Results of the WMT20 Metrics Shared Task

Nitika Mathur
The University of Melbourne
nmathur@student.unimelb.edu.au

Johnny Tian-Zheng Wei
University of Southern California,
jwei@umass.edu

Markus Freitag
Google Research
freitag@google.com

Qingsong Ma
Tencent-CSIG,
AI Evaluation Lab
qingsong.mqs@gmail.com

Ondřej Bojar
Charles University,
MFF ÚFAL
bojar@ufal.mff.cuni.cz

Abstract

This paper presents the results of the WMT20 Metrics Shared Task. Participants were asked to score the outputs of the translation systems competing in the WMT20 News Translation Task with automatic metrics. Ten research groups submitted 27 metrics, four of which are reference-less "metrics". In addition, we computed five baseline metrics, including SENTBLEU, BLEU, TER and CHRF using the SacreBLEU scorer. All metrics were evaluated on how well they correlate at the system-, document- and segment-level with the WMT20 official human scores.

We present an extensive analysis on influence of reference translations on metric reliability, how well automatic metrics score human translations, and we also flag major discrepancies between metric and human scores when evaluating MT systems. Finally, we investigate whether we can use automatic metrics to flag incorrect human ratings.

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 automatic metrics that score MT output by comparing them with a reference translation generated by human translators, who are instructed to translate "from scratch", without post-editing from MT. In addition, following last year's collaboration with the WMT Quality Estimation (QE) task, we also invited submissions of reference-free metrics that compare MT outputs directly with the source segment.

Similar to the last year's editions, the source, reference texts, and MT system outputs for the metric task come from the News Translation Task (Barrault et al., 2020, which we denote as Findings 2020). This year, the language pairs were English ↔ Chinese, Czech, German, Inuktitut, Japanese, Polish, Russian and Tamil. We further included systems participating in the WMT parallel corpus filtering task (Koehn et al., 2020): Khmer and Pashto to English.[2]

All metrics are evaluated based on their agreement with human evaluation. We evaluate metrics at three levels: comparing MT systems on the entire testset, segments (either sentences or short paragraphs), and new this year, documents. We introduce document-level evaluation to incentivize the development of metrics that are take into account broader context of evaluated sentences or paragraphs, following the recent emergence of document-level MT techniques.

Multiple References This year, we have two independently generated references for English ↔ German, English ↔ Russian, and Chinese → English. This lets us investigate the influence of references and the utility of multiple references. We instructed participants to score MT systems against the references individually as well as with all available references. In addition, we also supplied a set of references for English to German, that were generated by asking linguists to paraphrase the WMT reference as much as possible (Freitag et al., 2020). These references are designed to minimise translationese in the reference which could lead to metrics to be biased against systems that generate more natural text.

[1] http://www.statmt.org/wmt20/metrics-task.html

[2] Note that the metrics task inputs also included MT systems translating between German ↔ French in the News Translation Task, and English → Khmer and Pashto from the WMT parallel corpus filtering task. We are unable to evaluate metrics on these language pairs as human evaluation is not available

Proceedings of the 5th Conference on Machine Translation (WMT), pages 688–725
Online, November 19–20, 2020. ©2020 Association for Computational Linguistics

Evaluating Human Translations Given that we have multiple human translations, we asked participants to evaluate each human translation using the other as a reference. For these language-pairs, at least one of these human translations was included in the human evaluation, so we can directly evaluate metrics on how they rank the human translation compared to the MT systems.

Additional Human Evaluation Finally, we pose the question if some of the discrepancies between metrics and human scores can be explained by bad human ratings. We rerun some of the human evaluations by using the same template, but switching the rater pool from non-experts to professional linguists. In particular, we rerun human evaluation for a subset of translations where all metrics disagree with the WMT human evaluation. This experiment could reveal a new use case of automatic metrics and indicate that automatic metrics can be used to identify bad ratings in human evaluations.

We first give an overview of the task (Section 2) and summarize the baseline (Section 3.1) and submitted (Section 3.2) metrics. The results for system-, segment-, and document-level evaluation are provided in Sections 4, followed by a joint discussion Section 5. Section 6 describes our rerunning of human evaluation with linguists before we summarise our findings in Section 7.

We will release data, code and additional visualisations in the metrics package to be made available at http://www.statmt.org/wmt20/results.html

2 Task Setup

This year, we provided task participants with one test set for each examined language pair, i.e. a set of source texts (which are commonly ignored by MT metrics), corresponding MT outputs (these are the key inputs to be scored) and one or more reference translations.

In the system-level, metrics aim to correlate with a system's score which is an average over many human judgments of segment translation quality produced by the given system. In the segment-level, metrics aim to produce scores that correlate best with a human ranking judgment of two output translations for a given source segment. And finally, we also trial document-level evaluation this year. (more on the manual quality assessment in Section 2.3).

Segments are sentences for all language pairs except English ↔ German and Czech, and for English → Chinese, which do not contain sentence boundaries and are translated and evaluated at the paragraph-level.

Participants were free to choose which language pairs and tracks (system/segment/document and reference-based/reference-free) they wanted to take part in.

2.1 Source and Reference Texts

The source and reference texts we use are mainly sourced from this year's WMT News Translation Task (see Findings 2020).

The test set typically contains somewhere between 1000 and 2000 segments for each translation direction, with fewer segments for some paragraph-segmented test sets, and the English ↔ Inuktitut directions contain 2971 sentences.

All test sets are from the news domain, except the English ↔ Inuktitut datasets which have a mix of in-domain text from Canadian Parliament Hansards (1566 sentences) and out-of-domain news documents (1405 sentences).

We also have systems from the parallel corpus filtering task which are from the Wikipedia domain (also labelled *newstest2020* in the metrics test set). The Khmer → English and Pashto → English contain 2320 and 2719 sentences respectively.

The reference translations provided in *newstest2020* were created in the same direction as the MT systems were translating. The exceptions are English ↔ Inuktitut, Khmer → English and Pashto → English, where the testset is a mixture of "source-original" and "target-original" texts.

2.2 System Outputs

The results of the Metrics Task are affected by the actual set of MT systems participating in a given translation direction. On one hand, if all systems are very close in their translation quality, then even humans will struggle to rank them. This in turn will make the task for MT metrics very hard. On the other hand, if the task includes a wide range of systems of varying quality, correlating with humans should be generally easier. One can also expect that if the evaluated systems are of different types, they will exhibit different error patterns and various MT metrics can be differently sensitive to these patterns.

- **Parallel Corpus Filtering Task.** This task required participants to submit scores for each sentence in the provided noisy parallel texts. These scores were used to subsample sentence pairs, which was then used to train a neural machine translation system (fairseq). This was tested on a held-out subset of Wikipedia translations.

- **Regular News Tasks Systems.** These are all the other MT systems in the evaluation; differing in whether they are trained only on WMT provided data ("Constrained", or "Unconstrained") as in the previous years.

With all language pairs, in addition to the submissions to the task, the test sets also include translations from freely available web services (online MT systems), which are deemed unconstrained.

Overall, the results are based on 208 systems across 18 language pairs.

2.3 Manual Quality Assessment

Human scores were obtained using Direct Assessment, where annotators are asked to rate the adequacy of a translation compared to either the source segment or a reference translation of the same source. This year, human data was collected from reference-based evaluations (or "monolingual") and reference-free evaluations (or "bilingual"). The reference-based (monolingual) evaluations were crowdsourced, while the reference-less (bilingual) evaluations were mainly from MT researchers who committed their time to contribute to the manual evaluation for each submitted system to the translation task.

Finally, following reports that MT system translations might seem adequate when scored in isolation but not in context of the whole document, when possible, the ratings are collected for each segment with document context. Table 1 summarises the details of how human annotations were collected for various language-pairs at WMT 2020.

The English → Inuktitut dataset, which contains a mix of in-domain (Hansard) and out-of-domain (news) data, was only evaluated on out-of-domain segments, so for system level evaluation, we evaluate metric scores computed on the news domain only as well as the full test set.

See Findings 2020 for details on human evaluation.

2.3.1 System-level Golden Truth: DA

For the system-level evaluation, the collected continuous DA scores, standardized for each annotator, are averaged across all assessed segments for each MT system to produce a scalar rating for the system's performance.

The underlying set of assessed segments is different for each system. Thanks to the fact that the system-level DA score is an average over many judgments, mean scores are consistent and have been found to be reproducible (Graham et al., 2013). For more details see Findings 2020.

The score of an MT system is calculated as the average rating of the segments translated by the system.

2.3.2 Segment-level Golden Truth: DARR

Starting from Bojar et al. (2017), when WMT fully switched to DA, we had to come up with a solid golden standard for segment-level judgements. Standard DA scores are reliable only when averaged over sufficient number of judgments.[3]

Fortunately, when we have at least two DA scores for translations of the same source input, it is possible to convert those DA scores into a relative ranking judgement, if the difference in DA scores allows conclusion that one translation is better than the other. In the following, we denote these re-interpreted DA judgements as "DARR", to distinguish it clearly from the relative ranking ("RR") golden truth used in the past years.[4]

From the complete set of human assessments collected for the News Translation Task, all possible pairs of DA judgements attributed to distinct translations of the same source segment were converted into DARR better/worse judgements. Distinct translations of the same source input whose DA scores fell within 25 percentage points (which could have

[3] For segment-level evaluation, one would need to collect many manual evaluations of the exact same segment as produced by each MT system. Such a sampling would be however wasteful for the evaluation needed by WMT, so only some MT systems happen to be evaluated for a given input segment. In principle, we would like to return to DA's standard segment-level evaluation in future, where a minimum of 15 human judgements of translation quality are collected per translation and combined to get highly accurate scores for translations, but this would increase annotation costs.

[4] Since the analogue rating scale employed by DA is marked at the 0-25-50-75-100 points, we use 25 points as the minimum required difference between two system scores to produce DARR judgements. Note that we rely on judgements collected from known-reliable volunteers and crowd-sourced workers who passed DA's quality control mechanism. Any inconsistency that could arise from reliance on DA judgements collected from low quality crowd-sourcing is thus prevented.

