Results of WMT22 Metrics Shared Task: Stop Using BLEU – Neural Metrics Are Better and More Robust

Markus Freitag(1), **Ricardo Rei**(2,3,4), **Nitika Mathur**(5), **Chi-kiu Lo**(6), **Craig Stewart**(2), **Eleftherios Avramidis**(8), **Tom Kocmi**(7), **George Foster**(1), **Alon Lavie**(2) **and André F. T. Martins**(2,3,9)
(1)Google Research (2)Unbabel (3)INESC-ID (4)Instituto Superior Técnico
(5)Oracle Digital Assistant (6)National Research Council Canada (7)Microsoft
(8)German Research Center for Artificial Intelligence (DFKI) (9)Instituto de Telecomunicações
wmt22-metric@googlegroups.com

Abstract

This paper presents the results of the WMT22 Metrics Shared Task. Participants submitting automatic MT evaluation metrics were asked to score the outputs of the translation systems competing in the WMT22 News Translation Task on four different domains: news, social, e-commerce, and chat. All metrics were evaluated on how well they correlate with human ratings at the system and segment level. Similar to last year, we acquired our own human ratings based on expert-based human evaluation via Multidimensional Quality Metrics (MQM). This setup had several advantages, among other things: (i) expert-based evaluation is more reliable, (ii) we extended the pool of translations by 5 additional translations based on MBR decoding or rescoring which are challenging for current metrics.

In addition, we initiated a challenge set subtask, where participants had to create contrastive test suites for evaluating metrics' ability to capture and penalise specific types of translation errors.

Finally, we present an extensive analysis on how well metrics perform on three language pairs: English→German, English→Russian and Chinese→English. The results demonstrate the superiority of neural-based learned metrics and demonstrate again that overlap metrics like BLEU, SPBLEU or CHRF correlate poorly with human ratings. The results also reveal that neural-based metrics are significant better than non-neural metrics across different domains and challenges.

1 Introduction

The metrics shared task[1] has been a key component of WMT since 2008, serving as a way to validate the use of automatic MT evaluation metrics and drive the development of new metrics. We evaluate reference-based automatic metrics that score MT output by comparing the translations with a reference translation generated by human translators, who are instructed to translate "from scratch" without post-editing from MT. In addition, we also invited submissions of reference-free metrics (quality estimation metrics or QE metrics) that compare MT outputs directly with the source segments. All metrics are evaluated based on their agreement with human rating when scoring MT systems and human translations at the system or sentence level. The final ranking of this year's submitted primary metrics is shown in Table 1. We provide details in the remainder of the paper.

Metric	avg rank
METRICX XXL	1.20
COMET-22	1.32
UNITE	1.86
BLEURT-20	1.91
COMET-20	2.36
MATESE	2.57
COMETKIWI*	2.70
MS-COMET-22	2.84
UNITE-SRC*	3.03
YISI-1	3.27
COMET-QE*	3.33
MATESE-QE*	3.85
MEE4	3.87
BERTSCORE	3.88
MS-COMET-QE-22*	4.06
CHRF	4.70
F101SPBLEU	4.97
HWTSC-TEACHER-SIM*	5.17
BLEU	5.31
REUSE*	6.69

Table 1: Official ranking of all primary submissions of the WMT22 Metric Task. The final score is the weighted average ranking over 201 different scenarios. Metrics with * are reference-free metrics.

We implemented several changes to the methodology that was followed in previous years' editions:

- **Expert-based human evaluation**: Like last year, we collected our own human ratings for select language pairs (en→de, en→ru, zh→en) from professional translators via MQM (Lommel et al.,

[1] https://wmt-metrics-task.github.io/

Proceedings of the Seventh Conference on Machine Translation (WMT), pages 46–68
Abu Dhabi, December 7–8, 2022. ©2022 Association for Computational Linguistics

2014). Freitag et al. (2021a) showed that expert-based MQM evaluations produce more reliable[2] scores when compared to the DA-based human ratings acquired by the WMT Translation task. This step was necessary as Freitag et al. (2021a) showed that the DA-based ground-truth is already of lower quality than some of our submissions (Section 3).

- **Additional Training Data**: We encouraged the participants to make use of existing MQM annotations for newstest2020 (Freitag et al., 2021a)[3], and the MQM annotations from the WMT21 Metrics Task (Freitag et al., 2021b) to improve and/or test their metrics.

- **Additional MT systems**: The primary use case for automatic metrics is guiding research to translations that are better than what we can generate right now. To address this scenario, we not only want to evaluate metrics on MT output that we are currently capable of generating, but also on translations that are better than the current WMT submissions. For that we need to add alternative translations that cover a wider space of possible translations. To address this, we added MT systems that were generated with MBR decoding or reranking (Section 2.2).

- **Challenge sets subtask**: In the main metrics task, the metrics are evaluated on MT systems translating test sets drawn from large sources of continuous text. In an effort to have a more fine-grained analysis on the strengths and weaknesses of the metrics, we introduced the concept of challenge sets. A challenge set consists of contrasting MT outputs, which have been deliberately devised or selected to include correct and incorrect translations of particular phenomena, along with their respective reference translation. The evaluation of every metric in this setup depends on its ability to rank the correct translations higher than their corresponding incorrect ones. Whereas a first version of challenge sets appeared in last year's metrics shared task (Freitag et al., 2021b), this year they appear for the first time as a subtask in a decentralized manner. Inspired by the *Build it or break it: The Language Edition* shared task (Ettinger et al., 2017), participants (the *Breakers*) had to submit their own test suites to test the robustness of MT metrics to particular phenomena that they choose. Our first edition of this subtask (Section 8) received four challenge set submissions covering a wide range of phenomena and languages.

- **Meta Evaluation**: A main aim of the metrics task is to rank the overall performance of various metrics. This requires some way of aggregating scores across different settings (language pair, domain, granularity etc.), in order to provide a balanced picture. Correlations with human scores have different ranges in different settings, so averaging them is not a good solution. Last year, we adopted a proposal by Kocmi et al. (2021) that involves taking the microaverage of a metric's accuracy in making pairwise system-ranking decisions across different settings. This is easy to interpret and reflects a common use-case for metrics, but because we have only three language pairs, and thus relatively few pairwise comparisons, it tends to place many metrics into large significance clusters (eg, 8 metrics in the top cluster last year, including CHRF but excluding COMET). In an effort to better discriminate, and to represent a broader set of use-cases, this year we computed the average rank of each metric across a large set of tasks (Section 5). This statistic has a clear interpretation, is justified by social choice theory (Colombo et al., 2022), and makes it easy to zoom into different subsets of tasks to provide finer-grained characterizations. To reflect the importance of the accuracy metric from last year, we define it as a single highly-important task (out of 201 tasks in total), with an overall weight of 25%.

- **MTME**: Similar to last year, all results in this paper are calculated with MTME[4]. We want to encourage every metric developer to use this tool to calculate scores for consistency and comparability going forward.

Our main findings are:

- Out of 13 reference-based metrics **BLEU is ranked last**, followed by F200SPBLEU and CHRF.

[2] DA is unreliable for high-quality MT output; ranks human translations lower than MT; correlates poorly with metrics. Expert-based MQM ranks human translations higher than MT and correlates generally much better with automatic metrics.

[3] https://github.com/google/wmt-mqm-human-evaluation

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

- **Neural fine-tuned metrics are not only better, but also robust to different domains**. Furthermore, based on the results from the four submitted challenge sets, neural fine-tuned metrics exhibit superior performance when compared to lexical and embedding similarity metrics.

- Top performing metrics from previous years are still top-performers, being only outperformed by model ensembles or metrics based on considerably larger neural models.

- For the first time since 2008, there was no new purely lexical metric submission, which indicates that metric developers are moving away from lexical metrics.

The rest of the paper is organized as follows: Section 2 describes the additional MT systems. Section 3 presents an overview of the conducted expert-based human evaluation. Section 4 describes the metrics evaluated this year (baselines and participants). Section 5 describes the conducted meta-evaluation. Section 6 reports our main results. Section 7 summarizes our results for additional WMT22 Translation task language-pairs based on their Direct Assessment human evaluation. Section 8 presents a description of the submitted challenge sets along with their findings. Finally, Section 9 presents our most relevant conclusions.

2 Translation Systems

Similar to the previous years' editions, the source, reference texts, and MT system outputs for the metrics task are mainly derived from the WMT22 general MT Task. In addition to the MT system outputs from the WMT evaluation campaign, we added translations from six additional MT systems which we deemed interesting for evaluation.

2.1 WMT Test Sets

The general MT 2022 test set contains around 2000 segments for each translation direction. This year, the test sets cover 4 domains: news, social, conversational, and e-commerce. There are around 500 sentences for each domain resulting in reasonably balanced test sets. English sources are identical for both into-German and into-Chinese translation directions. The reference translations provided for the test sets are translated by professional translators. We have two reference translations for English→German and Chinese→English sponsored by Microsoft and one reference translation for English→Russian sponsored by Google. For more details regarding the news test sets, we refer the reader to the WMT22 General MT task findings paper (Kocmi et al., 2022a).

2.2 Additional MT Output

Similar to last year, we want to expand the pool of translations beyond the WMT submissions, which usually are quite similar to each other. We added translations based on M2M100 and translations generated with MBR decoding.

M2M100 1.2B As the field moves forward to large multilingual pre-trained models, we are interested in comparing such general-purpose large multilingual MT systems against direct submissions to the general MT task. Models such as MBART50 (Tang et al., 2021) and M2M100 (Fan et al., 2021) are publicly available, easy to use and have recently been used as baselines and/or as a backbone for new research. We tested both models on the newstest2021 and we decided to include M2M100 1.2B as an additional MT output as it yielded better automatic scores.

MBR Outputs Minimum Bayes Risk (MBR) decoding has recently gained attention in MT as a decision rule, with the potential to overcome some of the biases of MAP decoding in NMT (Eikema and Aziz, 2020; Müller and Sennrich, 2021; Eikema and Aziz, 2021; Freitag et al., 2022; Fernandes et al., 2022). MBR decoding centrally relies on a reference-based utility metric: its goal is to identify a hypothesis with a high estimated utility (expectation under model distribution) with the hope that a high estimated utility translates into a high actual utility (with respect to a human reference). MBR decoding is particularly interesting for reference-based metrics as it stress tests the metric, using it as a utility function.

This year, we added three different MBR runs using three different utility functions (BLEU, BLEURT-20, and COMET-20) as additional translations. Freitag et al. (2022) demonstrated that the translations generated with a neural-based utility (BLEURT-20, and COMET-20) generate translations that are not only better when compared to MAP decoding, but the resulting translations are also significantly different from both the beam search decoding and the MBR decoding output using BLEU as a utility function. To make it even more interesting for the metric task, for these MBR translation models we used a transformer-big baseline trained only on WMT22 bilingual training data. By not using the strongest NMT system, we hope to see interesting new errors in the translation output. To generate the candidate list for MBR decoding, we sampled 256 times from the model using unbiased ancestral sampling.

Reranking Outputs Complementary to MBR outputs, we were also interested in comparing and evaluating the quality produced by reranking approaches based on QE. Our hope is that QE based reranking would lead to translations that are lexically different than traditional beam search output and thus lead to more diverse translations for the same source sentences. For English→German and English→Russian we used the Fairseq WMT19 systems[5] (Ng et al., 2019) with Nucleus Sampling (Holtzman et al., 2019) to generate 200 candidate translations, from which we choose the best translation according to the Tune Reranker proposed in Fernandes et al. (2022). For Chinese→English we used the same process but replacing the NMT model with MBART50 (many-to-one) and using only 50 samples.

