. In *Proc. of AAAI*.

Alexander Molchanov. 2020. PROMT systems for WMT 2020 shared news translation task. In *Proc. of WMT*.

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 *Proc. of CoNLL*.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Ranking sentences for extractive summarization with reinforcement learning. In *Proc. of NAACL*.

Nathan Ng, Kyra Yee, Alexei Baevski, Myle Ott, Michael Auli, and Sergey Edunov. 2019. Facebook FAIR's WMT19 news translation task submission. In *Proc. of WMT*.

Csaba Oravecz, Katina Bontcheva, László Tihanyi, David Kolovratnik, Bhavani Bhaskar, Adrien Lardilleux, Szymon Klocek, and Andreas Eisele. 2020. eTranslation's submissions to the WMT 2020 news translation task. In *Proc. of WMT*.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proc. of NAACL Demonstrations*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proc. of ACL*.

Ramakanth Pasunuru and Mohit Bansal. 2018. Multi-reward reinforced summarization with saliency and entailment. In *Proc. of NAACL*.

Hao Peng, Nikolaos Pappas, Dani Yogatama, Roy Schwartz, Noah A. Smith, and Lingpeng Kong. 2021. Random feature attention. In *Proc. of ICLR*.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proc. of WMT*.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proc. of WMT*.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proc. of WMT*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *JLMR*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proc. of EMNLP*.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proc. of EMNLP*.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. 2015. ImageNet large scale visual recognition challenge. *IJCV*.

Roy Schwartz, Jesse Dodge, Noah A. Smith, and Oren Etzioni. 2019. Green AI. *CACM*.

Thomas Scialom, Sylvain Lamprier, Benjamin Piwowarski, and Jacopo Staiano. 2019. Answers unite! unsupervised metrics for reinforced summarization models. In *Proc. of EMNLP*.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proc. of ACL*.

Thibault Sellam, Dipanjan Das, and Ankur P Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proc. of ACL*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Improving neural machine translation models with monolingual data. In *Proc. of ACL*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. Neural machine translation of rare words with subword units. In *Proc. of ACL*.

Eva Sharma, Luyang Huang, Zhe Hu, and Lu Wang. 2019. An entity-driven framework for abstractive summarization. In *Proc. of EMNLP*.

Tingxun Shi, Shiyu Zhao, Xiaopu Li, Xiaoxue Wang, Qian Zhang, Di Ai, Dawei Dang, Xue Zhengshan, and Jie Hao. 2020. OPPO's machine translation systems for WMT20. In *Proc. of WMT*.

Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proc. of AMTA*.

Gabriel Stanovsky, Noah A. Smith, and Luke Zettlemoyer. 2019. Evaluating gender bias in machine translation. In *Proc. of ACL*.

Brian Thompson and Matt Post. 2020. Automatic machine translation evaluation in many languages via zero-shot paraphrasing. In *Proc. of EMNLP*.

Robert Tibshirani. 1994. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society, Series B*.

Antonio Toral, Sheila Castilho, Ke Hu, and Andy Way. 2018. Attaining the unattainable? reassessing claims of human parity in neural machine translation. In *Proc. of WMT*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proc. of NeurIPS*.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. CIDEr: Consensus-based image description evaluation. In *Proc. of CVPR*.

Elena Voita, Rico Sennrich, and Ivan Titov. 2019. When a good translation is wrong in context: Context-aware machine translation improves on deixis, ellipsis, and lexical cohesion. In *Proc. of ACL*.

Weiyue Wang, Jan-Thorsten Peter, Hendrik Rosendahl, and Hermann Ney. 2016. CharacTer: Translation edit rate on character level. In *Proc. of WMT*.

Daimeng Wei, Hengchao Shang, Zhanglin Wu, Zhengzhe Yu, Liangyou Li, Jiaxin Guo, Minghan Wang, Hao Yang, Lizhi Lei, Ying Qin, and Shiliang Sun. 2020. HW-TSC's participation in the WMT 2020 news translation shared task. In *Proc. of WMT*.

Liwei Wu, Xiao Pan, Zehui Lin, Yaoming Zhu, Mingxuan Wang, and Lei Li. 2020a. The Volctrans machine translation system for WMT20. In *Proc. of WMT*.