Language pairs	source/reference	crowd/researcher	document context
iu-en	reference	crowd	No
*-en except iu-en	reference	crowd	Yes
en-*, de-fr, fr-de	source	mix of crowd and researcher*	Yes

Table 1: Direct Assessment at WMT20. Note that researcher annotations can contain some amount of professional annotations

been deemed equal quality) were omitted from the evaluation of segment-level metrics. Conversion of scores in this way produced a large set of DARR judgements for all language pairs, shown in Table 2 due to combinatorial advantage of extracting DARR judgements from all possible pairs of translations of the same source input. We see that only km-en and ps-en can suffer from insufficient number of these simulated pairwise comparisons.

The DARR judgements serve as the golden standard for segment-level evaluation in WMT19.

2.3.3 Document-level Golden Truth: DARR

As segments were scored in document context, we can compute document scores as the average human rating of the segments in the document. We acknowledge that this may be an oversimplification. First of all, we are hoping that human assessors have indicated errors in document-level coherence at at least one of the affected segments, but we have no evidence that they actually do so. Second, document-level phenomena are rather scarce and averaging segment-level scores is likely to average out these sparse observations even if they were marked at individual sentences. And lastly, in some situations, lack of cross-sentence coherence can be so critical that any strategy of composing sentence-level scores is bound to downplay the severity of the error, see e.g. Vojtěchová et al. (2019). At the current point, we have nothing better to start with but we believe that better techniques will be proposed in the future.

Graham et al. (2017) recommend around averaging 100 annotations per document to obtain reliable document scores. Since the average number of assessments we have is much less than that, we compute the ground truth in the same way as the segment level evaluation.

We first compute document scores as the average of all segment scores in the document, which we denote as DOC-DA. We then generate DOC-DARR pairs of better and worse translations of the same source document when there is at least a 25 point difference in the raw DOC-DA scores. See Table 3 for details.

In case of DARR (which we denote as DOC-DARR), all language pairs suffer from insufficient number of these simulated pairwise comparisons.

Similar to segment-level evaluation, we use the Kendall Tau-like formula (Section 2) to evaluate metric agreement with humans on the generated pairwise DARR judgements.

Note that we do not include any human-translated segments in this evaluation. In addition, iu-en is excluded from document-level evaluation because its DA judgements were collected for isolated sentences.

3 Metrics

3.1 Baselines

We agree with the call to use SacreBLEU (Post, 2018) as the standard MT evaluation scorer. We no longer report scores of the metrics from the Moses scorer, which requires tokenized text. We use the following metrics from the SacreBLEU scorer as baselines, with the default parameters:

3.1.1 SacreBLEU baselines

- BLEU (Papineni et al., 2002a) is the precision of n-grams of the MT output compared to the reference, weighted by a brevity penalty to punish overly short translations. `BLEU+case.mixed+lang.LANGPAIR-+numrefs.1+smooth.exp+tok.13a-+version.1.4.14`

  We run SacreBLEU with the `--sentence-score` option to obtain sentence scores for SENTBLEU; this uses the same parameters as BLEU. Although not it's intended use, we also compute system- and document-level scores for SENTBLEU as the mean segment score.

- TER (Snover et al., 2006) measures the number of edits (insertions, deletions, shifts and substitutions) required

	DA>1	Ave	DA pairs	daRR
cs-en	664	11.3	39187	14018
de-en	785	11.0	43669	16584
iu-en	2620	4.5	26120	8162
ja-en	993	9.0	36169	15193
pl-en	1001	11.8	64670	21121
ru-en	991	10.0	44664	14024
ta-en	997	7.6	26662	12789
zh-en	2000	13.8	177492	62586
km-en	1963	3.2	8295	3706
ps-en	2204	3.1	7994	3507
en-cs	1418	10.3	68587	21121
en-de	1418	6.9	30567	9339
en-iu	1268	7.9	35384	13159
en-ja	1000	9.6	41576	12830
en-pl	1000	10.6	52003	17689
en-ru	1971	5.7	28274	8330
en-ta	1000	7.9	28974	9087
en-zh	1418	10.6	72581	12652

Table 2: Segment-level: Number of judgements for DA converted to daRR data; "DA>1" is the number of source input segments in the manual evaluation where at least two translations of that same source input segment received a DA judgement; "Ave" is the average number of translations with at least one DA judgement available for the same source input segment; "DA pairs" is the number of all possible pairs of translations of the same source input resulting from "DA>1"; and "daRR" is the number of DA pairs with an absolute difference in DA scores greater than the 25 percentage point margin.

	doc-DA>1	Ave	doc-DA pairs	doc-daRR
cs-en	102	11.4	6041	1424
de-en	118	11.0	6579	1866
ja-en	80	8.9	2850	790
pl-en	62	11.8	4012	635
ru-en	91	9.9	4077	753
ta-en	82	7.5	2126	684
zh-en	155	13.8	13897	3085
en-cs	130	10.2	6162	1442
en-de	130	6.9	2844	669
en-iu	35	7.8	969	203
en-ja	63	9.7	2686	469
en-pl	63	10.7	3359	677
en-ru	122	5.7	1768	387
en-ta	63	7.9	1834	389
en-zh	130	10.6	6667	651

Table 3: Document-level: Number of judgements for doc-DA converted to doc-daRR data; "doc-DA>1" is the number of source input documents in the manual evaluation where we have doc-DA scores for at least two translations of that same source input documents; "Ave" is the average number of translations with at least one doc-DA judgement available for the same source input document; "doc-DA pairs" is the number of all possible pairs of translations of the same source input resulting from "doc-DA>1"; and "doc-daRR" is the number of doc-DA pairs with an absolute difference in doc-DA scores greater than the 25 percentage point margin.
Note that iu-en is not included as document-context was not available for this evaluation.

to transform the MT output to the reference. `TER+lang.LANGPAIR-+tok.tercom-nonorm-punct-noasian-uncased+version.1.4.14`

- chrF (Popović, 2015) uses character n-grams instead of word n-grams to compare the MT output with the reference [5]. Version string: `chrF2+lang.LANGPAIR-+numchars.6+space.false-+version.1.4.14`.

3.1.2 chrF++

chrF++ (Popović, 2017) includes word unigrams and bigrams in addition to character ngrams. We ran the original Python implementation of the metric [6] with the default parameters `--ncorder 6 --nwworder 2 --beta 2`

3.2 Submissions

The rest of this section summarizes participating metrics.

3.2.1 BERT-base-L2, BERT-large-L2, mBERT-L2

The three baselines were obtained by fine-tuning BERT (Devlin et al., 2019) on the ratings of WMT Metrics years 2015 to 2018, using a regression loss. What distinguishes the metrics is the initial BERT checkpoint: BERT-base-L2 uses a 12-layer Transformer architecture pre-trained on English data, mBERT-L2 is similar but trained

[5] Note that the SacreBLEU scorer does not yet implement chrF with multiple references

[6] `chrF++.py` available at https://github.com/m-popovic/chrF

	metric	features	Learned	Scoring level seg	 doc	 sys	Citation/Participant	Availability
Baselines	sentBLEU	n-grams		●	⋆	⋆	Papineni et al. (2002a)	https://github.com/mjpost/sacrebleu
	BLEU	n-grams		–	–	●	Papineni et al. (2002a)	https://github.com/mjpost/sacrebleu
	TER	edit distance		●	⊘	●	Snover et al. (2006)	https://github.com/mjpost/sacrebleu
	chrF	character n-grams		●	⊘	●	Popović (2015)	https://github.com/mjpost/sacrebleu
	chrF++	character n-grams		●	⊘	●	Popović (2017)	https://github.com/m-popovic/chrF
Reference-based metrics	parBLEU	paraphrases		●	⋆	●	Univ of Edinburgh, Univ of Tartu, JHU Bawden et al. (2020)	not a public metric
	parchrF++	paraphrases		●	⋆	●	Univ of Edinburgh, Univ of Tartu, JHU Bawden et al. (2020)	not a public metric
	parESIM	paraphrases	yes	●	⊘	⊘	Univ of Edinburgh, Univ of Tartu, JHU Bawden et al. (2020)	not a public metric
	prism	paraphrases		●	⊘	⊘	Johns Hopkins University	https://github.com/thompsonb/prism
	CharacTER	character edit distance		●	⊘	⊘	RWTH Aachen Wang et al. (2016)	https://github.com/rwth-i6/CharacTER
	EED	character edit distance		●	⊘	⊘	RWTH Aachen Stanchev et al. (2019)	https://github.com/rwth-i6/ExtendedEditDistance
	SWSS+METEOR	semantic similarity		●	⊘	⊘	, Xu et al. (2020)	not a public metric
	MEE	word embeddings		●	⊘	⊘	IIIT - Hyderabad, Ananya Mukherjee and Sharma (2020)	not a public metric
	YiSi	contextual word embeddings		●	⊘	⊘	NRC Lo (2019, 2020)	http://chikiu-jackie-lo.org/home/index.php/yisi
	BERT-base-L2	contextual word embeddings	yes	●	⊘	⊘	Google (Devlin et al., 2019)	(BLEURT code, private checkpoint)
	BERT-large-L2,	contextual word embeddings	yes	●	⊘	⊘	Google (Devlin et al., 2019)	(BLEURT code, private checkpoint)
	mBERT-L2	contextual word embeddings	yes	●	⊘	⊘	Google (Devlin et al., 2019)	(BLEURT code, private checkpoint)
	BLEURT	contextual word embeddings	yes	●	⊘	⊘	Google (Devlin et al., 2019)	https://github.com/google-research/bleurt
	BLEURT-extended	contextual word embeddings	yes	●	⊘	⊘	Google (Devlin et al., 2019)	(BLEURT code, private checkpoint)
	YiSi-combii	contextual word embeddings	yes	●	⊘	⊘	Google (Devlin et al., 2019)	not a public metric
	bleurt-combi	contextual word embeddings	yes	●	⊘	⊘	Google (Devlin et al., 2019)	not a public metric
	COMET	predictor-estimator model	yes	●	●	⊘	Unbabel (Rei et al., 2020b)	https://github.com/Unbabel/COMET
	COMET-Rank	predictor-estimator model	yes	●	⊘	⊘	Unbabel (Rei et al., 2020b)	https://github.com/Unbabel/COMET
	COMET-HTER	predictor-estimator model	yes	●	●	⊘	Unbabel (Rei et al., 2020b)	https://github.com/Unbabel/COMET
	COMET-2R	predictor-estimator model	yes	●	⊘	⊘	Unbabel (Rei et al., 2020b)	https://github.com/Unbabel/COMET
	COMET-MQM	predictor-estimator model	yes	●	●	⊘	Unbabel (Rei et al., 2020b)	https://github.com/Unbabel/COMET
	BAQ, EQ	?	?	●	⊘	⊘	?	not a public metric
src-based	COMET-QE	predictor-estimator model	yes	●	⊘	⊘	Unbabel (Rei et al., 2020b)	https://github.com/Unbabel/COMET
	OpenKiwi-Bert	predictor-estimator model	yes	●	⊘	⊘	Unbabel Kepler et al. (2019)	https://github.com/Unbabel/OpenKiwi
	OpenKiwi-XLMR	predictor-estimator model	yes	●	⊘	⊘	Unbabel Kepler et al. (2019)	https://github.com/Unbabel/OpenKiwi
	YiSi-2	contextual word embeddings		●	⊘	⊘	NRC Lo and Larkin (2020)	