3 MQM Human Evaluation

Automatic metrics are usually evaluated by measuring correlations with human ratings. The quality of the underlying human ratings is critical and recent findings (Freitag et al., 2021a) have shown that crowd-sourced human ratings are not reliable for high quality MT output. Furthermore, an evaluation schema based on MQM (Lommel et al., 2014), which requires explicit error annotation, is preferable to an evaluation schema that only asks raters for a single scalar value per translation. Similar to last year, we decided to not use the human ratings from the WMT General MT task, and conducted our own MQM-based human evaluation on a subset of submissions and a subset of language pairs that are most interesting for evaluating current metrics. This not only had the advantage of more reliable ratings for a subset of language pairs, but also gave us the opportunity to add our own translations that might be challenging for current metrics and are not part of an WMT submission.

MQM is a general framework that provides a hierarchy of translation errors which can be tailored to specific applications. Google and Unbabel sponsored the human evaluation for this year's metrics task for a subset of language pairs using either professional translators (English→German, Chinese→English) or trusted and trained raters (English→Russian). The error annotation typology and guidelines used by Google's and Unbabel's annotators differ slightly and are described in the following two sections.

3.1 English→German and Chinese→English

Annotations for English→German and Chinese→English were sponsored and executed by Google, using 11 professional translators (7 for English→German, 4 for Chinese→English) having access to the full document context. Each segment gets annotated by a single rater. Instead of assigning a scalar value to each translation, annotators were instructed to label error spans within each segment in a document, paying particular attention to document context. Each error was highlighted in the text, and labeled with an error category and a severity. To temper the effect of long segments, we imposed a maximum of five errors per segment, instructing raters to choose the five most severe errors for segments containing more errors. Segments that are too badly garbled to permit reliable identification of individual errors are assigned a special *Non-translation* error. Error severities are assigned independent of category, and consist of *Major*, *Minor*, and *Neutral* levels, corresponding respectively to actual translation or grammatical errors, smaller imperfections and purely subjective opinions about the translation. Since we are ultimately interested in scoring segments, we adopt the weighting scheme shown in Table 2, in which segment-level scores can range from 0 (perfect) to 25 (worst). The final segment-level score is an average over scores from all annotators. For more details, exact annotator instructions and a list of error categories, we refer the reader to Freitag et al. (2021a) as the exact same setup was used for the WMT21 metrics task.

Severity	Category	Weight
Major	Non-translation all others	25 5
Minor	Fluency/Punctuation all others	0.1 1
Neutral	all	0

Table 2: Google's MQM error weighting.

[5] https://github.com/facebookresearch/fairseq/tree/main/examples/wmt19

3.2 English→Russian

The annotations for English→Russian were provided by Unbabel who utilized four professional, native language annotators with ample translation experience. Annotation was conducted using Unbabel's own proprietary variant of the MQM framework (Lommel et al., 2014) which is fully compliant with MQM 2.0, being the most recent iteration of the framework[6]. Annotation was split along the four domain boundaries with each of the annotators evaluating all of the systems for a single content type. Similarly to Google, the annotators were given the full document context (up to ten segments) and were instructed to identify (by highlighting) and classify errors in accordance with the MQM typology. Annotators were also asked to classify error severity; in addition to *Minor* and *Major* error severities used by Google, Unbabel also uses a *Critical* error severity. However, in the interest of maintaining consistency in evaluation, we calculated the MQM score in a manner compliant with the Google methodology outlined above. Specifically all annotated *Critical* errors were counted as *Major* and punctuation errors were weighted using the weighting scheme in Table 2.

3.3 Human Evaluation Results

As discussed in Section 1, we decided to run our own human evaluation in order to generate our golden-truth ratings and come to stronger conclusions about the quality of each automatic metric across all domains. However, this also meant that we were only able to evaluate a subset of the test sets. In Table 3, you can see the number of segments for each language pair and test set that we used for human evaluation. We followed a simple and consistent approach to downsample the data: we kept the first 10 sentences of each document. By doing this, we did not need to discard any documents and only needed to crop longer documents. An exception is Chinese→English where we evaluated the full test set.

language	news	social	ecomm.	conv.
en→de	300/511	340/512	230/530	445/484
en→ru	300/511	340/512	230/530	445/484
zh→en	505/505	503/503	518/518	349/349

Table 3: Numbers of MQM-annotated segments per domain.

The results of the MQM human evaluation can be seen in Table 4. Most of the reference translations are ranked first, except for refB for English→German. Not ranking the human evaluation on top of the MT output is usually a signal for a corrupt human evaluation. We double checked the annotation for refB and can confirm that the reference translation indeed contained some errors.

4 Baselines and Primary Submissions

We computed scores for several baseline metrics in order to compare submissions against previous well-studied metrics. We will start by describing those baselines and then we will describe the submissions from participating teams. An overview of the evaluated metrics can be seen in Table 5.

4.1 Baselines

SacreBLEU baselines We use the following metrics from the SacreBLEU (Post, 2018) as baselines:

- **BLEU (Papineni et al., 2002)** is based on the precision of n-grams between the MT output and its reference weighted by a brevity penalty. Using SacreBLEU we obtained sentence-BLEU values using the `sentence_bleu` Python function and for corpus-level BLEU we used `corpus_bleu` (both with default arguments[7]).

- **F101SPBLEU (Goyal et al., 2022)** and **F200SPBLEU (NLLB Team et al., 2022)** are BLEU scores computed with subword tokenization done by standardized Sentencepiece Models (Kudo and Richardson, 2018). We used the command line SacreBLEU to compute the sentence level F101SPBLEU[8] and F200SPBLEU[9] and we average those scores to obtain a corpus-level score.

- **CHRF (Popović, 2015)** uses character n-grams instead of word n-grams to compare the MT output with the reference. For CHRF we used the SacreBLEU `sentence_chrf` function (with default arguments[10]) for segment-level scores and we average those scores to obtain a corpus-level score.

[6] https://themqm.org/

[7] lnrefs.1|case.mixed|lang.LANGPAIR|tok.13a|smooth.exp |version.1.5.0

[8] nrefs:1|case:mixed|eff:yes|tok:flores101|smooth:exp| version:2.3.1

[9] nrefs:1|case:mixed|eff:yes|tok:flores200|smooth:exp| version:2.3.1

[10] chrF2|lang.LANGPAIR|nchars.6|space.false|version.1.5.0

	English→German ↓				
System	all	news	social	ecom.	conv.
refA	0.64	0.97	0.68	0.56	0.42
Online-W	0.79	0.95	0.74	0.93	0.65
refB	0.91	1.38	0.93	1.17	0.46
MBR-bleu	0.96	1.29	1.14	0.82	0.67
Online-B	1.04	1.44	1.27	0.88	0.67
JDExploreAcademy	1.05	1.36	1.21	1.20	0.64
MBR-comet	1.08	1.40	1.33	1.01	0.71
MBR-bleurt	1.11	1.55	1.41	0.72	0.78
Online-A	1.21	1.40	1.55	1.35	0.76
Online-G	1.22	1.78	1.51	1.17	0.66
Online-Y	1.30	1.99	1.45	1.02	0.86
QUARTZ	1.34	1.85	1.59	1.10	0.94
Lan-Bridge	1.41	2.43	1.72	1.09	0.65
OpenNMT	1.68	1.98	2.14	1.73	1.09
PROMT	1.76	2.41	1.94	1.56	1.27
M2M100	2.82	3.46	2.99	2.94	2.19

	Chinese→English ↓				
System	all	news	social	ecom.	conv.
refA	1.22	1.42	1.10	1.42	0.82
refB	2.00	2.18	1.83	1.69	0.96
Lan-Bridge	2.47	2.45	1.97	3.55	1.39
MBR-bleurt	2.51	2.52	2.06	3.68	1.55
Online-B	2.71	2.66	2.07	3.73	1.55
LanguageX	2.74	2.74	2.46	3.78	1.58
JDExploreAcademy	2.83	2.84	2.56	3.81	1.60
MBR-comet	2.87	2.88	2.63	3.98	1.61
Online-G	2.93	2.90	2.73	4.16	1.63
MBR-bleu	3.00	2.94	2.77	4.22	1.64
HuaweiTSC	3.09	2.96	2.80	4.30	1.68
AISP-SJTU	3.19	3.08	2.89	5.03	1.76
Online-Y	3.28	3.27	3.03	5.20	1.79
Online-A	3.73	3.49	3.48	5.39	2.04
Online-W	3.95	3.96	3.60	5.76	2.30
M2M100	6.82	7.47	5.78	9.37	3.61

	English→Russian ↓				
System	all	news	social	ecom.	conv.
refA	1.13	0.43	2.17	1.95	0.39
Online-W	1.37	1.35	2.96	0.90	0.41
MBR-bleu	1.85	1.57	4.01	1.39	0.63
Online-B	1.94	1.59	4.29	1.37	0.68
Online-G	2.03	1.50	4.33	1.88	0.71
JDExploreAcademy	2.09	1.14	4.63	2.23	0.71
MBR-comet	2.10	2.01	4.74	1.26	0.57
Lan-Bridge	2.34	2.14	5.49	1.49	0.51
Online-Y	2.55	2.06	5.79	1.66	0.86
Online-A	2.85	1.83	6.56	2.62	0.83
PROMT	2.94	2.04	6.88	2.55	0.73
HuaweiTSC	3.40	1.72	8.07	3.02	1.17
SRPOL	3.68	2.02	8.19	3.53	1.43
eTranslation	3.79	2.30	8.54	3.49	1.32
QUARTZ	4.06	3.82	7.02	5.03	1.46
M2M100	4.56	3.74	9.27	4.42	1.58

Table 4: MQM human evaluations for generaltest2022. Lower average error counts represent higher MT quality.

BERTScore (Zhang et al., 2020) leverages contextual embeddings from pre-trained transformers to create soft-alignments between words in candidate and reference sentences using cosine similarity. Based on the alignment matrix, BERTScore returns a precision, recall and F1 score. We used F1 without TF-IDF weighting.

YiSi-1 (Lo, 2019) is a MT evaluation metric that measures the semantic similarity between a machine translation and human references by aggregating the IDF-weighted lexical semantic similarities based on the contextual embeddings extracted from pre-trained language models (e.g. RoBERTa, CamemBERT, XLM-RoBERTa, etc.).

Bleurt (Sellam et al., 2020) is a learned metric that is fine-tuned to produce a DA for a given translation by encoding it jointly with its reference. We used the Bleurt20 checkpoint (Pu et al., 2021) which was trained on top of RemBERT using DA from previous shared tasks ranging 2015 to 2019 and additional synthetic data created from Wikipedia articles.

Comet (Rei et al., 2020) is a learnt metric that is fine-tuned to produce a z-standardized DA for a given translation by comparing its representation to source and reference embeddings. We used the default model `wmt20-comet-da` provided in version 1.1.2 which is trained on top of XLM-R large using data from from previous shared tasks ranging 2017 to 2019.