Shuangzhi Wu, Xing Wang, Longyue Wang, Fangxu Liu, Jun Xie, Zhaopeng Tu, Shuming Shi, and Mu Li. 2020b. Tencent neural machine translation systems for the WMT20 news translation task. In *Proc. of WMT*.

Yuxiang Wu and Baotian Hu. 2018. Learning to extract coherent summary via deep reinforcement learning. In *Proc. of AAAI*.

Jiacheng Xu and Greg Durrett. 2019. Neural extractive text summarization with syntactic compression. In *Proc. of EMNLP*.

Lei Yu, Laurent Sartran, Po-Sen Huang, Wojciech Stokowiec, Domenic Donato, Srivatsan Srinivasan, Alek Andreev, Wang Ling, Sona Mokra, Agustin Dal Lago, Yotam Doron, Susannah Young, Phil Blunsom, and Chris Dyer. 2020. The DeepMind Chinese–English document translation system at WMT2020. In *Proc. of WMT*.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2020a. PEGASUS: Pre-training with extracted gap-sentences for abstractive summarization. In *Proc. of ICML*.

Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. 2021. VinVL: Making visual representations matter in vision-language models. In *Proc. of CVPR*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020b. BERTScore: Evaluating text generation with BERT. In *Proc. of ICLR*.

Xingxing Zhang, Mirella Lapata, Furu Wei, and Ming Zhou. 2018. Neural latent extractive document summarization. In *Proc. of EMNLP*.

Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason J. Corso, and Jianfeng Gao. 2020. Unified vision-language pre-training for image captioning and VQA. In *Proc. of AAAI*.

Qingyu Zhou, Nan Yang, Furu Wei, Shaohan Huang, Ming Zhou, and Tiejun Zhao. 2018. Neural document summarization by jointly learning to score and select sentences. In *Proc. of ACL*.

Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul F. Christiano, and Geoffrey Irving. 2019. Fine-tuning language models from human preferences.

Appendices

A Case Studies of Evaluation Practice

Fig. 4 depicts breakdowns of evaluation metrics used in the papers on machine translation and summarization from NAACL and ACL 2021. We examined all papers whose title contains "machine translation" and "summarization." We see the clear gap between generation modeling and evaluation research; most researchers do not take advantage of recent metrics that correlate better with human judgments.

B Participating Generators

Here we list the generators submitted in the initial BILLBOARDS.

B.1 WMT20 ZH-EN

Hyperparameter	Value
label smoothing	0.1
# max tokens	1024
dropout rate	0.1
encoder embedding dim	512
encoder ffn dim	2048
# encoder attn heads	8
decoder embedding dim	512
decoder ffn dim	2048
# decoder attn heads	8
max source positions	1024
max target positions	1024
Adam lrate	5×10^{-4}
Adam β_1	0.9
Adam β_2	0.98
lr-scheduler	inverse square
warm-up lr	1×10^{-7}
# warmup updates	4000
# max updates	600K
# GPUs	8
length penalty	0.6

Table 4: Transformer-base `fairseq` hyperparameters and setting.

We use all 16 submissions for the WMT20 ZH-EN task (Barrault et al., 2020)[11] as well as our own three transformer baselines that were implemented in `fairseq` (Ott et al., 2019). Our baselines allow researchers to compare their translation models without resource-intensive techniques such as backtranslation (Sennrich et al., 2016a), model ensembling, and deep encoders (Kasai et al., 2021a). Tables 4 and 5 list the hyperprameters. We generally follow the setting from Vaswani et al. (2017).

[11] `https://www.statmt.org/wmt20/results.html`.

Hyperparameter	Value
label smoothing	0.1
# max tokens	4096
dropout rate	0.1
encoder embedding dim	1024
encoder ffn dim	4096
# encoder attn heads	16
decoder embedding dim	1024
decoder ffn dim	4096
# decoder attn heads	16
max source positions	1024
max target positions	1024
Adam lrate	5×10^{-4}
Adam β_1	0.9
Adam β_2	0.98
lr-scheduler	inverse square
warm-up lr	1×10^{-7}
# warmup updates	4000
# max updates	600K
# GPUs	8
length penalty	0.6

Table 5: Transformer-large and transformer-large-ensemble `fairseq` hyperparameters and setting. Transformer-large-ensemble ensembles four transformer-large models with different random initializations.