Table 4: Participants of WMT20 Metrics Shared Task. "●" denotes that the metric took part in (some of the language pairs) of the segment- and/or document- and/or system-level evaluation. "⊘" indicates that the document- and system-level scores are implied, simply taking arithmetic (macro-)average of segment-level scores. "–" indicates that the metric didn't participate the track (Seg/Doc/Sys-level). "⋆" indicates that we computed the metric's document or system score for this track as the macro-average of segment scores, though the metric is not defined this way. A metric is learned if it is trained on a QE or metric evaluation dataset (i.e. pretraining or parsers don't count, but training on WMT 2019 metrics task data does).

on Wikipedia data in 102 languages, and BERT-LARGE-L2 is English-only with 24 layers.

3.2.2 BLEURT, BLEURT-EXTENDED, YISI-COMBI, BLEURT-YISI-COMBI

BLEURT (Sellam et al., 2020a) is a BERT-based regression model trained twice: first on million synthetic pairs obtained by random perturbations, then on ratings from years 2015 to 2019 of the WMT Workshop. BLEURT-EXTENDED (Sellam et al., 2020b) is a BERT-based regression model trained on human ratings of years 2015 to 2019 of the WMT Workshop, combined with BERT-Chinese for to-Chinese sentence pairs. The main checkpoint is a 24-layer Transformer, trained on a mixture of Wikipedia articles and training data from WMT Newstest in 20 languages.

YISI-COMBI: We are using YISI-1 on an mBERT model that is fine tuned on WMT data for single reference submissions. We are using aggregating internal scores in YISI over different references for the final output for multi reference submission.

BLEURT-COMBI: We are using the same output as YISI-COMBI for single reference submissions. We are mixing YISI-1, YISI-2 and BLEURT scores for different references for the multi reference submission.

3.2.3 CHARACTER

CHARACTER (Wang et al., 2016), identical to the 2016 setup, is a character-level metric inspired by the commonly applied translation edit rate (TER). It is defined as the minimum number of character edits required to adjust a hypothesis, until it completely matches the reference, normalized by the length of the hypothesis sentence. CHARACTER calculates the character-level edit distance while performing the shift edit on word level. Unlike the strict matching criterion in TER, a hypothesis word is considered to match a reference word and could be shifted, if the edit distance between them is below a threshold value. The Levenshtein distance between the reference and the shifted hypothesis sequence is computed on the character level. In addition, the lengths of hypothesis sequences instead of reference sequences are used for normalizing the edit distance, which effectively counters the issue that shorter translations normally achieve lower TER. Similarly to other character-level metrics, CHARACTER is generally applied to nontokenized outputs and references, which also holds for this year's submission with one exception. This year tokenization was carried out for en-ru hypotheses and references before calculating the scores, since this results in large improvements in terms of correlations. For other language pairs, no tokenizer was used for pre-processing.

3.2.4 COMET

COMET* metrics (Rei et al., 2020b) were build using the Estimator model or the Translation Ranking model proposed in Rei et al. (2020a). Those neural models use XLM-RoBERTa to encode source, MT hypothesis and reference in the same cross-lingual space and then are optimised towards different objectives. COMET (main metric) is an Estimator model that regresses on Direct Assessments (DA) from 2017 to 2019 and COMET-2R is a variant of COMET (main metric) that was trained to handle multiple references at inference time. COMET-HTER and COMET-MQM follow the same architecture but regress on Human-mediated Translation Edit Rate (HTER) and a proprietary metric compliant with the Multidimensional Quality Metrics framework (MQM), respectively. COMET-Rank uses the Translation Ranking architecture to directly optimize the distance between "better" hypothesis and the respective source and reference, while pushing the "worse" hypothesis away. This Translation Ranking model was directly optimised on DA relative-ranks from 2017 to 2019. Finally, COMET-QE removes the reference at input and proportionately reduces the dimensions of the estimator network to accommodate the reduced input.

3.2.5 EED

EED (Stanchev et al., 2019) is a character-based metric, which builds upon CDER. It is defined as the minimum number of operations of an extension to the conventional edit distance containing a "jump" operation. The edit distance operations (insertions, deletions and substitutions) are performed at the character level and jumps are performed when a blank space is reached. Furthermore, the coverage of multiple characters in the hypothesis is penalised by the introduction of a coverage penalty. The sum of the length of the reference and the coverage penalty is used as the normalisation term.

3.2.6 MEE

MEE (Ananya Mukherjee and Sharma, 2020) is an automatic evaluation metric that leverages the similarity between embeddings of words in candidate and reference sentences to assess translation quality. Unigrams are matched based on their surface forms, root forms and meanings which aids to capture lexical, morphological and semantic equivalence. Semantic evaluation is achieved by using pretrained fasttext embeddings provided by Facebook to calculate the word similarity score between the candidate and the reference words. MEE computes evaluation score using three modules namely exact match, root match and synonym match. In each module, fmean-score is calculated using harmonic mean of precision and recall by assigning more weight to recall. The final translation score is obtained by taking average of fmean-scores from individual modules.

3.2.7 ESIM

Enhanced Sequential Inference Model (Chen et al., 2017) is a neural model proposed for Natural Language Inference that has been adapted for MT evaluation by Mathur et al. (2019). It uses cross-sentence attention and sentence matching heuristics to generate a representation of the translation and the reference, which is fed to a feedforward regressor. This year's scores were submitted by Bawden et al. (2020) as part of the submission on PARESIM.

3.3 OPENKIWI-BERT, OPENKIWI-XLMR

OPENKIWI-BERT and OPENKIWI-XLMR (Kepler et al., 2019) are state of the art Quality Estimation models developed for the WMT20 QE shared task and are trained with WMT Metrics data from 2017 to 2019.

3.3.1 PARBLEU, PARCHRF++, PARESIM

PARBLEU, PARCHRF++, and PARESIM (Bawden et al., 2020) are variants of their respective core metrics computed against the provided human reference and a set of automatically generated paraphrases. PARBLEU used five paraphrases, while the other two used only one. Both BLEU and CHRF++ have in-built support for multiple references. For ESIM, we calculate the score for each reference separately and then average them to get the final score.

3.3.2 PRISM

PRISM (Thompson and Post, 2020) is a many-many multilingual neural machine translation system trained on data for 39 language pairs, with data derived largely from WMT and Wikimatrix. It casts machine translation evaluation as a zero-shot paraphrasing task, producing segment-level scores by force-decoding between a system output and a reference, in both directions, and averaging the model scores. System-level scores are produced by averaging segment-level ones. For evaluation in Inuktikut, Khmer, Pashto, and Tamil, we used a "Prism44" model that was retrained after adding WMT-provided data for these languages to its original training data set. All other languages were evaluated with the original "Prism39" model.

3.3.3 SWSS+METEOR

SWSS (Semantically Weighted Sentence Similarity, Xu et al. 2020) is an approach to extracting semantic core words, which are words that carry important semantic meanings in sentences, and using them in MT evaluation. It uses UCCA (Universal Conceptual Cognitive Annotation), a semantic representation framework, to identify semantic core words, and then calculates sentence similarity scores on the overlap of semantic core words of sentence pairs. Taking sentence-level semantic structure information into consideration, SWSS can improve the performance of lexical metrics when combined with them. The submitted metric (SWSS+METEOR) is a weighted combination of SWSS and Meteor.

3.3.4 YISI-0, YISI-1, YISI-2

YISI (Lo, 2019, 2020) is a unified semantic MT quality evaluation and estimation metric for languages with different levels of available re-sources. YISI-1 is a reference-based MT evaluation metric that measures the semantic similarity between a ma-chine translation and human references by aggregating the idf-weighted lexical semantic similarities based on the contextual embeddings extracted from pretrained language models (BERT, CamemBERT, RoBERTa, XLM, XLM-RoBERTa, etc.) and optionally incorporating shallow semantic structures (denoted as YISI-1_SRL; not participating this year). YISI-0 is the degenerate version of YISI-1 that is ready-to-deploy to any language. It uses longest common character substring to measure the lexical similarity. YISI-2 (Lo and Larkin, 2020) is the bilingual, reference-less version for MT quality estimation, which uses bilingual mappings of the contextual embeddings extracted from pretrained language models (XLM or XLM-RoBERTa) to evaluate the crosslingual lexical semantic similarity between the input and

MT output. Like YISI-1, YISI-2 can exploit shallow semantic structures as well (denoted as YISI-2_SRL; does not participate this year).

3.4 Pre-processing

Since some metrics, such as BLEU, aim to achieve a strong positive correlation with human assessment, while error metrics, such as TER, aim for a strong negative correlation, in previous years we compare metrics via the absolute value $|r|$ of a given metric's correlation with human assessment. However, this can mask instances of true negative correlation for metrics that aim for a positive correlation (and vice-versa).