Comet-QE (Rei et al., 2021) is a reference-free learnt metric similar to Comet. We used the `wmt21-comet-qe-mqm`) model which was a top-performing metric from last year's shared task. This metric is first trained on z-standardized DA from 2017 to 2020 and then fine-tuned on z-standardized MQM from (Freitag et al., 2021a).

	metric	broad category	sup-erv.	ref. free	citation	availability (https://github.com/)
baselines	BLEU	lexical overlap			Papineni et al. (2002)	mjpost/sacrebleu
	F101SPBLEU	lexical overlap			Goyal et al. (2022)	mjpost/sacrebleu
	F200SPBLEU	lexical overlap			NLLB Team et al. (2022)	mjpost/sacrebleu
	CHRF	lexical overlap			Popović (2015)	mjpost/sacrebleu
	BERTSCORE	embedding similarity			Zhang et al. (2020)	Tiiiger/bert_score
	BLEURT	fine-tuned metric	✓		Sellam et al. (2020)	google-research/bleurt
	COMET	fine-tuned metric	✓		Rei et al. (2020)	Unbabel/COMET
	COMET-QE	fine-tuned metric	✓	✓	Rei et al. (2021)	Unbabel/COMET
	YISI-1	embedding similarity			Lo (2019)	chikiulo/yisi
primary submissions	COMET-22	fine-tuned metric	✓		Rei et al. (2022)	Unbabel/COMET
	COMETKIWI	fine-tuned metric	✓	✓	Rei et al. (2022)	Unbabel/COMET
	EE-BERTSCORE	embedding similarity			Liu et al. (2022)	(not available)
	KG-BERTSCORE	embedding similarity		✓	Liu et al. (2022)	(not available)
	MATESE	fine-tuned metric	✓		Perrella et al. (2022)	(not available)
	MATESE-QE	fine-tuned metric	✓	✓	Perrella et al. (2022)	(not available)
	MEE4	lexical & embedding similarity			Mukherjee and Shrivastava (2022b)	AnanyaCoder/WMT22Submission
	METRICX XXL	fine-tuned metric	✓			(not available)
	MS-COMET	fine-tuned metric	✓		Kocmi et al. (2022b)	MicrosoftTranslator/MS-Comet
	MS-COMET-QE	fine-tuned metric	✓	✓	Kocmi et al. (2022b)	MicrosoftTranslator/MS-Comet
	REUSE	embedding similarity		✓	Mukherjee and Shrivastava (2022a)	AnanyaCoder/WMT22Submission_REUSE
	TEACHER-SIM	fine-tuned metric	✓	✓	Liu et al. (2022)	(not available)
	SESCORE	fine-tuned metric			Xu et al. (2022)	xu1998hz/SEScore
	UNITE	fine-tuned metric	✓		Wan et al. (2022b)	NLP2CT/UniTE

Table 5: Baseline metrics and primary submissions for the metrics task. We categorize metrics into 3 major classes: lexical, embedding similarity and fine-tuned metrics. Regarding fine-tuned metrics we have metrics that use human quality scores such as DA or MQM and metrics that use synthetic labels for fine-tuning (3rd column).

4.2 Metric Submissions

The rest of this section summarizes participating metrics. The * symbol indicates that the metric is the primary submission of the research group.

COMET-22* (Rei et al., 2022) is an ensemble of two models; 1) COMET estimator model trained with Direct Assessments and 2) a newly proposed multitask model trained to predict sentence-level MQM scores along with OK/BAD word-level tags derived from annotation spans.

COMETKIWI* ensembles 2 QE models similarly to COMET-22; 1) classic Predictor-Estimator QE model trained on DAs ranging 2017 to 2019 and then fine-tuned on DAs from MLQE-PE (the official DA from the QE shared task) and 2) the same multitask model used in the COMET-22 submission but without access to a reference translation.

MS-COMET-22* and MS-COMET-QE-22* (Kocmi et al., 2022b) are built on top of COMET by Microsoft Research using proprietary data. This metric is trained on a several times larger set of human judgements compared to COMET-baseline, covering 113 languages and 15 domains. Furthermore, the authors propose filtering of human judgement with potentially low quality to further improve the model.

MS-COMET-22 evaluated source, MT hypothesis and human reference from the input, while MS-COMET-QE-22 calculated scores in quality estimation fashion with only source segment and MT hypothesis.

EE-BERTSCORE* (Liu et al., 2022) stands for Entropy Enhanced BERTSCORE and aims at achieving a more balanced system-level rating by assigning weights to segment-level scores produced by BERTSCORE. The weights are determined by the difficulty of a segment determined by the entropy between the hypothesis-reference pair.

KG-BERTSCORE (Liu et al., 2022) is a reference-free machine translation (MT) evaluation metric, which incorporates multilingual knowledge graph into BERTScore by linearly combining the results of BERTScore and bilingual named entity matching.

CROSS-QE (Liu et al., 2022) is a reference-free metric with a similar architecture to COMET-QE.

HWTSC-Teacher-Sim★ (Liu et al., 2022) is a reference-free metric by fine-tuning the multilingual Sentence BERT model paraphrase-multilingual-mpnet-base-v2

HWTSC-TLM (Liu et al., 2022) is a reference-free metric which only uses a target-side language model to score the system translations as input.

Matese★ (Perrella et al., 2022) and Matese-QE★ leverage transformer-based multilingual encoders to identify error spans in translations, and classify their severity between *Minor* and *Major*. The quality score returned for a translation is computed following the MQM error weighting used by Google (see Section 3.1).

Mee (Mukherjee et al., 2020) is an automatic evaluation metric that leverages the similarity between embeddings of words in candidate translation and the corresponding reference. Unigrams are matched based on their surface forms, root forms and meanings while semantic evaluation is achieved by using pretrained fasttext embeddings. MEE computes evaluation score using three modules namely exact match, root match and synonym match. In each module, fmean-score is calculated giving more weight to recall. Final score is the average of the three individual modules.

Mee2 and Mee4★ (Mukherjee and Shrivastava, 2022b) are improved versions of Mee focusing on computing contextual and syntactic equivalences along with lexical, morphological and semantic similarity. The intent is to capture fluency and context of the MT outputs along with their adequacy. Fluency is captured using syntactic similarity and context is captured using sentence similarity leveraging sentence embeddings. The final score is the weighted combination of three similarity scores: a) syntactic similarity achieved by modified Bleu score; b) lexical, morphological and semantic similarity: measured by explicit unigram matching; c) contextual similarity: sentence similarity scores from Language-Agnostic BERT model.

Reuse★ (Mukherjee and Shrivastava, 2022a) is a bilingual, unsupervised reference-free metric. It estimates the translation quality at chunk-level and sentence-level. Source and target sentence chunks are retrieved by using a multi-lingual chunker. Chunk-level similarity is computed by leveraging BERT contextual word embeddings and sentence similarity scores are calculated by leveraging sentence embeddings of Language-Agnostic BERT models. The final quality estimation score is obtained by mean pooling the chunk-level and sentence-level similarity scores.

MetricX XL and MetricX XXL★ are massive multi-task metrics, which fine-tune large language model checkpoints such as mT5 on a variety of human feedback data such as DA, MQM, QE, NLI and Summarization Eval. The resulting primary submission uses the MQM score outputted by a fine-tuned 30B mT5.

UniTE★ (Wan et al., 2022a,b) is a learnt metric that can possess the ability of evaluating translation outputs following all three evaluation scenarios, i.e., source-only, reference-only, and source-reference-combined. Following their previous work, the authors improve their models by pre-training on pseudo-labeled data examples, and applying data cropping and a ranking-based score normalization during fine-tuning. The resulting submission is an ensemble of two models trained with different backbone models (XLM-R and InfoXLM).

SEScore★ (Xu et al., 2022) is an unsupervised reference-based evaluation metric, which takes model output and reference to produce a quality score. SEScore is trained from a pre-trained language model (Ex. Roberta) on synthetic triples generated from raw text. The synthetic triples consist of (raw text, synthetic error text, pseudo score), corresponding to (reference, model output, human rating). The data used for training the metric is constructed by synthesising candidate sentences y' to mimic plausible errors by transforming raw input sentences multiple times. At each step, a random span of text is selected and new content is inserted, deleted or replaced. All these errors are non-overlapping. The authors name this data construction process "stratified error synthesis", which randomly samples a set of potential errors and stochastically applies them on a given sentence. The score assigned to the perturbed sentences is a raw count of the severities applied by each transformation. In the end, SEScore is a regression quality prediction model trained on synthetic triples. Since this process can be applied to raw data and the resulting model can be developed for any text generation domain.

5 Meta Evaluation

Our main goal in evaluating metrics is to establish a ranking that reflects a metric's accuracy across a broad range of settings and applications. Combining results across different settings is challenging because correlations with human gold scores have different ranges and may be subject to differing degrees of noise. There are also many ways of measuring correlation, with different strengths and weaknesses, and it is often not clear which is best in a given setting.

This year, our overall ranking is just each metric's average rank across a large number of "tasks". Unlike raw correlation scores, ranks are comparable across tasks. The resulting global ranking approximates the "Kemeny consensus" – the ranking with lowest aggregate Kendall distance to the per-task rankings – which in turn satisfies several criteria from social choice theory (Colombo et al., 2022). Our version has the following features:

- We use a large number of tasks which may contain overlapping information. For instance, on each dataset, we compute both Pearson and Kendall-Tau correlation, and treat these as separate tasks. This makes the overall ranking robust to quirks in particular correlations.

- To guard against inadvertent bias toward settings that have more tasks than others, we use a task weighting that reflects the relative importance of various attributes (language pair, domain, etc.).

- Within each task, we establish a ranking that includes ties to reflect statistical significance. This naturally up-weights tasks that are more discriminative. For instance, a task that yields the ranking 1, 1, 1, 1 will not affect the overall ranking at all, while a ranking of 1, 2, 3, 4 is a maximal vote.

- In order to indicate metric proximity, we report raw averages over (weighted) per-task ranks rather than the resulting ranking as advocated by Colombo et al. (2022). For instance, average ranks of 1.1, 1.2, 2.1, 3.9 indicate that the top two metrics perform similarly and the last metric is considerably worse; these details is lost in the global ranking 1, 2, 3, 4.

- We also report rankings on selected subsets of tasks to characterize metric behavior on attributes such as language or domain.

5.1 Tasks

Tasks are identified by unique value assignments for each of the following attributes: language, domain, level, include-human, averaging method, and correlation. These are as follows:

Language (4 values)

Language pairs include those for which we have MQM ratings – English→German, English→Russian, and Chinese→English – plus *All*, which indicates all pairs pooled together.

Domain (5 values)

We computed correlations on domain-specific portions of each test-set as well as on each test-set as a whole. All language pairs have the same set of domains: *conversation*, *e-commerce*, *news*, and *social*. We use *mixed* to refer to all domains together, *i.e.*, the whole test set.

Level (2 values)

For each domain (including *mixed*), we computed correlations at the *system* level and the *segment* level. Human scores for each domain are averages over the corresponding segments. For metric submissions that did not include domain-level scores, we computed similar averages.

Include-human (2 values)

We computed separate correlations over sets of outputs that exclude human references (include-human=*false*) and that include all available references (include-human=*true*) except the standard reference, which is never scored by metrics. The first scenario reflects the standard use-case for metrics; the second captures a future scenario in which MT output quality approaches human quality. Since English→Russian has only a single reference, it participates only in the first condition. For the other two language pairs we use the reference that was judged best by the MQM raters. Table 6 summarizes the use of reference translations for different language pairs.

language	best ref	scored ref
en→de	A	B
en→ru	A	{}
zh→en	A	B

Table 6: Use of reference translations.

language	domain	level	+human	averaging	correlation	tasks	weight
all (1/4)	mixed (1/1)	sys (1/1)	no (1/1)	none (1/1)	acc (1/1)	1	1/4
en-ru (1/4)	* (1/5)	sys (1/2)	no (1/1)	none (1/1)	P,K (1/2)	10	1/80
		seg (1/2)	no (1/1)	* (1/3)	P,K (1/2)	30	1/240
en-de,zh-en (1/4)	* (1/5)	sys (1/2)	* (1/2)	none (1/1)	P,K (1/2)	40	1/160
		seg (1/2)	* (1/2)	* (1/3)	P,K (1/2)	120	1/480
						201	

Table 7: Task weighting. Column entries are sets of values for the attribute in the heading, with * designating all possible values. Numbers in brackets show the weight assigned to each value in the set. Each line corresponds to a set of tasks that have the same weight: the product of all the per-attribute weights shown in brackets. *P* and *K* refer to Pearson and Kendall correlation, respectively.