We use `newstest-2019` as the dev. set and the official training data.[12] We apply Moses tokenization (Koehn et al., 2007) and BPE with 32K operations (Sennrich et al., 2016b) to English text. We tokenize Chinese text with the Jieba package,[13] following Hassan et al. (2018). Separately from English, BPE with 32K operations is then applied to Chinese. The decoder input and output embeddings are tied. Moses detokenization is applied to get the final outputs in the last step. We make the three models and preprocessed train/dev. data publicly available.[14] Table 6 lists all generators and their automatic evaluation scores from the top-performing metric (ensemble in this case).

B.2 WMT20 EN-DE

Similar to WMT20 ZH-EN, we use all 14 submissions for the WMT20 EN-DE task along with our three transformer baselines. The same hyperparameters are chosen as in WMT20 ZH-EN (Tables 4 and 5). We preprocess both English and German text by the Moses tokenizer and *joint* BPE with 32K operations. All embeddings are shared. We apply the Moses detokenizer to get the final outputs.

[12] `http://www.statmt.org/wmt20/translation-task.html`.

[13] `https://github.com/fxsjy/jieba`.

[14] `https://github.com/jungokasai/billboard/tree/master/baselines`.

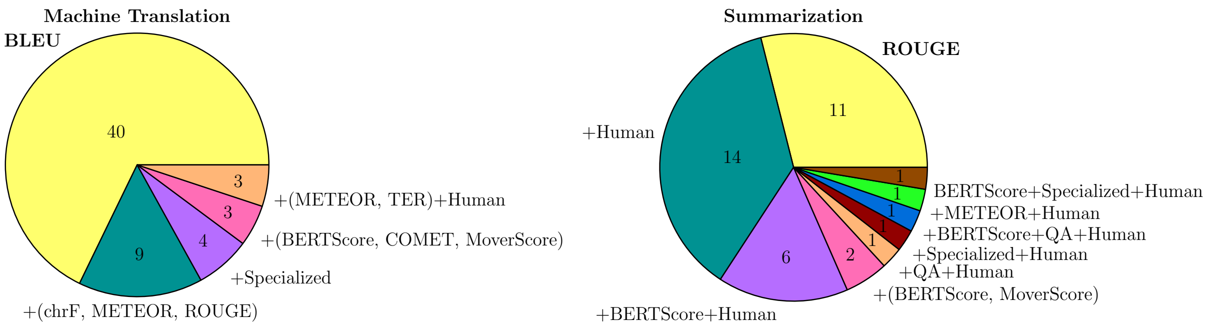

Figure 4: Breakdowns of evaluation metrics used in the papers on machine translation and summarization from NAACL and ACL 2021. We examined all papers whose title contains "machine translation" and "summarization" and disregarded papers primarily on evaluation metrics. "QA" metrics use a QA system to evaluate summaries (e.g., Eyal et al., 2019). "Specialized" indicates specialized evaluation in a particular dimension, rather than the overall generation quality, such as document-level evaluations on contrastive sets (Voita et al., 2019).

Generator	**Description**	**Score**
Huoshan Translate	Wu et al. (2020a)	78.85
THUNLP	Not available	78.81
Huawei TSC	Wei et al. (2020)	78.79
DeepMind	Yu et al. (2020)	78.76
WeChat AI	Meng et al. (2020)	78.75
Tencent Translation	Wu et al. (2020b)	78.74
DiDi NLP	Chen et al. (2020)	78.66
OPPO	Shi et al. (2020)	78.59
Online-B	Not available	78.36
SJTU-NICT	Li et al. (2020)	78.27
trans-large-ensemble	§B.1	77.35
trans-large	§B.1	76.98
Online-A	Not available	76.86
trans-base	§B.1	76.79
dong-nmt	Not available	76.74
Online-G	Not available	76.44
zlabs-nlp	Not available	75.79
Online-Z	Not available	75.05
WMT Biomed Baseline	Bawden et al. (2020)	73.89

Table 6: WMT20 ZH-EN generators and reference papers. The score column indicates the score from the metric that currently correlates best with the human judgments (**ensemble**).