For system, document and segment level scores, we reverse the sign of the score of error metrics prior to the comparison with human scores, whether on the system, document or segment level: higher scores have to indicate better translation quality.

4 Results

4.1 System-Level Evaluation

As in previous years, we employ the Pearson correlation (r) as the main evaluation measure for system-level metrics. The Pearson correlation is as follows:

$$r = \frac{\sum_{i=1}^{n}(H_i - \overline{H})(M_i - \overline{M})}{\sqrt{\sum_{i=1}^{n}(H_i - \overline{H})^2}\sqrt{\sum_{i=1}^{n}(M_i - \overline{M})^2}} \quad (1)$$

where H_i are human assessment scores of all systems in a given translation direction, M_i are the corresponding scores as predicted by a given metric. $\overline{H}$ and $\overline{M}$ are their means, respectively.

As recommended by Graham and Baldwin (2014), we employ Williams significance test (Williams, 1959) to identify differences in correlation that are statistically significant. Williams test is a test of significance of a difference in dependent correlations and therefore suitable for evaluation of metrics. Correlations not significantly outperformed by any other metric for the given language pair are highlighted in bold in all the results tables that show Pearson correlation of metric and human scores.

Pearson correlation is ideal for reporting whether metric scores have the same trend as human scores. In practice, we use metrics to make decisions comparing MT systems, and Kendall's Tau appears to be more close to this use case, as it directly checks whether the metric ordering of a pair of MT systems agrees with the human ordering. However, unlike Pearson correlation, it is not sensitive to whether the metric score differences correspond to the human score differences. We stay with Pearson correlation for the official results, but also report Kendall's Tau correlation in the appendix.

The calculation of Pearson correlation coefficient is dependent on the mean, which is very sensitive to outliers. So if we have systems whose scores are far away from the rest of the systems, the presence of these "outlier" systems can give a misleadingly high impression of the correlations, and potentially change ranking of metrics. To avoid this, we also report correlations over non-outlier systems only.

To remove outliers, we are guided by the robust outlier detection method proposed for MT metric evaluation by Mathur et al. (2020). This method, recommended by the statistics literature (Iglewicz and Hoaglin, 1993; Rousseeuw and Hubert, 2011; Leys et al., 2013) depends on the median and the median absolute deviation (MAD) which is the median of the absolute difference between each point and the median. The method removes systems whose human scores are greater than 2.5 MAD away from the median.

The cutoff of 2.5 is subjective, and Leys et al. (2013) suggest the guidelines of using 3 (very conservative), 2.5 (moderately conservative) or 2 (poorly conservative), and recommends 2.5. For some language pairs, we override the 2.5 cutoff for systems that are close to the cutoff. We give examples in Section 5, and list all identified outliers in Table 15 in the Appendix.

4.1.1 System-Level Results

Tables 5 and 6 provide the system-level correlations of metrics. These tables include results for all MT systems, and in cases where we detect outliers, we also report correlation without outliers.

This year, we also carry out an extended analysis of the impact of (multiple) human references, see the following paragraphs.

Scoring Human Translation In this section, we investigate how well the metric submissions score human translations. We have five language pairs where two reference translations were provided by WMT. The manual DA scoring of News Translation Task included all the out-of-English human references in the evaluation along with the MT systems.

	cs-en		de-en		ja-en		pl-en		ru-en		ta-en		zh-en		iu-en		km-en	ps-en
	10		9		7		13		10		12		15		9		7	6
sentBLEU	0.844	**0.800**	0.978	**0.786**	**0.974**	**0.851**	**0.502**	**0.284**	**0.916**	**0.833**	0.925	**0.829**	0.948	**0.950**	0.649	**0.469**	0.969	**0.888**
BLEU	**0.851**	**0.800**	0.985	0.778	**0.969**	0.826	**0.549**	**0.355**	0.884	0.761	0.916	**0.807**	0.956	**0.957**	0.569	0.348	0.969	**0.888**
TER	**0.845**	**0.783**	0.993	**0.766**	**0.974**	0.752	**0.586**	**0.346**	**0.904**	**0.829**	0.805	**0.795**	0.956	0.911	**0.733**	**0.616**	**0.973**	**0.935**
chrF++	**0.867**	**0.804**	0.997	0.699	**0.974**	**0.871**	**0.538**	**0.328**	0.894	**0.833**	**0.953**	**0.830**	**0.975**	**0.955**	0.726	0.392	0.983	0.900
chrF	**0.872**	**0.806**	**0.997**	0.687	0.968	**0.861**	**0.528**	**0.312**	0.890	0.831	**0.951**	**0.828**	**0.976**	**0.954**	0.729	0.337	0.978	0.898
parBLEU	**0.834**	**0.774**	0.986	**0.838**	**0.970**	**0.833**	**0.562**	**0.342**	0.877	0.744	0.908	**0.801**	0.958	**0.953**	0.624	0.398	**0.971**	**0.939**
parCHRF++	**0.865**	**0.810**	**0.998**	0.708	**0.974**	**0.877**	**0.551**	**0.347**	0.885	0.823	**0.942**	**0.825**	**0.976**	**0.956**	0.720	0.296	**0.985**	0.899
CharacTER	0.844	**0.812**	**0.998**	0.687	**0.970**	**0.895**	**0.522**	**0.325**	**0.927**	**0.869**	**0.965**	**0.880**	0.964	**0.950**	**0.763**	**0.410**	**0.977**	0.841
EED	**0.884**	**0.838**	**0.997**	**0.752**	**0.974**	**0.904**	**0.538**	**0.299**	**0.926**	**0.872**	**0.958**	**0.862**	0.956	0.932	**0.821**	**0.587**	**0.990**	**0.930**
YiSi-0	**0.876**	**0.825**	**0.998**	**0.786**	**0.972**	**0.867**	0.453	**0.207**	**0.938**	**0.874**	**0.968**	**0.861**	0.956	0.918	**0.831**	**0.563**	**0.986**	**0.932**
SWSS+METEOR	–		–		**0.978**	**0.919**	**0.472**	**0.212**	**0.925**	**0.876**	**0.967**	**0.862**	0.959	0.926	**0.766**	**0.545**	**0.990**	**0.946**
MEE	**0.861**	**0.822**	0.995	0.712	**0.982**	**0.900**	**0.464**	**0.295**	**0.927**	**0.878**	**0.950**	**0.835**	0.952	**0.948**	**0.771**	**0.562**	0.970	0.878
prism	0.818	0.720	**0.998**	**0.775**	**0.974**	**0.869**	**0.502**	**0.269**	**0.908**	**0.839**	0.898	**0.788**	0.957	**0.945**	**0.833**	**0.616**	0.950	**0.966**
YiSi-1	**0.832**	**0.746**	**0.998**	**0.783**	**0.982**	**0.868**	**0.543**	**0.316**	**0.915**	**0.833**	0.925	**0.797**	0.961	**0.942**	**0.834**	**0.590**	**0.977**	**0.953**
BERT-base-L2	0.775	0.693	0.997	0.791	0.971	0.789	**0.552**	**0.328**	0.919	**0.836**	0.909	0.746	**0.967**	0.929	0.704	0.145	0.967	**0.945**
BERT-large-L2	**0.784**	0.695	0.990	0.800	**0.974**	**0.784**	0.520	0.282	**0.925**	**0.843**	0.901	0.760	0.962	0.928	0.744	0.211	0.959	**0.950**
mBERT-L2	**0.798**	0.715	0.995	**0.824**	**0.969**	**0.811**	**0.555**	**0.302**	**0.908**	0.805	0.887	0.740	0.959	**0.935**	**0.837**	**0.530**	**0.980**	**0.938**
BLEURT	**0.792**	0.725	0.996	0.770	**0.978**	**0.820**	**0.591**	**0.371**	**0.924**	**0.844**	0.906	0.768	**0.966**	**0.931**	0.771	0.320	**0.984**	**0.955**
BLEURT-extended	0.771	0.668	0.985	**0.818**	0.961	**0.772**	**0.551**	0.298	0.900	0.797	0.897	0.743	0.945	0.931	0.789	0.359	**0.985**	**0.942**
esim	**0.790**	0.716	**0.998**	0.808	**0.983**	**0.822**	**0.591**	**0.358**	**0.928**	**0.834**	0.885	**0.801**	**0.963**	0.910	**0.807**	**0.514**	0.929	0.929
parESIM-1	**0.788**	0.712	**0.998**	**0.835**	**0.983**	**0.819**	**0.591**	**0.363**	**0.926**	**0.828**	0.885	**0.797**	**0.963**	0.910	**0.800**	**0.495**	0.929	0.929
COMET	0.783	0.694	**0.998**	0.773	0.964	**0.828**	**0.591**	**0.345**	**0.923**	**0.836**	0.880	**0.764**	0.952	**0.931**	**0.852**	**0.605**	0.971	0.941
COMET-2R	0.777	0.697	**0.998**	0.772	0.964	**0.818**	**0.584**	**0.332**	**0.924**	**0.843**	0.881	**0.770**	0.949	**0.928**	**0.872**	**0.644**	0.970	**0.949**
COMET-HTER	0.738	0.661	0.995	**0.767**	0.912	**0.702**	0.446	0.231	0.867	0.741	0.726	0.595	0.809	0.873	**0.770**	**0.464**	0.901	0.862
COMET-MQM	0.728	0.612	0.991	**0.684**	0.906	0.707	0.424	0.222	0.858	0.746	0.767	0.617	0.784	0.862	**0.841**	**0.631**	0.914	0.880
COMET-Rank	0.705	0.534	0.964	**0.757**	0.923	**0.793**	0.483	**0.284**	0.868	0.732	0.787	0.664	0.877	**0.909**	0.158	0.214	0.911	0.855
BAQ_dyn	–		–		–		–		–		–		0.956	0.928	–		–	–
BAQ_static	–		–		–		–		–		–		**0.960**	**0.933**	–		–	–
COMET-QE	0.755	0.622	0.939	**0.805**	0.892	**0.585**	0.447	0.218	0.883	0.773	0.795	0.672	0.847	0.887	**0.685**	**0.661**	0.896	0.832
OpenKiwi-Bert	0.726	**0.698**	0.989	**0.741**	0.735	0.546	0.355	0.187	**0.862**	0.695	0.645	0.469	0.625	0.774	-0.126	-0.671	0.751	0.753
OpenKiwi-XLMR	0.760	**0.680**	0.995	0.701	0.931	0.714	0.442	0.171	0.859	0.697	0.792	0.659	0.905	0.899	0.271	-0.577	0.880	0.865
YiSi-2	0.764	0.640	0.988	0.404	**0.971**	**0.776**	**0.437**	**0.230**	**0.825**	**0.814**	0.849	**0.761**	**0.964**	0.933	**0.676**	**0.371**	0.790	**0.942**