Averaging (3 values)

At the segment level, metric and human scores are naturally represented as system × segment matrices. However, correlations operate over pairs of vectors rather than pairs of matrices. There are three ways to resolve the problem: flatten the matrices into single vectors, compute average correlations over matching pairs of row vectors, or compute average correlations over matching pairs of column vectors. We designate these as *none*, *system*, and *segment* averaging, respectively. They measure a metric's ability to rate an arbitrarily-chosen (system, segment) pair, an arbitrary segment for a fixed system, and different system outputs for the same segment. Last year we used only the first alternative; this year include all three. System-level correlations do not require averaging, since their inputs are vectors in the first place.

Correlation (3 values)

We computed three correlations: system-level pairwise ranking *accuracy* (as proposed by Kocmi et al., 2021), *Pearson* and *Kendall*. Accuracy was used only for a single task in which all language pairs were pooled (language=*All*), while Pearson and Kendall were used for all other tasks. Pearson correlation tests linear fit with MQM scores, a stringent but reasonable criterion since we expect these scores to conform to a linear scale (for example, a translation with two minor errors is twice as bad as one with only a single error). Pearson has well-known drawbacks (Mathur et al., 2020), notably sensitivity to outliers, which we minimized by choosing only relatively high-performing systems. Like accuracy, Kendall is based on pairwise score comparisons, and thus reflects a common ranking use-case. It is susceptible to noise in gold pairwise rankings, for which a common strategy is to discard pairs judged not to be significantly different. We did not take this into account, relying instead on our significance tests for metric (rather than system) rankings.

5.2 Task Weighting

As explained in the previous section, attributes are not independent. For instance, there are three averaging methods for segment-level tasks, but only one for system-level tasks. If all tasks were weighted equally, this would have the undesirable consequence of making segment-level correlations count for 3× as much as system-level correlations when determining the overall ranking.

To avoid this, we used a hierarchical weighting scheme. We first ordered the attributes as listed in the previous section, then distributed weights evenly among all permissible values at each step of the hierarchy. The results are shown in Table 7. There are a total of 201 tasks, of which the accuracy task for all language pairs receives a weight of 1/4, with the remaining mass of 3/4 distributed among tasks whose individual weights vary between 1/80 and 1/420.

In Figures 1 through 4, we show analyses of how metric performance varies along different dimensions (attributes) such as language, domain, etc.. To do this, we partition tasks according to the values of the selected attribute, re-normalizing their global weights so they sum to 1 for each partition. We then compute weighted average ranks for each partition separately, in the same fashion as the overall ranking.

5.3 Per-task Ranking

For each task, we compare all pairs of metrics, and determine whether the difference in their correlation scores is significant according to the PERM-

BOTH hypothesis test of Deutsch et al. (2021), using 1000 re-sampling runs, and setting $p = 0.05$. For the averaging methods, sampling is performed separately for each row or column vector prior to averaging.

We then assign ranks as follows. Starting with the highest-scoring metric, we move down the list of metrics in descending order by score, and assign rank 1 to all metrics until we encounter the first metric that is significantly different from any that have been visited so far. That metric is assigned rank 2, and the process is repeated. This continues until all metrics have been assigned a rank.

6 Main Results

As we have seen in Section 5, the main results are defined across different settings including system-level and segment-level tasks. Nonetheless, since the main use case of automatic metrics is to rank systems, system-level accuracy has a 1/4 weight on the final score with the remaining 3/4 distributed over 200 different settings.

Table 1 shows the official ranking of all primary submissions over the 201 different settings. A key observation is that neural metrics perform significantly better than lexical metrics. Of the 20 evaluated metrics, BLEU and SPBLEU are ranked 19th and 17th respectively. On the other hand, fine-tuned neural baseline metrics such as COMET-20 and BLEURT-20 are still ranked above several of the new primary submissions. They are outperformed only by submissions based on models that are considerably larger[11]. Figure 1 shows the ranking split by the different language pairs. The trend is very similar for all language pairs. While METRICX XXL performs best for En→De and En→Ru, COMET-22 performs best for Zh→En.

One open question about neural metrics has been their ability to generalise to new domains, since most training and testing data from previous years were based on News data. In Figure 2 we present the performance of each metric across four domains: news, social, conversational, and e-commerce. Similar to last year, we observe that the neural metrics perform better than lexical overlap metrics across all four domains.

Figure 3 shows the average rankings when grouped separately by system-level and segment-level tasks. Many metrics fall into the same significance cluster when evaluated on the system-level as we only have a very limited number of MT systems. Nevertheless, we observe that the metric rankings are largely stable across both granularities and that METRICX XXL and COMET-22 perform best on both the segment-level and system-level tasks. The differences are more prevalent in the segment-level task, though.

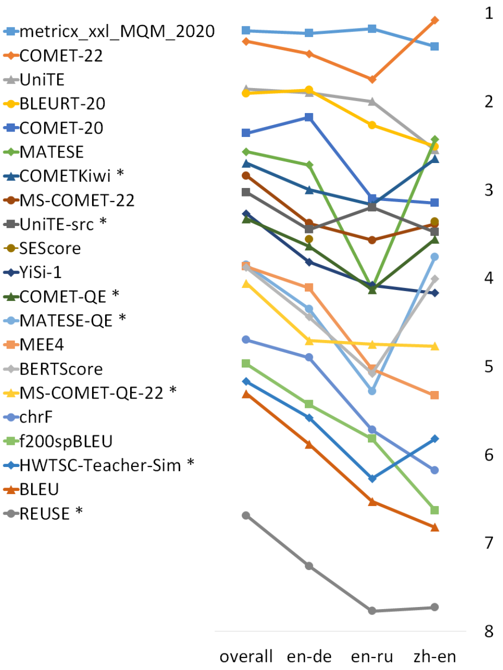

Figure 1: Weighted ranking of metrics' correlation with human grouped by translation directions.

In Figure 4, we compare the rankings when including human translations as MT systems (with human) or just considering MT submission (without human). Overall, the majority of metrics show

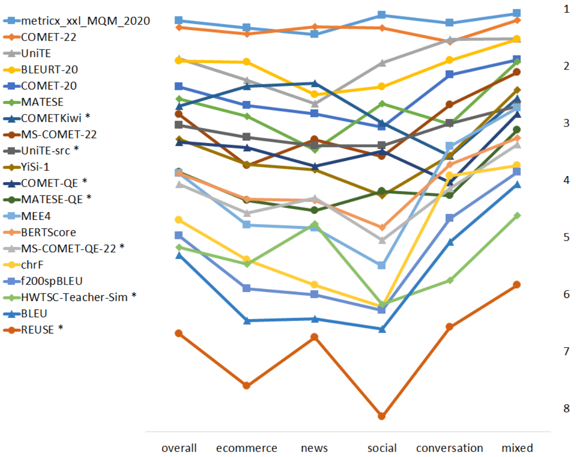

Figure 2: Weighted ranking of metrics' correlation with human grouped by domains.

[11] Both UNITE and COMET-22 are ensembles of two models trained on XLM-R variants while METRICX XXL uses mT5 XXL as a backbone

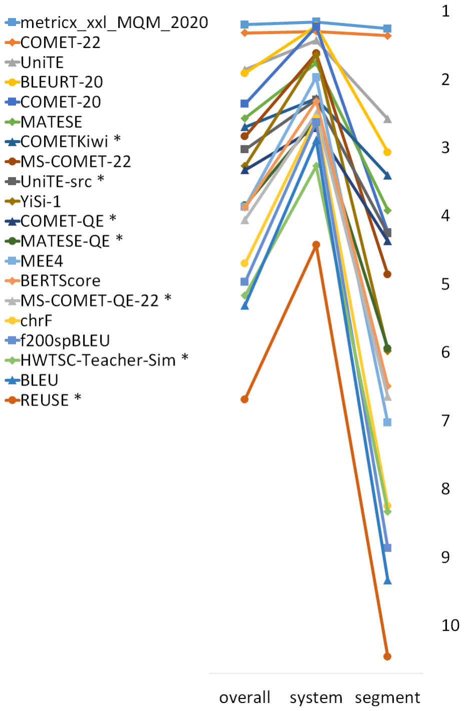

Figure 3: Weighted ranking of metrics' correlation with human grouped by granularity levels.

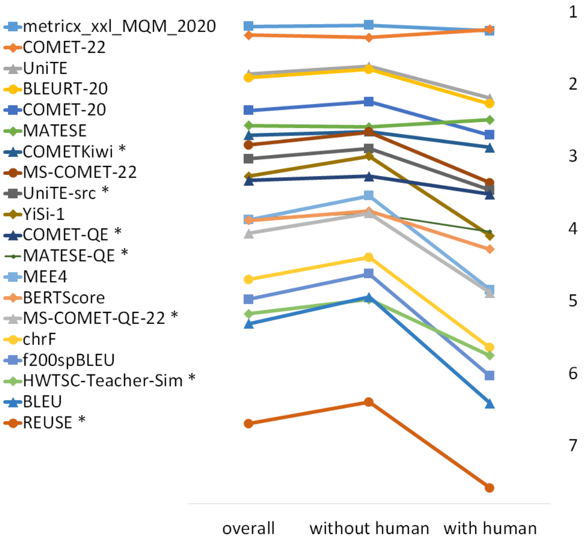

Figure 4: Weighted ranking of metrics' correlation with human grouped by candidate pools (with or without human translations).

lower correlation when we include human translations, except COMET-22 and MATESE.

7 Direct Assessment Human Evaluation

In addition to our MQM annotations and as a contrastive evaluation to cover more language pairs, we look into the performance of metrics when compared to the human evaluation campaign conducted by the General MT shared task (Kocmi et al., 2022a), who ran human evaluation for all 21 translation directions and WMT22 submissions. Last year, we decided to exclude the human ratings by the WMT main task as they were of lower quality than the best automatic metrics. However, the GeneralMT task improved their evaluation methodology in particular for all from-English and non-English translation directions and implemented the Scalar Quality Metric (SQM) which has been shown to have high correlation with MQM on at least the system-level (Freitag et al., 2021a). The GeneralMT task used two different human evaluation methodologies depending on the language pair: reference-based Direct Assessment (Ref. DA) (Graham et al., 2013) and SQM style source-based DA (DA+SQM) (Kocmi et al., 2022a).

Ref. DA has been used for all into-English translation directions and asks human raters to judge each system translation against human reference translation on a 0–100 scale. This technique does not use bilingual speakers and is evaluated by non-professional crowd workers. In order to increase quality of assessment, there are several quality control items. Out of all collected human annotations, 63% have been removed due to failing quality control.

DA+SQM asks bilingual raters to annotate system translations against original sources on a 0–100 labeled scale. The scale is marked with seven points representing expected quality. In this setting, Kocmi et al. (2022a) evaluated all from-English and non-English translation directions. They used mainly professional raters.