Table 7 shows the generators and their automatic evaluation scores from the top-performing metric (ensemble).

B.3 CNNDM Summarization

We submit all 26 models from Fabbri et al. (2021).[15] Table 8 shows all models and their automatic evaluation scores from the top-performing metric (COMET).

[15] https://github.com/Yale-LILY/SummEval.

Generator	**Description**	**Score**
Tohoku-AIP-NTT	Kiyono et al. (2020)	90.50
Tencent Translate	Wu et al. (2020b)	90.43
OPPO	Shi et al. (2020)	90.42
eTranslation	Oravecz et al. (2020)	90.39
Online-B	Not available	90.38
Huoshan Translate	Wu et al. (2020a)	90.32
AFRL	Gwinnup and Anderson (2020)	90.16
Online-A	Not available	90.12
UEDIN	Germann (2020)	89.98
PROMT NMT	Molchanov (2020)	89.66
trans-large	§B.2	89.60
trans-large-ensemble	§B.2	89.59
trans-base	§B.2	89.35
Online-Z	Not available	89.26
Online-G	Not available	88.98
zlabs-nlp	Not available	88.65
WMT Biomed Baseline	Bawden et al. (2020)	88.23

Table 7: WMT20 EN-DE generators and reference papers. The score column indicates the score from the metric that currently correlates best with the human judgments (**ensemble**).

B.4 MSCOCO Image Captioning

We submit the four strong models from the literature (Kasai et al., 2022).[16] They share similar pipeline structure but vary in model architecture, (pre)training data, model size, and (pre)training objective. Table 9 shows the models with their papers and automatic scores from the top-performing metric (**RefCLIP-S**).

[16] https://github.com/jungokasai/THumB/tree/master/mscoco.
[17] Model with CIDEr optmization, https://github.com/microsoft/Oscar/blob/master/VinVL_MODEL_ZOO.md#Image-Captioning-on-COCO.
[18] Model with CIDEr optmization.
[19] Model with cross-entropy optimization, https://vision-explorer.allenai.org/image_captioning.

Generator	Description	Score
Lead-3	First 3 sentences	-0.011
T5	Raffel et al. (2020)	-0.030
BART	Lewis et al. (2020)	-0.032
Pegasus-dynamic-mix	Zhang et al. (2020a)	-0.044
RNES	Wu and Hu (2018)	-0.049
Unified-ext-abs	Hsu et al. (2018)	-0.056
Pegasus-huge-news	Zhang et al. (2020a)	-0.056
REFRESH	Narayan et al. (2018)	-0.067
ROUGESal	Pasunuru and Bansal (2018)	-0.073
Human-H	Highlights	-0.075
NEUSUM	Zhou et al. (2018)	-0.083
BanditSum	Dong et al. (2018)	-0.083
LATENT	Zhang et al. (2018)	-0.099
Closed-book-decoder	Jiang and Bansal (2018)	-0.112
Multi-task-Ent-QG	Guo et al. (2018)	-0.117
Pointer-Generator	See et al. (2017)	-0.144
UniLM	Dong et al. (2019)	-0.151
Bottom-Up	Gehrmann et al. (2018)	-0.160
JEC	Xu and Durrett (2019)	-0.167
Fast-abs-rl	Chen and Bansal (2018)	-0.189
NeuralTD	Böhm et al. (2019)	-0.215
Improve-abs	Kryściński et al. (2018)	-0.329
BertSum-abs	Liu and Lapata (2019)	-0.341
STRASS	Bouscarrat et al. (2019)	-0.405
GPT-2-zero-shot	Ziegler et al. (2019)	-0.441
SENECA	Sharma et al. (2019)	-0.735

Table 8: CNNDM summarization generators and reference papers. They are from Fabbri et al. (2021), but we apply detokenization (Bird et al., 2009) and/or truecasing (Manning et al., 2014) to standardize the model outputs for better, reproducible evaluations. The score column indicates the score from the metric that currently correlates best with the human judgments (**COMET**).