Table 5: Pearson correlation of to-English system-level metrics with DA human assessment over MT systems using the *newstest2020* references. For language pairs that contain outlier systems, we also show correlation after removing outlier systems (“-out”). Correlations of metrics not significantly outperformed by any other for that language pair are highlighted in bold.

	en-cs		en-de		en-ja		en-pl		en-ru	en-ta		en-zh	en-iu$_{\text{FULL}}$		en-iu$_{\text{NEWS}\#}$	
	All	-out	All	-out	All	-out	All	-out	All	All	-out	All	All	-out	All	-out
	12	10	14	11	11	9	14	11	9	15	12	12	11	8	11	8
SENTBLEU	0.840	0.436	0.934	0.823	0.946	**0.976**	0.950	**0.772**	**0.981**	0.881	0.852	0.927	0.129	0.047	0.075	**0.172**
BLEU	0.825	0.390	0.928	0.825	0.945	**0.980**	0.943	**0.743**	**0.980**	0.880	0.829	0.928	0.163	0.131	0.074	**0.111**
TER	0.814	0.339	0.941	**0.848**	0.297	0.801	0.893	0.553	0.064	0.870	**0.883**	-0.213	0.384	**0.133**	0.357	**0.083**
CHRF++	0.833	0.349	0.958	**0.850**	0.952	0.945	**0.956**	**0.783**	**0.983**	0.929	**0.880**	0.878	0.328	0.128	0.315	**0.098**
CHRF	0.826	0.313	**0.962**	**0.862**	0.951	0.964	**0.957**	**0.793**	**0.982**	0.937	**0.890**	0.923	0.350	**0.122**	0.336	**0.091**
PARBLEU	0.870	0.543	0.910	0.774	0.869	0.813	0.948	**0.760**	**0.959**	0.871	**0.849**	0.962	0.194	**0.464**	0.126	**0.306**
PARCHRF++	0.860	0.438	0.957	**0.845**	0.955	0.951	**0.953**	**0.818**	**0.975**	–		0.948	–		–	
CHARACTER	0.807	0.269	**0.961**	**0.868**	0.951	**0.936**	0.935	**0.726**	0.961	0.957	0.851	0.905	0.503	0.008	0.515	**0.121**
EED	0.817	0.271	**0.965**	**0.869**	0.955	0.965	**0.962**	**0.789**	**0.980**	**0.959**	**0.913**	0.928	0.519	0.043	0.483	**0.122**
MEE	0.875	0.495	0.954	0.820	–		**0.952**	**0.733**	0.724	0.906	0.861	–	0.287	0.094	0.242	**0.113**
YISI-0	0.797	0.270	0.953	**0.889**	0.967	**0.972**	**0.953**	**0.783**	**0.971**	0.929	**0.897**	0.362	0.525	0.015	0.505	**0.095**
PRISM	0.949	0.805	0.958	**0.851**	0.932	0.921	**0.958**	**0.742**	0.724	0.863	0.452	0.221	**0.957**	**-0.043**	**0.945**	**0.088**
YISI-1	0.922	0.664	**0.971**	**0.887**	0.969	0.967	**0.964**	**0.714**	0.926	**0.973**	**0.909**	0.959	0.554	-0.217	0.523	-0.014
YISI-COMBI	–		0.971	0.868	–		–		–	–		–	–		–	
BLEURT-YISI-COMBI	–		0.971	0.868	–		–		–	–		–	–		–	
MBERT-L2	0.946	0.782	**0.970**	**0.861**	0.977	**0.969**	**0.976**	**0.775**	0.946	**0.944**	0.834	0.934	–		–	
BLEURT-EXTENDED	**0.989**	**0.960**	**0.969**	**0.870**	0.944	0.953	**0.982**	**0.828**	**0.980**	0.940	0.814	0.928	0.823	**0.122**	0.762	**0.155**
ESIM	0.908	0.575	**0.979**	**0.894**	**0.993**	**0.981**	**0.969**	0.698	**0.967**	0.937	0.833	0.972	0.814	**0.365**	0.760	0.418
PARESIM-1	0.919	0.635	**0.974**	**0.886**	**0.989**	**0.971**	**0.968**	0.705	**0.964**	0.937	0.833	**0.983**	0.814	**0.365**	0.760	0.418
COMET	0.978	0.926	**0.972**	**0.863**	0.974	**0.969**	**0.981**	**0.800**	0.925	0.944	0.798	0.007	0.860	0.028	**0.858**	0.152
COMET-2R	**0.983**	0.942	**0.972**	**0.869**	**0.986**	**0.978**	**0.982**	**0.803**	0.872	**0.959**	0.852	-0.066	0.848	-0.008	**0.867**	**0.177**
COMET-HTER	0.976	0.917	0.951	**0.852**	**0.989**	**0.974**	**0.974**	**0.763**	0.803	0.925	0.681	-0.073	**0.900**	0.142	**0.888**	0.092
COMET-MQM	0.974	0.910	0.881	0.840	0.974	**0.965**	0.967	**0.766**	0.788	0.910	0.641	0.084	0.870	0.129	**0.867**	0.172
COMET-RANK	0.959	0.868	0.877	**0.860**	0.931	0.928	0.957	**0.760**	0.676	0.876	0.511	0.540	0.283	0.099	0.392	**0.252**
BAQ_DYN	–		–		–		–		–	–		0.904	–		–	
BAQ_STATIC	–		–		–		–		–	–		**0.958**	–		–	
EQ_DYN	–		–		–		–		–	–		0.948	–		–	
EQ_STATIC	–		–		–		–		–	–		**0.976**	–		–	
COMET-QE	**0.989**	**0.974**	0.903	0.831	0.953	**0.955**	0.969	**0.804**	0.807	0.887	0.622	0.375	**0.905**	**0.578**	**0.928**	**0.651**
OPENKIWI-BERT	0.920	0.830	0.852	**0.829**	0.363	0.783	0.903	0.450	0.834	0.846	0.370	0.551	0.573	-0.602	0.808	**0.194**
OPENKIWI-XLMR	0.972	0.911	**0.968**	0.814	**0.992**	**0.976**	0.957	**0.638**	0.875	0.910	0.676	-0.010	0.513	-0.668	0.680	-0.358
YISI-2	0.714	0.353	0.899	0.552	0.854	0.646	0.470	-0.107	0.584	0.922	**0.923**	-0.215	0.802	**-0.257**	**0.830**	**0.065**

Table 6: Pearson correlation of out-of-English system-level metrics with DA human assessment over MT systems using the newstest2020 references; For language pairs that contain outlier systems, we also show correlation after removing outlier systems. Correlations of metrics not significantly outperformed by any other for that language pair are highlighted in bold.

# The English → Inuktitut human evaluation only contained the news subset, so we recompute en-iu system scores of metrics on the news subset of the testset (1405 sentences). Note that the scores of PARBLEU and PARCHRF were computed as average of segment scores

	en-de	en-de$_P$	en-de$_B$	en-de$_P$	en-zh	en-zh$_B$	de-en	ru-en	zh-en
HID	Human-B	Human-B	Human-A	Human-A	Human-B	Human-A	Human-B	Human-B	Human-B
N	12	12	12	12	13	13	10	11	16
sentBLEU	0.441	0.851	0.639	0.676	0.647	0.837	0.437	0.797	0.917
BLEU	0.458	**0.868**	0.672	0.665	0.658	0.814	0.480	0.738	**0.938**
TER	0.233	0.495	0.577	0.695	-0.131	-0.138	0.466	**0.812**	0.850
chrF++	0.555	**0.917**	0.748	0.650	0.592	0.805	0.437	0.815	**0.947**
chrF	0.599	**0.919**	0.772	0.645	0.651	0.812	0.442	0.821	**0.948**
parBLEU	0.349	0.676	0.580	0.682	0.569	0.787	0.498	0.716	0.926
parCHRF++	0.573	**0.890**	0.748	0.698	0.559	0.776	0.447	0.803	**0.950**
CharacTER	0.472	**0.890**	0.736	0.638	0.687	0.850	0.410	**0.856**	**0.938**
EED	0.447	**0.898**	0.685	0.646	0.679	0.830	0.466	**0.861**	0.910
YiSi-0	0.514	**0.892**	0.728	0.724	0.244	0.274	0.566	**0.860**	0.898
SWSS+METEOR	–	–	–	–	–	–	–	**0.866**	0.914
MEE	0.512	**0.886**	0.719	0.642	–	–	0.399	**0.855**	**0.941**
prism	0.472	0.727	0.731	0.742	0.157	0.166	0.591	**0.837**	**0.942**
YiSi-1	0.640	**0.895**	**0.830**	0.697	0.773	**0.916**	**0.713**	**0.822**	**0.943**
Yisi-combi	0.607	0.891	0.801	0.702	–	–	–	–	–
bleurt-Yisi-combi	0.607	0.891	0.801	0.702	–	–	–	–	–
BERT-base-L2	–	–	–	–	–	–	**0.785**	**0.813**	**0.922**
BERT-large-L2	–	–	–	–	–	–	**0.794**	**0.819**	**0.923**
mBERT-L2	**0.845**	0.876	**0.875**	**0.810**	0.868	**0.907**	**0.748**	0.789	**0.925**
BLEURT	–	–	–	–	–	–	0.754	**0.823**	**0.923**
BLEURT-extended	**0.888**	**0.896**	**0.883**	**0.838**	0.865	**0.910**	**0.811**	0.757	0.914
esim	0.719	**0.920**	**0.870**	**0.744**	0.837	**0.924**	0.765	**0.819**	**0.911**
paresim-1	0.687	**0.905**	**0.856**	**0.763**	0.822	**0.910**	**0.798**	**0.815**	**0.911**
COMET	0.854	**0.894**	**0.879**	**0.822**	0.078	0.062	**0.759**	**0.821**	**0.916**
COMET-2R	0.820	**0.866**	**0.877**	**0.865**	0.009	-0.003	**0.756**	**0.837**	**0.911**
COMET-HTER	0.840	0.871	**0.869**	**0.851**	0.006	-0.001	**0.761**	0.718	0.857
COMET-MQM	0.839	0.876	0.859	0.825	0.158	0.154	**0.682**	0.722	0.846
COMET-Rank	0.782	**0.870**	**0.830**	0.794	0.578	0.565	**0.709**	0.725	**0.896**
BAQ_dyn	–	–	–	–	0.739	–	–	–	0.236
BAQ_static	–	–	–	–	**0.915**	–	–	–	0.239
EQ_dyn	–	–	–	–	0.729	–	–	–	–
EQ_static	–	–	–	–	**0.925**	–	–	–	–
COMET-QE	**0.885**	**0.885**	**0.844**	**0.844**	0.473	0.481	**0.806**	0.749	0.865
OpenKiwi-Bert	0.741	0.741	**0.835**	**0.835**	0.487	0.521	**0.655**	0.682	0.742
OpenKiwi-XLMR	0.736	0.736	0.795	**0.795**	0.053	0.050	0.660	0.694	**0.893**
YiSi-2	-0.333	-0.333	-0.039	-0.039	-0.190	-0.198	0.123	0.513	0.882