We present system-level accuracy results in Table 8. The ranking generated based on accuracy scores when taking the DA+SQM annotation as ground truths is comparable to the primary results in Table 1, ranking METRICX XXL as the best performing metric followed by UNITE and COMET-22. Similarly, it ranks n-gram matching metrics (BLEU, CHRF, F101SPBLEU) among worst performing metrics. This confirms the main findings from MQM evaluation.

On the other hand, accuracy scores taking ref. DA as the ground truth, result in a very different ranking of the metrics. It ranks n-gram matching metrics as the top performing metrics. This suggest that the technique does not evaluate systems well

Number of languages	13	6
Number of system pairs	564	329
Human judgement style	DA+SQM	ref. DA
MetricX XXL	**0.862** (1)	0.620 (11)
UniTE	0.849 (2)	0.623 (10)
Comet-22	0.842 (3)	0.626 (9)
CometKiwi*	0.835 (4)	0.617 (12)
MS-Comet-22	0.833 (5)	0.626 (9)
Bleurt-20	0.830 (6)	0.650 (5)
Comet-20	0.826 (7)	0.635 (8)
MS-Comet-qe-22*	0.824 (8)	0.641 (7)
Comet-QE*	0.821 (9)	0.605 (13)
UniTE-src*	0.800 (10)	0.623 (10)
YiSi-1	0.785 (11)	0.660 (3)
BERTScore	0.764 (12)	0.666 (2)
chrF	0.762 (13)	0.666 (2)
EE_BERTScore	0.750 (14)	0.647 (6)
f101spBleu	0.748 (15)	**0.669 (1)**
HWTSC-Teacher-Sim*	0.720 (16)	0.568 (15)
Bleu	0.707 (17)	0.653 (4)
Reuse*	0.344 (18)	0.584 (14)

Table 8: System-level pairwise accuracy for WMT style human evaluation. Numbers in brackets show rank of metrics given human judgement style. The highest score is present bolded.

and instead human crowd workers are incentivized to quickly compare the surface forms of translation against reference without understanding. We would advise metric developers and researchers running human evaluations not to use reference-based DA, especially when evaluated with non-professional crowd workers.

8 Challenge Sets Subtask

The challenge sets subtask is inspired by the *Build it or break it: The Language Edition* shared task (Ettinger et al., 2017) which aimed at testing the generalizability of NLP systems beyond the distributions of their training data. With that said, our goal is to encourage researchers to build a set of test sets that measure metrics' ability to detect different targeted phenomena that might not be well represented in traditional test sets used to evaluate metrics.

This subtask is made of three consecutive phases; 1) the *Breaking Round*, 2) the *Scoring Round* and 3) the *Analysis Round*:

1. In the *Breaking Round*, the challenge set participants (*Breakers*) submit their challenge sets composed of contrastive examples for different phenomena with source sentences (s), incorrect translations ($\hat{t}$), correct translations (t) and references (r).

2. In the *Scoring Round* the metrics participants from the main task (the *Builders*) are asked to score all translations with their metrics without knowing which ones are correct or incorrect. Also, in this phase the organisers score all data with the baseline metrics.

3. Finally, after gathering all metric scores, the data is returned to the *Breakers* for the *Analysis round*, where they look at which metrics are able to correctly rank the correct translations above the incorrect ones for the different phenomena being tested.

We had a total of 4 submissions to this shared task, covering a wide range of phenomena and 146 different language pairs. Table 9 provides an overview of the submitted challenge sets. A short description of every submission follows:

ACES The ACES (Translation Accuracy Challenge Sets; Amrhein et al., 2022) results from a collaboration between the University of Zurich with the University of Edinburgh. This challenge set, highly inspired by the MQM framework, consists of 36,499 examples, covering 146 language pairs and 68 phenomena, ranging from simple perturbations at the word/character level to more complex errors based on discourse and real-world knowledge. The data was created artificially for some error types and manually for others.

Their analysis aimed to reveal the extent to which metrics take into account the source sentence context and the surface-level overlap with the reference, and if they profit by using multilingual embeddings. Finally, they recommend that one considers a) **combining metrics with different strengths** and b) explicitly **modelling additional language-specific information** beyond what is available via multilingual embeddings.

SMAUG The challenge set based on Sentence-level Multilingual data Augmentation (SMAUG; Alves et al., 2022), submitted by Unbabel and IST evaluates the robustness of MT metrics to 5 different types of translation errors; Named entity errors, numerical errors, meaning errors, insertion of content and content missing. These errors are created by perturbing reference translations and then curated by the authors. The challenge set covers 3

challenge set	method	lang. pairs	pheno-mena	items	citation	availability (https://github.com/)
ACES	automatic	146	68	36,499	Amrhein et al. (2022)	EdinburghNLP/ACES
DFKI-CS	semi-autom.	2	107	19,347	Avramidis and Macketanz (2022)	DFKI-NLP/mt-testsuite
HwTsc-CS	semi-autom.	1	5	721	Chen et al. (2022)	HwTsc/Challenge-Set-for-MT-Metrics
SMAUG	automatic	3	5	632	Alves et al. (2022)	Unbabel/smaug

Table 9: Overview of the participations at the challenge sets task

language pairs and contains close to 50 high-quality examples for each phenomenon.

In this challenge set the authors show that there has been a promising progress in terms of detecting these critical errors when compared to last year's metric submissions. Nevertheless, errors related to **named entities and numbers were found to pose a challenge for several tested metrics**. Also, due to a high variance in the observed results across all the error types it becomes **hard to predict performance of current methods with respect to untested translation errors**.

HWTSC Challenge Set The challenge set submitted by Huawei Translation Services Center (Chen et al., 2022) aims at examining metrics ability to handle synonyms and to discern critical errors in translations. This challenge set is composed of 721 zh-en examples for 5 different error types; Named entity errors, numerical errors, time & date errors, wrong unit conversions and Affirmation/Negation errors. The underlying data is either WMT 21 or Flores 101 which covers two distinct domains, News and Wikipedia respectively. To create alternative translations the authors used in-house translators (performing post-edit) and to create the adversarial translations they used LIST (Alzantot et al., 2018).

The authors of this challenge set conclude that although embedding-based metrics perform relatively well on discerning sentence-level negation/affirmation errors, they **perform poorly on relating synonyms**. Additionally they find that the generalizability of some metrics is compromised, as they are **susceptible to different text styles**.

DFKI Challenge Set The submission by DFKI (Avramidis and Macketanz, 2022) employs a linguistically motivated challenge set that includes about 20,000 items extracted from 145 MT systems for two language directions (German⇔English). It is based on a test suite (Macketanz et al., 2022) that covers more than 100 linguistically-motivated phenomena organized in 14 categories.

The best performing metrics are YiSi-1, BERTScore and COMET-22 for German-English, and UniTE, UniTE-ref, MetricX-xl-DA-2019 and MetricX-xxl-DA-2019 for English-German. Metrics in both directions are performing worst when it comes to **named-entities & terminology** and particularly **measuring units**. Particularly in German-English they are weak at detecting issues at **punctuation, polar questions, relative clauses, dates** and **idioms**. In English-German, they perform worst at **present progressive of transitive verbs, future II progressive of intransitive verbs, simple present perfect of ditransitive verbs** and **focus particles**.

9 Conclusion

This paper summarizes the results of the WMT22 shared task on automated machine translation evaluation, the Metrics Shared Task. We presented an extensive analysis on how well metrics perform on our three main language pairs: English→German, English→Russian and Chinese→English. The results, based on 201 different tasks, demonstrated the superiority of neural-based learned metrics over overlap-based metrics like Bleu, spBleu or chrF. These results are confirmed with DA+SQM human judgement. Although this was already the case in the previous years' Metric Shared Tasks, we further strengthened the case for neural-based fine-tuned metrics by demonstrating their superiority across four different domains. In addition, we initiated a challenge set subtask, where participants had to create contrastive test suites for evaluating metrics' ability to capture and penalise specific types of translation errors.

10 Ethical Considerations

MQM annotations and additional reference translations in this paper are done by professional translators. They are all paid at professional rates.

Organizers from the National Research Council

Canada and Unbabel have submitted to this task the frozen stable versions of their metrics (YiSi and COMET) dated before this year's shared task and publicly available. Newer versions of COMET were developed without using any of the test set, test suite or challenge sets.

11 Acknowledgments

Results for this shared task would not be possible without tight collaboration with the organizers of the General MT Task. We are grateful to Google and Unbabel for sponsoring and overseeing the human evaluation.

Ricardo Rei and André F. T. Martins are supported by the P2020 program (MAIA: contract 045909) and by European Union's Horizon Europe Research and Innovation Actions (UTTER: contract 101070631). Additionally, André F. T. Martins is supported by the European Research Council (ERC StG DeepSPIN, 758969) and Fundação para a Ciência e Tecnologia (contract UIDB/50008/2020).

Eleftherios Avramidis is supported by the German Ministry of Education and Research through the research program SocialWear (grant num. 01IW2000).

References

Duarte M. Alves, Ricardo Rei, Ana C. Farinha, José G. C. de Souza, and André F. T. Martins. 2022. Robust MT evaluation with Sentence-level Multilingual data Augmentation. In *Proceedings of the Seventh Conference on Machine Translation*, Abu Dhabi. Association for Computational Linguistics.

Moustafa Alzantot, Yash Sharma, Ahmed Elgohary, Bo-Jhang Ho, Mani Srivastava, and Kai-Wei Chang. 2018. Generating natural language adversarial examples. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2890–2896, Brussels, Belgium. Association for Computational Linguistics.

Chantal Amrhein, Nikita Moghe, and Liane K. Guillou. 2022. ACES: Translation Accuracy Challenge Sets for Evaluating Machine Translation Metrics. In *Proceedings of the Seventh Conference on Machine Translation*, Abu Dhabi. Association for Computational Linguistics.

Eleftherios Avramidis and Vivien Macketanz. 2022. Linguistically motivated evaluation of machine translation metrics based on a challenge set. In *Proceedings of the Seventh Conference on Machine Translation*, Abu Dhabi. Association for Computational Linguistics.

Xiaoyu Chen, Daimeng Wei, Hengchao Shang, Zongyao Li, Zhanglin Wu, Zhengzhe Yu, Ting Zhu, Mengli Zhu, Ning Xie, Lizhi Lei, Shimin Tao, Hao Yang, and Ying Qin. 2022. Exploring Robustness of Machine Translation Metrics: A Study of Twenty-Two Automatic Metrics in the WMT22 Metric Task. In *Proceedings of the Seventh Conference on Machine Translation*, Abu Dhabi. Association for Computational Linguistics.

Pierre Colombo, Nathan Noiry, Ekhine Irurozki, and Stéphan Clémençon. 2022. What are the best systems? new perspectives on nlp benchmarking. *arXiv preprint arXiv:2202.03799*.

Daniel Deutsch, Rotem Dror, and Dan Roth. 2021. A statistical analysis of summarization evaluation metrics using resampling methods. *arXiv preprint arXiv:2104.00054*.

Bryan Eikema and Wilker Aziz. 2020. Is MAP decoding all you need? the inadequacy of the mode in neural machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4506–4520, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Bryan Eikema and Wilker Aziz. 2021. Sampling-based minimum bayes risk decoding for neural machine translation.

Allyson Ettinger, Sudha Rao, Hal Daumé III, and Emily M. Bender. 2017. Towards linguistically generalizable NLP systems: A workshop and shared task. In *Proceedings of the First Workshop on Building Linguistically Generalizable NLP Systems*, pages 1–10, Copenhagen, Denmark. Association for Computational Linguistics.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2021. Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research*, 22(1).