Generator	Description	Score
VinVL-large[17]	Zhang et al. (2021)	83.78
VinVL-base[18]	Zhang et al. (2021)	83.45
Unified-VLP	Zhou et al. (2020)	82.59
Up-Down[19]	Anderson et al. (2018)	80.63

Table 9: MSCOCO image captioning generators and reference papers. The score column indicates the score from the metric that currently correlates best with the human judgments (**RefCLIP-S**).

C Participating Metrics

Table 10 discusses details and configurations of the automatic metrics that we implement in our initial BILLBOARDS.

Metric	Description	Refs.	Src.	Cont.
BLEU[20]	Papineni et al. (2002)	✓	✗	✗
ROUGE-3[21]	Lin (2004)	✓	✗	✗
ROUGE-L	Lin (2004)	✓	✗	✗
METEOR	Banerjee and Lavie (2005)	✓	✗	✗
TER[22]	Snover et al. (2006)	✓	✗	✗
METEOR[23]	Banerjee and Lavie (2005)	✓	✗	✗
chrF[24]	Popović (2015)	✓	✗	✗
CIDEr[25]	Vedantam et al. (2015)	✓	✗	✗
SPICE	Anderson et al. (2016)	✓	✗	✗
CharacTER[26]	Wang et al. (2016)	✓	✗	✗
chrF++	Popović (2017)	✓	✗	✗
SummaQA[27]	Scialom et al. (2019)	✗	✓	✓
BERTScore	Zhang et al. (2020b)	✓	✗	✓
BLEURT[28]	Sellam et al. (2020)	✓	✗	✓
COMET[29]	Rei et al. (2020)	✓	✓	✓
COMET-QE	Rei et al. (2020)	✗	✓	✓
Prism-ref[30]	Thompson and Post (2020)	✓	✗	✓
Prism-src	Thompson and Post (2020)	✗	✓	✓
CLIP-S[31]	Hessel et al. (2021)	✗	✓	✓
RefCLIP-S	Hessel et al. (2021)	✓	✓	✓
RefOnlyC	Kasai et al. (2022)	✓	✗	✓

Table 10: Automatic metrics and their reference papers. The refs., src., and cont. columns indicate whether they use references, input source features, and pretrained contextual representations (e.g., BERT; Devlin et al., 2019), respectively.

D Additional Ensemble Metric Ablations

Seen in Table 11 are ablation studies for the ensemble metrics where one of the three selected metrics is removed at a time. Dropping one metric often has no impact on the correlation score, suggesting that these metrics are highly redundant and capture similar aspects of the output quality. BILLBOARDS encourage researchers to explore ways to diversify automatic evaluations by updating the ensemble metric every time a new metric is submitted.

[20] SACREBLEU implementation of sentence-level BLEU-4; https://github.com/mjpost/sacreBLEU/blob/v1.2.12/sacrebleu.py#L999.
[21] https://pypi.org/project/rouge-score/.
[22] https://github.com/mjpost/sacrebleu.
[23] https://www.nltk.org/_modules/nltk/translate/meteor_score.html.
[24] https://github.com/m-popovic/chrF.
[25] https://github.com/salaniz/pycocoevalcap.
[26] https://github.com/rwth-i6/CharacTER.
[27] https://github.com/ThomasScialom/summa-qa.
[28] https://huggingface.co/metrics/bleurt.
[29] https://github.com/Unbabel/COMET/.
[30] https://github.com/thompsonb/prism.
[31] https://github.com/salaniz/pycocoevalcap.

ZH-EN	–	COMET	COMET-QE	BLEURT
	0.61	0.61	0.57	0.61
EN-DE	–	COMET	COMET-QE	Prism-ref
	0.51	0.52	0.52	0.52
CNNDM	–	COMET	COMET-QE	BERTScore
	0.29	0.23	0.31	0.31
COCO	–	RefCLIP-S	RefOnlyC	CIDEr
	0.45	0.44	0.42	0.43

Table 11: Correlations from ensemble ablation studies. One of the three selected metrics is removed at a time, and a new Lasso regression model is trained on the remaining metrics. The bigger the correlation drop is, the bigger the contribution is from the removed metric. COMET-QE is a referenceless metric.