Table 7: Evaluating Human translation: Pearson correlation of metrics with DA human assessment for all MT systems plus Human translation. The subscript *B* represents an alternate reference, *P* represents a paraphrased reference. N is the total number of MT systems (excluding outliers) and HID is the identity of the human translation evaluated. Correlations of metrics not significantly outperformed by any other for that language pair are highlighted in bold.

For to-English language pairs, only the secondary human reference translations were manually scored with DA as the primary human reference translation was shown to the monolingual annotators.

For these language pairs, the metrics can score a human translation by using the other one as the reference translation. For simplicity, we add the second human reference translation to the list of translation outputs and observe how its scoring by the given metric affects the correlation.

Table 7 shows how well the metrics correlate with the WMT human evaluation when including human translations as additional output. In most cases, the correlation decreases as metrics struggle to correctly score translations that are different from MT systems. Metrics that rely on fine-tuning on existing human assessments from the previous WMT campaigns (e.g. BLEURT, ESIM, COMET) can handle human translations much better on average. Also, the Paraphrased references help the lexical metrics correctly identify the high quality of human translations.

We present a deeper analysis of how metrics score human translations in Section 5.1.2. We base this discussion on scatterplots of human vs metric scores. We include scatterplots of selected metrics in Appendix B.

Influence of References Rewarding multiple alternative translations is the primary motivation behind multiple-reference based evaluation. It is generally assumed that using multiple reference translation for automatic evaluation is helpful as we cover a wider space of possible translations (Papineni et al., 2002b; Dreyer and Marcu, 2012; Bojar et al., 2013). Nevertheless, new studies (Freitag et al., 2020) showed that multi-reference evaluation does not improve the correlation for high quality output anymore. Since we have multiple references available for five language pairs, we can look at how much the choice of reference(s) influences correlation.

Table 8 compares metric correlations on the primary reference set *newstest2020*, alternative reference *newstestB2020*, paraphrased reference *newstestP2020* (only for English-German), or using all available references *newstestM2020*. We only report system-level correlations of metrics on MT systems after discarding outliers.

4.2 Segment- and Document-Level Evaluation

Segment-level evaluation relies on the manual judgements collected in the News Translation Task evaluation. This year, again we were unable to follow the methodology outlined in Graham et al. (2015) for evaluating of segment-level metrics because the sampling of segments did not provide sufficient number of assessments of the same segment. We therefore convert pairs of DA scores for competing translations to DARR better/worse preferences as described in Section 2.3.2. We further follow the same process to generate DARR ground truth for documents, as we do not have enough annotations to obtain accurate human scores.

We measure the quality of metrics' scores against the DARR golden truth using a Kendall's Tau-like formulation, which is an adaptation of the conventional Kendall's Tau coefficient. Since we do not have a total order ranking of all translations, it is not possible to apply conventional Kendall's Tau given the current DARR human evaluation setup (Graham et al., 2015).

Our Kendall's Tau-like formulation, τ, is as follows:

$$\tau = \frac{|Concordant| - |Discordant|}{|Concordant| + |Discordant|} \qquad (2)$$

where $Concordant$ is the set of all human comparisons for which a given metric suggests the same order and $Discordant$ is the set of all human comparisons for which a given metric disagrees. The formula is not specific with respect to ties, i.e. cases where the annotation says that the two outputs are equally good.

The way in which ties (both in human and metric judgement) were incorporated in computing Kendall τ has changed across the years of WMT Metrics Tasks. Here we adopt the version used in WMT17 DARR evaluation. For a detailed discussion on other options, see also Macháček and Bojar (2014).

Whether or not a given comparison of a pair of distinct translations of the same source input, s_1 and s_2, is counted as a concordant (Conc) or disconcordant (Disc) pair is defined by the following matrix:

In previous years, we used bootstrap resampling (Koehn, 2004; Graham et al., 2014) to estimate confidence intervals for our Kendall's Tau formulation, and metrics with non-overlapping 95% confidence

	en-de 11	en-de$_B$ 11	en-de$_P$ 11	en-de$_M$ 11	en-zh 12	en-zh$_B$ 12	en-zh$_M$ 12	de-en 9	de-en$_B$ 9	de-en$_M$ 9	ru-en 10	ru-en$_B$ 10	ru-en$_M$ 10	zh-en 15	zh-en$_B$ 15	zh-en$_M$ 15
SENTBLEU	0.823	0.837	0.815	**0.827**	0.927	0.911	0.919	**0.786**	0.763	**0.788**	**0.833**	**0.850**	**0.837**	**0.950**	0.928	**0.944**
BLEU	0.825	0.844	0.830	0.822	0.928	0.899	0.913	0.778	0.797	**0.805**	0.761	0.780	0.775	**0.957**	**0.934**	**0.949**
TER	**0.848**	**0.860**	**0.859**	**0.852**	-0.213	-0.200	-0.203	**0.766**	**0.744**	**0.758**	**0.829**	**0.832**	**0.853**	0.911	0.875	0.911
CHRF++	**0.850**	**0.866**	**0.876**	**0.858**	0.878	0.915	0.885	0.699	0.681	0.704	**0.833**	**0.839**	**0.843**	**0.955**	**0.948**	**0.952**
CHRF	**0.862**	**0.874**	**0.883**	–	0.923	0.912	–	0.687	0.683	–	0.831	**0.839**	–	**0.954**	**0.947**	–
PARBLEU	0.774	0.796	0.724	0.794	0.962	0.955	0.959	**0.838**	**0.831**	**0.829**	0.744	0.767	0.756	**0.953**	**0.934**	**0.945**
PARCHRF++	**0.845**	**0.863**	**0.865**	**0.856**	0.948	**0.966**	0.896	0.708	0.704	0.669	0.823	**0.834**	**0.832**	**0.956**	**0.950**	**0.956**
CHARACTER	**0.868**	**0.889**	**0.835**	**0.878**	0.905	0.908	0.901	0.687	0.696	**0.713**	**0.869**	**0.853**	**0.873**	**0.950**	**0.942**	**0.949**
EED	**0.869**	**0.871**	**0.867**	**0.867**	0.928	0.923	0.930	**0.752**	**0.747**	**0.752**	**0.872**	**0.868**	**0.879**	0.932	0.922	0.932
YISI-0	**0.889**	**0.882**	**0.873**	**0.886**	0.362	0.273	0.332	**0.786**	**0.790**	**0.794**	**0.874**	**0.867**	**0.880**	0.918	0.911	0.918
SWSS+METEOR	–	–	–	–	–	–	–	–	–	–	**0.876**	**0.891**	**0.891**	0.926	0.923	0.929
MEE	0.820	0.833	0.830	**0.820**	–	–	–	0.712	0.674	0.712	**0.878**	**0.876**	**0.876**	**0.948**	**0.940**	**0.948**
PRISM	**0.851**	**0.854**	**0.839**	**0.851**	0.221	0.178	0.221	**0.775**	**0.763**	**0.770**	**0.839**	**0.841**	**0.842**	**0.945**	**0.949**	**0.945**
YISI-1	**0.887**	**0.886**	**0.888**	**0.885**	0.959	0.959	0.960	**0.783**	**0.781**	**0.781**	**0.833**	**0.837**	**0.838**	**0.942**	**0.942**	**0.943**
YISI-COMBI	0.868	0.873	0.876	**0.876**	–	–	–	–	–	–	–	–	–	–	–	–
BLEURT-YISI-COMBI	0.868	0.873	0.876	**0.876**	–	–	–	–	–	–	–	–	–	–	–	–
BERT-BASE-L2	–	–	–	–	–	–	–	0.791	**0.798**	**0.802**	**0.836**	**0.833**	**0.835**	0.929	**0.936**	**0.933**
BERT-LARGE-L2	–	–	–	–	–	–	–	0.800	**0.801**	**0.812**	**0.843**	**0.844**	**0.850**	0.928	**0.935**	**0.932**
MBERT-L2	**0.861**	**0.862**	0.841	**0.865**	0.934	0.925	0.936	**0.824**	**0.825**	**0.834**	0.805	0.813	0.816	**0.935**	**0.938**	**0.939**
BLEURT	–	–	–	–	–	–	–	0.770	**0.769**	0.780	**0.844**	**0.847**	**0.850**	**0.931**	**0.936**	**0.935**
BLEURT-EXTENDED	**0.870**	**0.870**	**0.860**	**0.867**	0.928	0.923	0.925	**0.818**	**0.805**	**0.812**	0.797	0.793	0.795	0.931	0.932	0.932
ESIM	**0.894**	**0.900**	**0.887**	**0.898**	0.972	**0.975**	0.976	0.808	**0.798**	**0.804**	**0.834**	**0.842**	**0.839**	0.910	**0.920**	0.916
PARESIM-1	**0.886**	**0.897**	**0.878**	**0.890**	**0.983**	**0.983**	**0.985**	**0.835**	**0.807**	**0.822**	**0.828**	**0.840**	**0.835**	0.910	0.918	0.915
COMET	**0.863**	**0.864**	**0.858**	**0.864**	0.007	-0.014	-0.004	0.773	**0.769**	0.772	**0.836**	**0.836**	**0.836**	**0.931**	**0.936**	**0.934**
COMET-2R	**0.869**	**0.869**	**0.866**	**0.875**	-0.066	-0.076	-0.075	0.772	**0.764**	0.771	**0.843**	**0.842**	**0.843**	**0.928**	**0.930**	**0.929**
COMET-HTER	**0.852**	**0.855**	**0.848**	**0.853**	-0.073	-0.075	-0.074	**0.767**	**0.769**	**0.768**	0.741	0.744	0.742	0.873	0.869	0.871
COMET-MQM	0.840	0.844	0.836	0.842	0.084	0.076	0.080	**0.684**	**0.686**	**0.685**	0.746	0.750	0.748	0.862	0.860	0.861
COMET-RANK	**0.860**	0.839	**0.831**	**0.852**	0.540	0.507	0.530	**0.757**	0.582	**0.723**	0.732	0.743	0.757	**0.909**	**0.908**	**0.919**