Patrick Fernandes, António Farinhas, Ricardo Rei, José De Souza, Perez Ogayo, Graham Neubig, and Andre Martins. 2022. Quality-aware decoding for neural machine translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1396–1412, Seattle, United States. Association for Computational Linguistics.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021a. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.

Markus Freitag, David Grangier, Qijun Tan, and Bowen Liang. 2022. High quality rather than high model probability: Minimum Bayes risk decoding with neural metrics. *Transactions of the Association for Computational Linguistics*, 10:811–825.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, George Foster, Alon Lavie, and Ondřej Bojar. 2021b. Results of the WMT21 metrics shared task: Evaluating metrics with expert-based human evaluations on TED and news domain. In *Proceedings of the Sixth Conference on Machine Translation*, pages 733–774, Online. Association for Computational Linguistics.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

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

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. In *International Conference on Learning Representations*.

Tom Kocmi, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Thamme Gowda, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Rebecca Knowles, Philipp Koehn, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Michal Novák, Martin Popel, Maja Popović, and Mariya Shmatova. 2022a. Findings of the 2022 conference on machine translation (wmt22). In *Proceedings of the Seventh Conference on Machine Translation*. Association for Computational Linguistics.

Tom Kocmi, Christian Federmann, Roman Grundkiewicz, Marcin Junczys-Dowmunt, Hitokazu Matsushita, and Arul Menezes. 2021. To ship or not to ship: An extensive evaluation of automatic metrics for machine translation.

Tom Kocmi, Hitokazu Matsushita, and Christian Federmann. 2022b. MS-COMET: Larger Filtered Human Annotations Help Metric Performance. In *Proceedings of the Seventh Conference on Machine Translation*, Abu Dhabi. Association for Computational Linguistics.

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

Yilun Liu, Xiaosong Qiao, Zhanglin Wu, Su Chang, Min Zhang, Yanqing Zhao, shimin tao Song Peng, Hao Yang, Ying Qin, Jiaxin Guo, Minghan Wang, Yinglu Li, Peng Li, and Xiaofeng Zhao. 2022. Partial Could Be Better Than Whole: HW-TSC 2022 Submission for the Metrics Shared Task. In *Proceedings of the Seventh Conference on Machine Translation*, Abu Dhabi. Association for Computational Linguistics.

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

Arle Lommel, Hans Uszkoreit, and Aljoscha Burchardt. 2014. Multidimensional Quality Metrics (MQM) : A Framework for Declaring and Describing Translation Quality Metrics. *Tradumàtica*, pages 0455–463.

Vivien Macketanz, Eleftherios Avramidis, Aljoscha Burchardt, He Wang, Renlong Ai, Shushen Manakhimova, Ursula Strohriegel, Sebastian Möller, and Hans Uszkoreit. 2022. A linguistically motivated test suite to semi-automatically evaluate German–English machine translation output. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 936–947, Marseille, France. European Language Resources Association.

Nitika Mathur, Timothy Baldwin, and Trevor Cohn. 2020. Tangled up in bleu: Reevaluating the evaluation of automatic machine translation evaluation metrics. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4984–4997.

Ananya Mukherjee, Hema Ala, Manish Shrivastava, and Dipti Misra Sharma. 2020. Mee: an automatic metric for evaluation using embeddings for machine translation. In *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 292–299. IEEE.

Ananya Mukherjee and Manish Shrivastava. 2022a. REUSE: REference-free UnSupervised quality Estimation Metric. In *Proceedings of the Seventh Conference on Machine Translation*, Abu Dhabi. Association for Computational Linguistics.

Ananya Mukherjee and Manish Shrivastava. 2022b. Unsupervised Embedding-based Metric for MT Evaluation with Improved Human Correlation. In *Proceedings of the Seventh Conference on Machine Translation*, Abu Dhabi. Association for Computational Linguistics.

Mathias Müller and Rico Sennrich. 2021. Understanding the properties of minimum Bayes risk decoding in neural machine translation. 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 259–272, Online. Association for Computational Linguistics.

Nathan Ng, Kyra Yee, Alexei Baevski, Myle Ott, Michael Auli, and Sergey Edunov. 2019. Facebook FAIR's WMT19 news translation task submission. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 314–319, Florence, Italy. Association for Computational Linguistics.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No Language Left Behind: Scaling Human-Centered Machine Translation.

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

Stefano Perrella, Lorenzo Proietti, Alessandro Scirè, Niccolò Campolungo, and Roberto Navigli. 2022. Machine Translation Evaluation as a Sequence Tagging Problem. In *Proceedings of the Seventh Conference on Machine Translation*, Abu Dhabi. Association for Computational Linguistics.

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

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

Amy Pu, Hyung Won Chung, Ankur Parikh, Sebastian Gehrmann, and Thibault Sellam. 2021. Learning compact metrics for MT. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 751–762, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. COMET-22: Unbabel-IST 2022 Submission for the Metrics Shared Task. In *Proceedings of the Seventh Conference on Machine Translation*, Abu Dhabi. Association for Computational Linguistics.

Ricardo Rei, Ana C Farinha, Chrysoula Zerva, Daan van Stigt, Craig Stewart, Pedro Ramos, Taisiya Glushkova, André F. T. Martins, and Alon Lavie. 2021. Are references really needed? unbabel-IST 2021 submission for the metrics shared task. In *Proceedings of the Sixth Conference on Machine Translation*, pages 1030–1040, Online. Association for Computational Linguistics.

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

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

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2021. Multilingual translation from denoising pre-training. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3450–3466, Online. Association for Computational Linguistics.

Yu Wan, Keqin Bao, Dayiheng Liu, Baosong Yang, Derek F. Wong, Lidia S. Chao, Wenqiang Lei, and Jun Xie. 2022a. Alibaba-Translate China's Submission for WMT2022 Metrics Shared Task. In *Proceedings of the Seventh Conference on Machine Translation*, Abu Dhabi. Association for Computational Linguistics.

Yu Wan, Dayiheng Liu, Baosong Yang, Haibo Zhang, Boxing Chen, Derek Wong, and Lidia Chao. 2022b. UniTE: Unified translation evaluation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8117–8127, Dublin, Ireland. Association for Computational Linguistics.

Wenda Xu, Yilin Tuan, Yujie Lu, Michael Saxon, Lei Li, and William Yang Wang. 2022. Not all errors are equal: Learning text generation metrics using stratified error synthesis. *arXiv preprint arXiv:2210.05035*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Task	Accuracy	en-de	en-de	en-ru	zh-en	zh-en
Human Translation Included	No	Yes	No	No	Yes	No
metricx_xl_DA_2019	0.865	0.908	0.905	0.977	0.966	0.982
metricx_xxl_DA_2019	0.865	0.907	0.901	0.982	0.961	0.984
metricx_xxl_MQM_2020	0.850	0.862	0.847	0.949	0.924	0.920
BLEURT-20	0.847	0.691	0.719	0.959	0.909	0.938
metricx_xl_MQM_2020	0.843	0.848	0.832	0.927	0.920	0.914
COMET-22	0.839	0.761	0.771	0.900	0.947	0.942
COMET-20	0.836	0.812	0.876	0.936	0.964	0.970
UniTE	0.828	0.642	0.624	0.888	0.922	0.914
MS-COMET-22	0.828	0.634	0.695	0.809	0.918	0.909
UniTE-ref	0.818	0.652	0.632	0.831	0.902	0.892
MATESE	0.810	0.647	0.617	0.757	0.869	0.856
YiSi-1	0.792	0.506	0.626	0.881	0.867	0.935
MEE4	0.788	0.404	0.537	0.792	0.818	0.905
COMETKiwi*	0.788	0.592	0.674	0.763	0.795	0.866
HuaweiTSC_EE_BERTScore_0.8_With_Human	0.785	0.354	0.463	0.818	0.903	0.960
HuaweiTSC_EE_BERTScore_0.8_Without_Human	0.785	0.338	0.451	0.818	0.900	0.957
Cross-QE*	0.781	0.643	0.661	0.806	0.817	0.870
HuaweiTSC_EE_BERTScore_0.5_With_Human	0.781	0.287	0.400	0.792	0.938	0.953
COMET-QE*	0.781	0.480	0.502	0.468	0.544	0.569
HuaweiTSC_EE_BERTScore_0.5_Without_Human	0.774	0.246	0.370	0.795	0.930	0.942
BERTScore	0.774	0.338	0.428	0.811	0.843	0.924
HuaweiTSC_EE_BERTScore_0.3_With_Human	0.759	0.243	0.356	0.754	0.945	0.943
UniTE-src*	0.759	0.509	0.509	0.779	0.791	0.874
MEE2	0.759	0.360	0.479	0.811	0.753	0.872
MS-COMET-QE-22*	0.755	0.417	0.539	0.672	0.799	0.897
MATESE-QE*	0.748	0.363	0.337	0.637	0.741	0.767
MEE	0.748	0.358	0.445	0.823	0.727	0.824
f101spBLEU	0.745	0.210	0.298	0.816	0.613	0.718
f200spBLEU	0.741	0.230	0.283	0.819	0.614	0.728
HuaweiTSC_EE_BERTScore_0.3_Without_Human	0.737	0.189	0.316	0.761	0.931	0.926
chrF	0.734	0.159	0.346	0.815	0.647	0.630
BLEU	0.708	0.038	0.179	0.724	0.579	0.594
HWTSC-TLM*	0.697	0.311	0.428	0.597	0.368	0.460
HWTSC-Teacher-Sim*	0.686	0.290	0.385	0.675	0.294	0.356
KG-BERTScore*	0.664	0.369	0.400	0.612	0.617	0.743
REUSE*	0.347	-0.514	-0.465	-0.349	-0.330	-0.142
SEScore	–	0.581	0.660	–	0.920	0.944

Table 10: Pearson correlation of all metrics with system-level MQM scores for the three main language pairs. Rows are sorted by the system-level pairwise accuracy across the three language pairs. Primary submissions are bolded, and baselines are underlined. Reference-free metrics are indicated using an asterisk.

A Language-Specific Results Tables

Language-specific results are given in Table 10 and Table 11. Each page contains results for scores over all domains over a single granularity (system or segment).

For all tables, the correlations are calculated on metric scores comparing MT system translations with Reference A, and any additional human reference translations are not included.

For segment level correlation, we report results on the "none" averaging method, where we flatten the matrices into single vectors before computing the Kendall Tau correlation.