E Additional Mixed-Effects Analysis

Table 12 presents fixed-effect coefficients that measure how much each automatic metric *overrates* machines over humans (§2.3). With some exceptions in CNNDM summarization, almost all automatic metrics *underrate* human generations (significantly positive coefficients). Table 13 swaps the roles of human-generated text, but we still see similar patterns: almost all metrics overrate machines over humans, but the problem is mitigated in COMET-QE, a referenceless, quality estimation metric. This confirms that our findings hold independently of the design choice.

F Crowdworker vs. Rubric-based Expert Evaluations

Seen in Table 14 are examples where crowdworker evaluators (Barrault et al., 2020) and professional translators (Freitag et al., 2021) disagree: crowdworkers give lower scores to the human-generated translations than the machine-generated ones. The first case requires document-level context to properly evaluate. Document-level context and diversity in high-quality human translations can mislead crowdworkers.

ZH-EN	COMET-QE	Ensemble	COMET	BLEURT	BERTScore	CharacTER	MoverScore	METEOR
	$0.13_{\pm 0.01}$	$0.26_{\pm 0.01}$	$0.27_{\pm 0.02}$	$0.32_{\pm 0.02}$	$0.52_{\pm 0.02}$	$0.56_{\pm 0.02}$	$0.57_{\pm 0.02}$	$0.57_{\pm 0.02}$
	Prism-ref	chrF	TER	chrF++	ROUGE-3	BLEU	ROUGE-L	Prism-src
	$0.58_{\pm 0.02}$	$0.58_{\pm 0.02}$	$0.59_{\pm 0.02}$	$0.60_{\pm 0.02}$	$0.61_{\pm 0.02}$	$0.62_{\pm 0.02}$	$0.64_{\pm 0.02}$	$1.13_{\pm 0.02}$
EN-DE	COMET-QE	Ensemble	COMET	MoverScore	chrF	chrF++	BLEU	CharacTER
	$-0.17_{\pm 0.02}$	$0.03_{\pm 0.02}$	$0.08_{\pm 0.02}$	$0.22_{\pm 0.03}$	$0.29_{\pm 0.02}$	$0.32_{\pm 0.02}$	$0.33_{\pm 0.03}$	$0.33_{\pm 0.03}$
	BERTScore	Prism-ref	TER	Prism-src				
	$0.43_{\pm 0.02}$	$0.44_{\pm 0.02}$	$0.49_{\pm 0.03}$	$1.46_{\pm 0.03}$				
CNNDM	TER	COMET	Ensemble	BERTScore	MoverScore	COMET-QE	CharacTER	BLEURT
	$-0.58_{\pm 0.14}$	$-0.17_{\pm 0.12}$	$-0.16_{\pm 0.12}$	$-0.04_{\pm 0.12}$	$-0.03_{\pm 0.11}$	$0.02_{\pm 0.11}$	$0.14_{\pm 0.15}$	$0.25_{\pm 0.12}$
	SummaQA	ROUGE-L	BLEU	Prism-ref	chrF	chrF++	ROUGE-3	METEOR
	$0.27_{\pm 0.10}$	$0.33_{\pm 0.13}$	$0.37_{\pm 0.11}$	$0.38_{\pm 0.12}$	$0.43_{\pm 0.13}$	$0.45_{\pm 0.13}$	$0.49_{\pm 0.11}$	$0.53_{\pm 0.12}$
COCO	CLIP-S	RefCLIP-S	CharacTER	chrF	ROUGE-3	chrF++	RefOnlyC	Ensemble
	$-0.04_{\pm 0.05}$	$0.09_{\pm 0.06}$	$0.13_{\pm 0.07}$	$0.18_{\pm 0.07}$	$0.22_{\pm 0.07}$	$0.23_{\pm 0.07}$	$0.24_{\pm 0.06}$	$0.24_{\pm 0.06}$
	SPICE	METEOR	BLEU	CIDEr	ROUGE-L	BERTScore	TER	MoverScore
	$0.25_{\pm 0.07}$	$0.32_{\pm 0.07}$	$0.39_{\pm 0.07}$	$0.43_{\pm 0.06}$	$0.44_{\pm 0.07}$	$0.45_{\pm 0.06}$	$0.45_{\pm 0.07}$	$0.51_{\pm 0.05}$