Table 8: Influence of references: Pearson correlation of metrics with DA human assessment for MT systems excluding outliers in WMT2020 for all language-pairs with multiple references; correlations of metrics not significantly outperformed by any other for that language pair are highlighted in bold. The subscript *B* represents a secondary reference, *P* represents a paraphrased reference, *M* represents all available references.
Note that we exclude reference-free metrics from this table, so the winners are not comparable with the main tables.

		Metric		
		$s_1 < s_2$	$s_1 = s_2$	$s_1 > s_2$
Human	$s_1 < s_2$	Conc	Disc	Disc
	$s_1 = s_2$	–	–	–
	$s_1 > s_2$	Disc	Disc	Conc

intervals are identified as having statistically significant difference in performance. The tests are inconclusive for most metric pairs this year and we do not include them in the paper.

4.2.1 Segment-Level Results

Results of the segment-level human evaluation for translations sampled from the News Translation Task are shown in Tables 9 and 10, We expect that comparing between segments translated by two MT systems that are far apart in quality would be a relatively easier task for automatic metrics. So we also include results after discarding segments that were translated by outlier systems.

Note that we do not include any human-translated segments in this evaluation.

4.3 Document-level Results

Results of the document-level human evaluation for translations sampled from the News Translation Task are shown in Tables 11 and 12.

5 Discussion

5.1 System-Level Results

In general, there is no clear best metric this year across all language pairs. For most language pairs, the William's significance test results in large clusters of metrics. The set of "winners" according to the test (i.e., the metrics that are not outperformed by any other metric) are typically not consistent across language pairs.

The sample of systems we employ to evaluate metrics is often small, as few as six MT systems for Pashto → English, for example. This can lead to inconclusive results, as identification of significant differences in correlations of metrics is unlikely at such a small sample size. Furthermore, Williams test takes into account the correlation between each pair of metrics, in addition to the correlation between the metric scores themselves, and this latter correlation increases the likelihood of a significant difference being identified. In extreme cases, the test would have low power when comparing a metric that doesn't correlate well with other metrics, resulting in this metric not being outperformed by other metrics despite having a much lower value of correlation.

To strengthen the conclusions of our evaluation, in past years (Bojar et al., 2016, 2017; Ma et al., 2018), we included significance test results for large hybrid-super-samples of systems 10K hybrid systems were created per language pair, with corresponding DA human assessment scores by sampling pairs of systems from the News Translation Task, creating hybrid systems by randomly selecting each candidate translation from one of the two selected systems. However, as WMT human annotations are collected with document context in 2020, this style of hybridization is susceptible to breaking cross-segment references in MT outputs and it would be unreasonable to shuffle individual segments. The creation of hybrid systems would need to be done by sampling documents instead of segments from all sets of systems. Finally, it is possible that including documents translated by outlier systems might falsely lead to high correlations. We believe that this merits further investigation based on data from previous of metrics tasks, and we do not attempt it this year.

In the rest of this section, we present analysis of various aspects of system-level evaluation based on scatterplots of all metrics. Appendix B contains scatterplots of metrics for each language pair. We include BLEU, chrF, the "best" reference-based metric and the "best" reference-free metric (we acknowledge that this is not the best way to define the best metric, but we choose the metric that is most highly correlated with humans on the set of all MT systems after removing outliers).

5.1.1 Influence of Domain in English → Inuktitut

English → Inuktitut training data was the Canadian Hansards domain, and the development data contained a small amount of news data. The test set was a mix of in-domain data from the Hansards and news documents. The evaluation was only done on the out-of-domain news documents, so we also look at metric scores computed only on the subset of news sentences.

Figure 1 shows that BLEU scores on the out-of-domain dataset are considerably smaller than the full dataset, showing that MT systems have a higher quality on the in-domain dataset. The relative scores of metrics remain mostly stable when we compare scores on the full test set to scores on

	cs-en		de-en		iu-en		ja-en		km-en	pl-en		ps-en	ru-en		ta-en		zh-en	
	all	all-out	all	all-out	all	all-out	all	all-out	all	all	all-out	all	all	all-out	all	all-out	all	all-out
	14018	9461	16584	6185	8162	5381	15193	6286	3706	21121	17979	3507	14024	11020	12789	8749	62586	53610
sentBLEU	0.068	0.057	0.413	-0.025	0.182	0.170	0.188	0.061	0.226	-0.024	-0.046	0.096	-0.005	-0.038	0.162	0.069	0.093	0.060
TER	-0.04	-0.06	0.355	-0.137	0.021	0.012	0.044	-0.077	0.125	-0.172	-0.196	-0.036	-0.117	-0.154	0.046	-0.063	-0.01	-0.047
chrF++	0.090	0.075	0.435	0.013	0.246	0.251	0.245	0.115	0.275	0.034	0.009	0.145	0.054	0.018	0.186	0.098	0.130	0.096
chrF	0.086	0.072	0.438	0.018	0.254	0.260	0.242	0.109	0.267	0.028	0.003	0.144	0.049	0.012	0.186	0.096	0.132	0.098
parBLEU	0.058	0.038	0.415	-0.039	0.167	0.161	0.198	0.074	0.203	-0.025	-0.049	0.100	-0.011	-0.052	0.159	0.064	0.095	0.059
parChrF++	0.096	0.082	0.436	0.009	0.232	0.235	0.247	0.117	0.267	0.027	0.002	0.147	0.044	0.007	0.184	0.095	0.132	0.099
CharacTER	0.090	0.087	0.440	0.011	0.214	0.220	0.221	0.106	0.248	0.023	-0.002	0.172	0.057	0.028	0.138	0.078	0.123	0.093
EED	0.091	0.078	0.440	0.013	0.256	0.258	0.235	0.116	0.271	0.045	0.022	0.149	0.053	0.018	0.198	0.103	0.129	0.093
YiSi-0	0.072	0.065	0.441	0.024	0.261	0.263	0.241	0.121	0.268	0.035	0.013	0.140	0.065	0.030	0.183	0.089	0.127	0.093
SWSS+METEOR	–		–		0.226	0.218	0.228	0.086	0.264	0.011	-0.016	0.130	0.048	0.010	0.205	0.120	0.133	0.099
MEE	0.063	0.045	0.402	-0.04	0.134	0.126	0.187	0.064	0.206	-0.084	-0.105	0.078	-0.041	-0.084	0.114	0.032	0.083	0.050
YiSi-1	0.117	0.103	0.468	0.051	0.253	0.260	0.277	0.128	0.316	0.042	0.023	0.147	0.091	0.049	0.248	0.162	0.146	0.115
BERT-base-L2	0.103	0.087	0.454	0.026	0.238	0.229	0.263	0.129	0.295	0.032	0.013	0.159	0.087	0.037	0.223	0.135	0.141	0.113
BERT-large-L2	0.102	0.087	0.456	0.025	0.251	0.249	0.262	0.114	0.314	0.044	0.027	0.151	0.094	0.047	0.245	0.157	0.133	0.102
mBERT-L2	0.119	0.111	0.442	0.001	0.244	0.235	0.251	0.120	0.312	0.047	0.029	0.151	0.083	0.036	0.227	0.139	0.133	0.104
BLEURT	0.126	0.118	0.456	0.015	0.258	0.256	0.265	0.123	0.327	0.057	0.040	0.207	0.093	0.046	0.230	0.145	0.137	0.107
BLEURT-extended	0.127	0.113	0.448	0.004	0.259	0.259	0.271	0.124	0.330	0.044	0.019	0.161	0.101	0.057	0.246	0.165	0.137	0.107
esim	0.110	0.103	0.454	0.031	0.241	0.233	0.239	0.119	0.300	0.058	0.045	0.147	0.084	0.044	0.208	0.117	0.138	0.108
paresim-1	0.105	0.098	0.464	0.051	0.249	0.241	0.242	0.121	0.292	0.066	0.055	0.149	0.089	0.049	0.213	0.123	0.139	0.111
COMET	0.129	0.112	0.485	0.090	0.281	0.271	0.274	0.127	0.298	0.099	0.085	0.158	0.156	0.117	0.241	0.163	0.171	0.142
COMET-2R	0.120	0.107	0.479	0.101	0.257	0.251	0.268	0.120	0.308	0.098	0.085	0.144	0.148	0.110	0.253	0.177	0.163	0.136
COMET-HTER	0.103	0.087	0.481	0.088	0.198	0.199	0.241	0.095	0.269	0.080	0.067	0.116	0.131	0.098	0.227	0.151	0.135	0.113
COMET-MQM	0.108	0.097	0.483	0.100	0.215	0.209	0.259	0.112	0.282	0.080	0.066	0.141	0.137	0.102	0.227	0.158	0.141	0.117
COMET-Rank	0.099	0.096	0.470	0.061	0.188	0.181	0.235	0.086	0.228	0.073	0.057	0.107	0.118	0.082	0.199	0.112	0.142	0.117
COMET-QE	0.091	0.072	0.410	0.042	0.031	0.020	0.153	0.048	0.148	0.039	0.029	0.092	0.084	0.049	0.163	0.099	0.088	0.070
OpenKiwi-Bert	0.036	0.029	0.379	0.013	-0.005	-0.009	0.110	0.000	0.168	-0.033	-0.043	0.076	-0.033	-0.067	0.118	0.052	0.029	0.020
OpenKiwi-XLMR	0.093	0.079	0.463	0.074	0.056	0.031	0.220	0.086	0.244	0.059	0.051	0.106	0.092	0.065	0.188	0.109	0.115	0.089
YiSi-2	0.068	0.054	0.413	0.006	0.039	0.028	0.204	0.074	0.214	0.048	0.042	0.073	0.070	0.056	0.199	0.113	0.116	0.084
prism	0.143	0.135	0.475	0.057	0.255	0.254	0.272	0.146	0.304	0.109	0.093	0.165	0.145	0.111	0.237	0.151	0.167	0.138
BAQ_dyn	–		–		–		–		–	–		–	–		–		0.119	0.089
BAQ_static	–		–		–		–		–	–		–	–		–		0.119	0.087