B Correlations with WMT Human Evaluation

Correlations with WMT Direct Assessment Human scores are given in the following tables, with results for language pairs evaluated using reference-based Direct Assessment (Ref. DA) (Graham et al., 2013), followed by results for language pairs evaluated using SQM style source-based DA (DA+SQM) (Kocmi et al., 2022a). Since most language pairs contained only a single reference, we used reference A for all pairs, and report results only for scoring MT output (omitting additional scored references for language pairs where these were available). System-level correlations use Pearson and segment-level scores use Kendall. For simplicity, both statistics are computed over raw rater scores, with no traditional difference-25

Task	(sys) Accuracy	en-de	en-de	en-ru	zh-en	zh-en
Human Translation Included	No	Yes	No	No	Yes	No
metricx_xl_DA_2019	0.865	0.356	0.362	0.393	0.383	0.392
metricx_xxl_DA_2019	0.865	0.355	0.361	0.405	0.377	0.386
metricx_xxl_MQM_2020	0.850	0.356	0.360	0.420	0.421	0.427
BLEURT-20	0.847	0.338	0.344	0.359	0.352	0.361
metricx_xl_MQM_2020	0.843	0.362	0.367	0.383	0.416	0.423
COMET-22	0.839	0.361	0.368	0.400	0.420	0.428
COMET-20	0.836	0.312	0.319	0.330	0.325	0.332
UniTE	0.828	0.362	0.369	0.378	0.351	0.357
MS-COMET-22	0.828	0.277	0.283	0.351	0.335	0.341
UniTE-ref	0.818	0.356	0.362	0.374	0.354	0.361
MATESE	0.810	0.323	0.323	0.279	0.382	0.389
YiSi-1	0.792	0.229	0.235	0.227	0.288	0.296
MEE4	0.788	0.236	0.243	0.210	0.189	0.194
COMETKiwi*	0.788	0.283	0.290	0.359	0.352	0.364
HuaweiTSC_EE_BERTScore_0.8_With_Human	0.785	–	–	–	–	–
HuaweiTSC_EE_BERTScore_0.8_Without_Human	0.785	–	–	–	–	–
Cross-QE*	0.781	0.259	0.263	0.310	0.368	0.378
HuaweiTSC_EE_BERTScore_0.5_With_Human	0.781	–	–	–	–	–
COMET-QE*	0.781	0.277	0.281	0.341	0.356	0.365
HuaweiTSC_EE_BERTScore_0.5_Without_Human	0.774	–	–	–	–	–
BERTScore	0.774	0.226	0.232	0.192	0.307	0.316
HuaweiTSC_EE_BERTScore_0.3_With_Human	0.759	–	–	–	–	–
UniTE-src*	0.759	0.283	0.287	0.342	0.332	0.343
MEE2	0.759	0.238	0.244	0.201	0.197	0.201
MS-COMET-QE-22*	0.755	0.226	0.233	0.305	0.277	0.287
MATESE-QE*	0.748	0.242	0.244	0.229	0.328	0.337
MEE	0.748	0.187	0.192	0.148	0.149	0.149
f101spBLEU	0.745	0.169	0.174	0.135	0.143	0.145
f200spBLEU	0.741	0.176	0.180	0.153	0.139	0.140
HuaweiTSC_EE_BERTScore_0.3_Without_Human	0.737	–	–	–	–	–
chrF	0.734	0.208	0.214	0.168	0.146	0.147
BLEU	0.708	0.164	0.169	0.140	0.143	0.145
HWTSC-TLM*	0.697	0.087	0.092	0.121	0.079	0.086
HWTSC-Teacher-Sim*	0.686	0.150	0.155	0.143	0.264	0.272
KG-BERTScore*	0.664	0.126	0.129	0.111	0.214	0.219
REUSE*	0.347	0.057	0.065	0.078	0.116	0.130
SEScore	–	0.261	0.266	–	0.324	0.331

Table 11: Kendall Tau correlation of all metrics with segment-level MQM scores for the three main language pairs. Rows are sorted by the system-level pairwise accuracy across the three language pairs. Primary submissions are bolded, and baselines are underlined. Reference-free metrics are indicated using an asterisk.

filtering.[12]

[12] The traditional recipe made little difference in overall correlation patterns.

Task Incl. Human Translation	Accuracy False	cs-en False	de-en False	ja-en False	ru-en False	uk-en False	zh-en False
f200spBLEU	0.669	0.812	0.405	0.949	0.831	0.714	0.517
chrF	0.666	0.806	0.354	0.983	0.827	0.688	0.568
BERTScore	0.666	0.825	0.440	0.988	0.851	0.717	0.396
YiSi-1	0.660	0.824	0.443	0.989	0.847	0.708	0.415
f101spBLEU	0.660	0.810	0.406	0.944	0.830	0.718	0.521
BLEU	0.653	0.801	0.352	0.934	0.843	0.648	0.563
BLEURT-20	0.650	0.833	0.458	0.990	0.849	0.733	0.266
HWTSC_EE_BERTScore_0.8_Without_Human	0.647	0.824	0.442	0.989	0.858	0.714	0.417
HWTSC_EE_BERTScore_0.3_Without_Human	0.647	0.808	0.391	0.987	0.876	0.678	0.437
HWTSC_EE_BERTScore_0.3_With_Human	0.647	0.799	0.390	0.987	0.876	0.680	0.412
HWTSC_EE_BERTScore_0.8_With_Human	0.644	0.820	0.440	0.989	0.858	0.715	0.411
HWTSC_EE_BERTScore_0.5_With_Human	0.644	0.808	0.410	0.988	0.870	0.696	0.416
HWTSC_EE_BERTScore_0.5_Without_Human	0.644	0.815	0.411	0.988	0.870	0.695	0.434
MS-COMET-QE-22*	0.641	0.769	0.395	0.990	0.867	0.699	0.312
COMET-20	0.635	0.827	0.424	0.989	0.847	0.723	0.330
metricx_xxl_DA_2019	0.635	0.831	0.469	0.987	0.850	0.730	0.148
UniTE-ref	0.629	0.822	0.440	0.982	0.855	0.727	0.167
MS-COMET-22	0.626	0.807	0.419	0.990	0.858	0.701	0.108
COMET-22	0.626	0.821	0.446	0.976	0.857	0.714	0.135
metricx_xl_DA_2019	0.623	0.833	0.468	0.987	0.851	0.730	0.157
Cross-QE*	0.623	0.791	0.415	0.989	0.863	0.719	0.129
UniTE	0.623	0.832	0.431	0.984	0.852	0.728	0.195
UniTE-src*	0.623	0.777	0.402	0.989	0.863	0.703	0.210
metricx_xl_MQM_2020	0.620	0.821	0.487	0.978	0.856	0.718	-0.039
metricx_xxl_MQM_2020	0.620	0.823	0.490	0.978	0.856	0.715	-0.061
COMETKiwi*	0.617	0.787	0.409	0.984	0.862	0.718	0.181
COMET-QE*	0.605	0.811	0.443	0.981	0.864	0.744	-0.006
REUSE*	0.584	0.200	0.194	0.990	0.683	0.150	0.531
HWTSC-TLM*	0.578	0.822	0.356	0.980	0.842	0.695	0.083
HWTSC-Teacher-Sim*	0.568	0.804	0.322	0.985	0.848	0.691	-0.011
KG-BERTScore*	0.568	0.539	0.052	0.989	0.805	0.516	0.264
MEE	–	–	–	–	–	–	0.578
MEE2	–	–	–	–	–	–	0.511
MEE4	–	–	–	–	–	–	0.455
SEScore	–	–	–	–	–	–	0.331
MATESE	–	–	–	–	–	–	0.013
MATESE-QE*	–	–	–	–	–	–	0.013

Table 12: System-level Pearson correlation with crowdsourced Ref. DA scores. Rows are sorted by the system-level pairwise accuracy across all language pairs. Primary submissions are bolded, and baselines are underlined. Reference-free metrics are indicated using an asterisk.

System-level Metric accuracy and correlations with REFDA scores contradict the main results. We strongly recommend against using Ref. DA scores to evaluate MT metrics.

Task	(sys) Accuracy	cs-en	de-en	ja-en	ru-en	uk-en	zh-en
Incl. Human Translation	False	False	False	False	False	False	False
f200spBLEU	0.669	0.043	0.010	0.085	0.018	0.006	0.026
chrF	0.666	0.042	0.017	0.083	0.015	0.003	0.025
BERTScore	0.666	0.039	0.011	0.084	0.019	0.003	0.020
YiSi-1	0.660	0.037	0.012	0.087	0.018	0.004	0.020
f101spBLEU	0.660	0.042	0.010	0.085	0.020	0.008	0.026
BLEU	0.653	0.043	0.009	0.081	0.014	0.007	0.024
BLEURT-20	0.650	0.036	0.018	0.085	0.014	0.002	0.013
HWTSC_EE_BERTScore_0.8_Without_Human	0.647	–	–	–	–	–	–
HWTSC_EE_BERTScore_0.3_Without_Human	0.647	–	–	–	–	–	–
HWTSC_EE_BERTScore_0.3_With_Human	0.647	–	–	–	–	–	–
HWTSC_EE_BERTScore_0.8_With_Human	0.644	–	–	–	–	–	–
HWTSC_EE_BERTScore_0.5_With_Human	0.644	–	–	–	–	–	–
HWTSC_EE_BERTScore_0.5_Without_Human	0.644	–	–	–	–	–	–
MS-COMET-QE-22*	0.641	0.022	0.011	0.088	-0.002	0.003	0.001
COMET-20	0.635	0.034	0.018	0.084	0.014	-0.002	0.009
metricx_xxl_DA_2019	0.635	0.040	0.019	0.086	0.015	0.005	0.008
UniTE-ref	0.629	0.032	0.018	0.084	0.009	0.004	0.005
MS-COMET-22	0.626	0.030	0.013	0.081	0.007	-0.000	0.004
COMET-22	0.626	0.031	0.019	0.079	0.013	0.002	0.002
metricx_xl_DA_2019	0.623	0.036	0.016	0.085	0.014	0.002	0.007
Cross-QE*	0.623	0.015	0.011	0.087	0.003	0.001	-0.000
UniTE	0.623	0.036	0.019	0.084	0.012	0.004	0.006
UniTE-src*	0.623	0.026	0.018	0.087	0.001	0.003	0.007
metricx_xl_MQM_2020	0.620	0.025	0.013	0.079	0.010	0.004	-0.002
metricx_xxl_MQM_2020	0.620	0.026	0.014	0.079	0.011	0.002	-0.003
COMETKiwi*	0.617	0.028	0.011	0.091	0.001	0.004	0.002
COMET-QE*	0.605	0.010	0.020	0.076	-0.005	-0.002	0.003
REUSE*	0.584	0.002	0.009	0.091	-0.007	0.000	0.011
HWTSC-TLM*	0.578	0.030	0.011	0.097	0.013	0.001	0.013
HWTSC-Teacher-Sim*	0.568	0.018	0.016	0.098	0.007	0.007	0.001
KG-BERTScore*	0.568	0.010	0.007	0.087	-0.012	0.008	-0.002
MEE	–	–	–	–	–	–	0.020
MEE2	–	–	–	–	–	–	0.021
MEE4	–	–	–	–	–	–	0.021
SEScore	–	–	–	–	–	–	0.013
MATESE	–	–	–	–	–	–	-0.009
MATESE-QE*	–	–	–	–	–	–	-0.006

Table 13: Segment-level Kendall-like correlation with crowdsourced Ref. DA scores. Rows are sorted by the system-level pairwise accuracy across all language pairs. Primary submissions are bolded, and baselines are underlined. Reference-free metrics are indicated using an asterisk.

The segment level Kendal-like correlations of all metrics with Ref. DA scores are all very close to zero, and these numbers are completely meaningless. We strongly recommend against using Ref. DA scores to evaluate MT metrics.