Table 12: Fixed-effect coefficients β_0 from the linear mixed-effects analysis that measures how much automatic metrics **overrate** machine text over human, as compared to human raters (§2.3). $\beta_0 = 0$ is neutral, and statistical significance is indicated by red (positive) or blue text (negative). The subscripts indicate 90% confidence intervals. COMET-QE, Prism-src, SummaQA and CLIP-S are referenceless metrics. In both WMT20 ZH-EN and WMT20 EN-DE, Human-B is evaluated as human-generated translations. Human-A (WMT20 ZH-EN) and Human-A and Human-P (WMT20 EN-DE) are used as the reference set for reference-based metrics.

ZH-EN	COMET-QE	Ensemble	COMET	BLEURT	TER	BERTScore	ROUGE-3	Prism-ref
	$0.03_{\pm 0.01}$	$0.07_{\pm 0.01}$	$0.08_{\pm 0.02}$	$0.09_{\pm 0.02}$	$0.23_{\pm 0.02}$	$0.24_{\pm 0.02}$	$0.24_{\pm 0.02}$	$0.25_{\pm 0.02}$
	CharacTER	ROUGE-L	chrF	MoverScore	METEOR	chrFpp	BLEU	Prism-src
	$0.25_{\pm 0.02}$	$0.26_{\pm 0.02}$	$0.27_{\pm 0.02}$	$0.27_{\pm 0.02}$	$0.29_{\pm 0.02}$	$0.29_{\pm 0.02}$	$0.30_{\pm 0.02}$	$0.79_{\pm 0.02}$
EN-DE	COMET-QE	Ensemble	COMET	MoverScore	Prism-ref	chrF	BERTScore	CharacTER
	$-0.09_{\pm 0.02}$	$-0.07_{\pm 0.02}$	$-0.06_{\pm 0.03}$	$0.02_{\pm 0.02}$	$0.18_{\pm 0.02}$	$0.20_{\pm 0.02}$	$0.21_{\pm 0.02}$	$0.22_{\pm 0.02}$
	chrF++	BLEU	TER	Prism-src				
	$0.22_{\pm 0.02}$	$0.23_{\pm 0.02}$	$0.32_{\pm 0.02}$	$1.38_{\pm 0.03}$				

Table 13: Fixed-effect coefficients β_0 from the linear mixed-effects analysis that measures how much automatic metrics **overrate** machine text over human, as compared to human raters (§2.3). **The roles of human translations are swapped**: Human-A is evaluated, and Human-B (WMT20 ZH-EN) and Human-B and Human-P (WMT20 EN-DE) are used as the reference. We still see similar patterns to Table 12: almost all automatic metrics overrate machines over humans, but the problem is less severe in the referenceless metric of COMET-QE.

	WMT20 ZH-EN	
Source	希望兴安省继续为白俄罗斯企业提供便利条件。	凭的是相机而动的时势驾驭。
Huoshan	It is hoped that **Xing'an Province** will continue to provide convenient conditions for Belarusian enterprises.	It is based on the current situation of the **camera**.
Human-A	He hoped that **Hung Yen Province** would continue to provide convenient conditions for Belarusian enterprises.	This relies on the ability to seize opportunities.
Human-B	He hoped that **this** could continue in the future.	It is based on the observation of various situations at different times.

Table 14: Examples where crowdsource evaluators (Barrault et al., 2020) and professional translators (Freitag et al., 2021) disagree: crowdworkers give lower scores to the human-generated translations than the machine-generated ones. The first case requires document-level context to properly evaluate. 兴安省 is Hung Yen Province in Vietnam in this context, but there is entity ambiguity. (Xing'an Province that existed in the Republic of China.) The second one illustrates the diversity of human generations that misleads crowdworkers.