Task Incl. Human Translation	accuracy True	cs-uk False	en-cs False	en-de False	en-hr False	en-ja False	en-liv False	en-ru False	en-uk False	en-zh False	liv-en False	sah-ru False	uk-cs False	zh-en False
metricx_xl_MQM_2020	0.862	0.989	0.833	0.674	0.981	0.885	0.819	0.950	0.916	0.809	0.973	1.000	0.964	0.829
metricx_xxl_MQM_2020	0.862	0.989	0.853	0.713	0.961	0.885	0.913	0.963	0.939	0.838	0.907	1.000	0.961	0.807
metricx_xxl_DA_2019	0.860	0.984	0.861	0.748	0.972	0.913	0.936	0.954	0.943	0.841	0.994	1.000	0.944	0.887
metricx_xl_DA_2019	0.853	0.985	0.837	0.731	0.979	0.920	0.931	0.937	0.926	0.813	0.997	1.000	0.942	0.903
UniTE	0.849	0.990	0.837	0.514	0.985	0.923	0.905	0.930	0.919	0.811	0.999	1.000	0.948	0.890
COMET-22	0.842	0.987	0.831	0.593	0.939	0.902	0.922	0.922	0.913	0.765	0.994	1.000	0.959	0.893
UniTE-ref	0.840	0.989	0.849	0.523	0.978	0.918	0.896	0.937	0.921	0.814	1.000	1.000	0.944	0.877
Cross-QE*	0.835	0.976	0.792	0.614	0.966	0.904	-0.395	0.936	0.910	0.714	0.984	1.000	0.951	0.835
COMETKiwi*	0.835	0.976	0.832	0.525	0.931	0.923	0.616	0.867	0.902	0.761	0.993	1.000	0.962	0.876
MS-COMET-22	0.833	0.978	0.763	0.265	0.940	0.927	0.942	0.906	0.785	0.822	0.999	1.000	0.962	0.856
BLEURT-20	0.830	0.989	0.832	0.707	0.973	0.907	0.956	0.931	0.937	0.665	0.998	1.000	0.951	0.906
COMET-20	0.826	0.985	0.739	0.626	0.974	0.915	0.957	0.914	0.910	0.744	0.998	1.000	0.953	0.913
MS-COMET-QE-22*	0.824	0.965	0.682	-0.047	0.905	0.940	0.911	0.822	0.702	0.822	0.999	1.000	0.953	0.833
COMET-QE*	0.821	0.922	0.828	0.522	0.781	0.881	0.015	0.941	0.881	0.709	0.998	1.000	0.921	0.818
UniTE-src*	0.800	0.955	0.765	0.396	0.966	0.921	-0.225	0.913	0.892	0.752	0.993	1.000	0.957	0.829
YiSi-1	0.785	0.960	0.632	0.747	0.921	0.929	0.986	0.804	0.889	0.509	0.987	1.000	0.950	0.932
HWTSC_EE_BERTScore_0.8_With_Human	0.780	0.938	0.489	0.674	0.860	0.921	0.958	0.703	0.828	0.612	0.997	1.000	0.960	0.950
HWTSC_EE_BERTScore_0.8_Without_Human	0.777	0.937	0.511	0.669	0.843	0.923	0.957	0.703	0.828	0.584	0.997	1.000	0.960	0.944
HWTSC_EE_BERTScore_0.5_With_Human	0.771	0.943	0.361	0.625	0.867	0.893	0.960	0.686	0.826	0.734	0.990	1.000	0.935	0.949
HWTSC_EE_BERTScore_0.5_Without_Human	0.766	0.943	0.438	0.609	0.821	0.903	0.960	0.688	0.826	0.652	0.989	1.000	0.935	0.933
BERTScore	0.764	0.935	0.482	0.648	0.890	0.932	0.969	0.702	0.825	0.412	0.993	1.000	0.965	0.937
chrF	0.762	0.927	0.689	0.811	0.920	0.931	0.969	0.813	0.895	0.210	0.988	1.000	0.979	0.881
HWTSC_EE_BERTScore_0.3_With_Human	0.750	0.933	0.203	0.552	0.869	0.857	0.961	0.655	0.793	0.774	0.984	1.000	0.908	0.931
HWTSC-TLM*	0.748	0.880	0.811	0.001	0.574	0.837	0.428	0.821	0.578	0.667	0.947	1.000	0.913	0.307
f101spBLEU	0.748	0.883	0.567	0.690	0.901	0.866	0.991	0.698	0.825	0.197	0.983	1.000	0.976	0.886
f200spBLEU	0.748	0.888	0.549	0.656	0.906	0.862	0.989	0.690	0.814	0.208	0.979	1.000	0.974	0.891
HWTSC_EE_BERTScore_0.3_Without_Human	0.732	0.935	0.327	0.521	0.806	0.875	0.961	0.659	0.794	0.678	0.982	1.000	0.909	0.910
HWTSC-Teacher-Sim*	0.720	0.850	0.516	-0.072	0.839	0.842	0.706	0.580	0.499	0.591	0.900	1.000	0.941	0.363
BLEU	0.707	0.890	0.666	0.493	0.919	0.282	0.804	0.649	0.752	0.065	0.979	1.000	0.982	0.859
KG-BERTScore*	0.484	0.366	-0.845	-0.128	0.331	0.065	-0.734	0.250	-0.834	-0.123	0.458	-1.000	0.900	0.466
REUSE*	0.344	0.042	-0.923	-0.409	0.223	-0.096	-0.713	-0.859	-0.873	-0.202	0.644	-1.000	0.653	-0.158
MATESE	–	–	–	0.460	–	–	–	0.891	–	–	–	–	–	0.843
MEE	–	–	–	0.746	–	–	–	0.767	–	–	–	–	–	0.823
MEE2	–	–	–	0.774	–	–	–	0.765	–	–	–	–	–	0.876
MEE4	–	–	–	0.738	–	–	–	0.744	–	–	–	–	–	0.897
MATESE-QE*	–	–	–	0.151	–	–	–	0.876	–	–	–	–	–	0.846
SEScore	–	–	–	0.138	–	–	–	–	–	–	–	–	–	0.924

Table 14: System-level Pearson correlation with WMT source-based DA+SQM scores. Rows are sorted by the system-level pairwise accuracy across all language pairs. Primary submissions are bolded, and baselines are underlined. Reference-free metrics are indicated using an asterisk.

Task Incl. Human Translation	(sys) Accuracy True	cs-uk False	en-cs False	en-de False	en-hr False	en-ja False	en-liv False	en-ru False	en-uk False	en-zh False	liv-en False	sah-ru False	uk-cs False	zh-en False
metricx_xl_MQM_2020	0.862	0.276	0.273	0.133	0.289	0.218	0.006	0.288	0.285	0.141	0.147	0.384	0.266	0.260
metricx_xxl_MQM_2020	0.862	0.295	0.306	0.143	0.310	0.242	0.021	0.312	0.315	0.156	0.108	0.382	0.288	0.268
metricx_xxl_DA_2019	0.860	0.315	0.299	0.152	0.318	0.239	0.109	0.317	0.331	0.149	0.208	0.500	0.305	0.244
metricx_xl_DA_2019	0.853	0.306	0.299	0.143	0.306	0.237	0.126	0.308	0.322	0.148	0.218	0.503	0.296	0.251
UniTE	0.849	0.311	0.314	0.135	0.320	0.256	0.107	0.284	0.335	0.148	0.230	0.515	0.305	0.211
COMET-22	0.842	0.309	0.317	0.127	0.316	0.230	0.078	0.303	0.328	0.156	0.186	0.470	0.326	0.271
UniTE-ref	0.840	0.315	0.315	0.131	0.315	0.253	0.094	0.280	0.338	0.149	0.235	0.517	0.308	0.210
Cross-QE*	0.835	0.232	0.267	0.085	0.208	0.241	-0.075	0.225	0.232	0.137	0.137	0.300	0.238	0.254
COMETKiwi*	0.835	0.288	0.295	0.111	0.255	0.202	0.017	0.255	0.299	0.129	0.173	0.359	0.287	0.231
MS-COMET-22	0.833	0.276	0.298	0.114	0.292	0.235	0.108	0.281	0.307	0.141	0.214	0.445	0.253	0.238
BLEURT-20	0.830	0.292	0.291	0.140	0.274	0.221	0.091	0.283	0.317	0.133	0.227	0.497	0.276	0.218
COMET-20	0.826	0.280	0.279	0.133	0.298	0.244	0.135	0.280	0.297	0.141	0.237	0.488	0.270	0.214
MS-COMET-QE-22*	0.824	0.247	0.245	0.093	0.261	0.179	0.151	0.249	0.260	0.127	0.167	0.341	0.214	0.220
COMET-QE*	0.821	0.225	0.258	0.089	0.249	0.181	-0.050	0.240	0.265	0.119	0.125	0.235	0.238	0.232
UniTE-src*	0.800	0.283	0.301	0.116	0.293	0.253	-0.015	0.256	0.322	0.150	0.179	0.314	0.288	0.215
YiSi-1	0.785	0.223	0.173	0.084	0.225	0.212	0.120	0.191	0.211	0.076	0.212	0.439	0.203	0.168
HWTSC_EE_BERTScore_0.8_With_Human	0.780	–	–	–	–	–	–	–	–	–	–	–	–	–
HWTSC_EE_BERTScore_0.8_Without_Human	0.777	–	–	–	–	–	–	–	–	–	–	–	–	–
HWTSC_EE_BERTScore_0.5_With_Human	0.771	–	–	–	–	–	–	–	–	–	–	–	–	–
HWTSC_EE_BERTScore_0.5_Without_Human	0.766	–	–	–	–	–	–	–	–	–	–	–	–	–
BERTScore	0.764	0.200	0.165	0.091	0.215	0.177	0.119	0.173	0.177	0.072	0.217	0.438	0.190	0.188
chrF	0.762	0.195	0.147	0.085	0.185	0.142	0.101	0.153	0.177	0.051	0.184	0.430	0.171	0.071
HWTSC_EE_BERTScore_0.3_With_Human	0.750	–	–	–	–	–	–	–	–	–	–	–	–	–
HWTSC-TLM*	0.748	0.122	0.059	0.035	0.076	0.081	0.051	0.102	0.098	0.023	0.100	0.105	0.062	0.039
f101spBLEU	0.748	0.154	0.131	0.070	0.179	0.131	0.098	0.124	0.145	0.049	0.146	0.372	0.162	0.074
f200spBLEU	0.748	0.160	0.133	0.069	0.176	0.133	0.089	0.132	0.155	0.050	0.148	0.383	0.162	0.069
HWTSC_EE_BERTScore_0.3_Without_Human	0.732	–	–	–	–	–	–	–	–	–	–	–	–	–
HWTSC-Teacher-Sim*	0.720	0.116	0.115	0.049	0.148	0.119	0.032	0.104	0.177	0.056	0.068	0.054	0.090	0.168
BLEU	0.707	0.133	0.136	0.069	0.179	0.038	0.036	0.122	0.150	0.032	0.139	0.361	0.150	0.077
KG-BERTScore*	0.484	0.175	0.085	0.040	0.154	0.082	-0.098	0.087	0.121	0.054	0.016	0.013	0.137	0.149
REUSE*	0.344	0.155	0.049	0.031	0.079	0.107	-0.110	0.049	0.101	0.081	0.032	0.108	0.115	0.093
MATESE	–	–	–	0.106	–	–	–	0.198	–	–	–	–	–	0.225
MEE	–	–	–	0.071	–	–	–	0.118	–	–	–	–	–	0.078
MEE2	–	–	–	0.092	–	–	–	0.171	–	–	–	–	–	0.117
MEE4	–	–	–	0.091	–	–	–	0.181	–	–	–	–	–	0.104
MATESE-QE*	–	–	–	0.083	–	–	–	0.173	–	–	–	–	–	0.220
SEScore	–	–	–	0.091	–	–	–	–	–	–	–	–	–	0.213

Table 15: Segment-level Kendall-like correlation with DA+SQM scores. Rows are sorted by the system-level pairwise accuracy across all language pairs. Primary submissions are bolded, and baselines are underlined. Reference-free metrics are indicated using an asterisk.