Table 9: Segment-level metric results for to-English language pairs: Kendall's Tau formulation of segment-level metric scores with DARR scores. For language pairs that contain outlier systems, we also show correlation after discarding segments translated by outlier systems

	en-cs		en-de		en-iu		en-ja	en-pl		en-ru	en-ta		en-zh
	all	all-out	all	all-out	all	all-out	all	all	all-out	all	all	all-out	all
	21121	10283	9339	4637	13159	5490	12830	17689	9316	8330	9087	3695	12652
sentBLEU	0.432	0.194	0.303	0.155	0.206	-0.084	0.479	0.153	0.067	0.051	0.398	0.206	0.396
TER	0.317	0.067	0.182	0.044	-0.071	-0.337	-0.591	0.003	-0.094	-0.121	0.203	0.019	-0.36
chrF++	0.478	0.228	0.367	0.215	0.338	0.075	0.506	0.255	0.154	0.156	0.579	0.349	0.388
chrF	0.472	0.229	0.379	0.224	0.344	0.095	0.506	0.250	0.150	0.153	0.589	0.359	0.400
parBLEU	0.460	0.226	0.299	0.136	0.212	-0.051	0.052	0.183	0.088	0.062	0.340	0.178	0.356
parCHRF++	0.492	0.253	0.355	0.192	–		0.527	0.272	0.167	0.176	–		0.398
CharacTER	0.413	0.195	0.311	0.179	0.309	0.108	0.471	0.198	0.107	0.143	0.525	0.270	0.339
EED	0.458	0.210	0.363	0.203	0.361	0.109	0.515	0.248	0.151	0.155	0.587	0.342	0.393
YiSi-0	0.432	0.191	0.349	0.212	0.362	0.101	0.484	0.233	0.132	0.151	0.547	0.336	0.319
MEE	0.411	0.157	0.289	0.128	-0.074	-0.272	–	0.125	0.025	0.027	0.373	0.168	–
YiSi-1	0.550	0.320	0.427	0.263	0.251	0.082	0.568	0.349	0.209	0.256	0.669	0.440	0.463
YISI-COMBI	–		0.399	0.224	–		–	–		–	–		–
BLEURT-COMBI	–		0.399	0.224	–		–	–		–	–		–
mBERT-L2	0.567	0.359	0.361	0.202	–		0.541	0.350	0.212	0.246	0.587	0.334	0.432
BLEURT-EXTENDED	0.689	0.517	0.447	0.278	0.359	0.112	0.533	0.430	0.271	0.305	0.643	0.419	0.460
ESIM	0.469	0.253	0.347	0.195	0.122	-0.018	0.522	0.312	0.203	0.224	0.599	0.363	0.391
PARESIM-1	0.475	0.257	0.343	0.197	0.122	-0.018	0.510	0.324	0.209	0.230	0.599	0.363	0.396
COMET	0.668	0.487	0.468	0.324	0.322	0.078	0.624	0.462	0.316	0.344	0.671	0.457	0.432
COMET-2R	0.669	0.512	0.463	0.321	0.326	0.078	0.630	0.445	0.294	0.343	0.676	0.463	0.434
COMET-HTER	0.665	0.500	0.440	0.303	0.331	0.088	0.601	0.427	0.274	0.292	0.640	0.411	0.411
COMET-MQM	0.666	0.490	0.423	0.275	0.313	0.078	0.588	0.424	0.271	0.281	0.635	0.413	0.388
COMET-Rank	0.629	0.408	0.379	0.217	0.297	0.097	0.569	0.388	0.207	0.229	0.588	0.342	0.380
COMET-QE	0.614	0.470	0.347	0.233	-0.04	-0.051	0.470	0.360	0.211	0.264	0.514	0.320	0.346
OpenKiwi-Bert	0.262	0.142	0.168	0.058	-0.115	-0.233	-0.529	0.153	0.035	0.164	0.169	0.022	0.077
OpenKiwi-XLMR	0.607	0.417	0.369	0.224	0.060	0.009	0.553	0.347	0.189	0.279	0.604	0.354	0.377
YiSi-2	0.187	0.104	0.296	0.171	0.146	0.073	0.383	0.115	0.052	0.146	0.545	0.332	0.152
PRISM	0.619	0.414	0.447	0.280	0.452	0.195	0.579	0.414	0.274	0.283	0.448	0.211	0.397
BAQ_DYN	–		–		–		–	–		–	–		0.351
BAQ_STATIC	–		–		–		–	–		–	–		0.344
EQ_DYN	–		–		–		–	–		–	–		0.356
EQ_STATIC	–		–		–		–	–		–	–		0.409

Table 10: Segment-level metric results for out-of-English language pairs: Kendall's Tau formulation of segment-level metric scores with DARR scores; For language pairs that contain outlier systems, we also show correlation after discarding segments translated by outlier systems

	cs-en		de-en		iu-en		ja-en		pl-en		ru-en		ta-en		zh-en	
	all	all-out	all	all-out	all	all-out	all	all-out	all	all-out	all	all-out	all	all-out	all	all-out
	1424	955	1866	495	36	24	790	311	635	529	753	581	684	440	3085	2618
SENTBLEU	0.104	0.058	0.601	-0.055	0.611	0.417	0.413	0.125	0.096	0.059	0.113	0.026	0.330	0.150	0.211	0.153
TER	0.115	0.068	0.621	-0.002	0.611	0.500	0.370	0.048	0.024	-0.059	0.089	-0.009	0.383	0.214	0.197	0.141
CHRF++	0.135	0.110	0.624	0.006	0.500	0.333	0.435	0.158	0.071	0.036	0.169	0.088	0.395	0.209	0.199	0.139
CHRF	0.126	0.091	0.626	0.002	0.611	0.417	0.453	0.209	0.065	0.021	0.195	0.112	0.395	0.200	0.209	0.154
PARBLEU	0.100	0.045	0.630	-0.002	0.556	0.417	0.428	0.138	0.065	0.032	0.086	-0.019	0.368	0.177	0.201	0.143
PARCHRF++	0.117	0.081	0.642	0.042	0.611	0.417	0.438	0.164	0.087	0.040	0.171	0.095	0.412	0.232	0.203	0.146
CHARACTER	0.059	0.049	0.646	0.079	0.500	0.250	0.410	0.145	0.090	0.051	0.187	0.122	0.371	0.196	0.219	0.166
EED	0.105	0.064	0.633	0.006	0.722	0.583	0.430	0.125	0.080	0.017	0.174	0.088	0.395	0.200	0.206	0.148
YISI-0	0.052	0.022	0.616	0.006	0.556	0.333	0.425	0.125	0.071	0.036	0.187	0.098	0.409	0.223	0.196	0.139
SWSS+METEOR	–		–		0.722	0.583	0.377	0.029	0.109	0.047	0.211	0.129	0.447	0.291	0.201	0.141
MEE	0.126	0.114	0.618	-0.006	0.444	0.250	0.438	0.190	0.014	-0.013	0.137	0.053	0.398	0.245	0.198	0.140
YISI-1	0.136	0.114	0.640	0.034	0.667	0.500	0.420	0.119	0.109	0.062	0.150	0.033	0.450	0.300	0.210	0.153
BERT-BASE-L2	0.164	0.139	0.654	0.075	0.778	0.667	0.430	0.151	0.046	-0.013	0.179	0.064	0.398	0.223	0.206	0.149
BERT-LARGE-L2	0.131	0.091	0.642	0.030	0.722	0.583	0.418	0.119	0.027	-0.028	0.195	0.084	0.439	0.291	0.185	0.124
MBERT-L2	0.149	0.118	0.621	-0.006	0.833	0.750	0.433	0.158	0.033	-0.036	0.232	0.126	0.418	0.259	0.216	0.162
BLEURT	0.154	0.125	0.641	0.038	0.667	0.500	0.420	0.100	0.039	-0.009	0.227	0.115	0.418	0.259	0.197	0.141
BLEURT-EXTENDED	0.140	0.114	0.633	0.014	0.833	0.750	0.430	0.113	0.077	0.006	0.243	0.143	0.412	0.245	0.198	0.141
ESIM	0.135	0.110	0.670	0.164	0.722	0.583	0.400	0.087	0.039	-0.017	0.174	0.064	0.404	0.236	0.203	0.148
PARESIM-1	0.119	0.093	0.670	0.156	0.722	0.583	0.392	0.055	0.033	-0.021	0.171	0.060	0.401	0.232	0.208	0.154
COMET	0.142	0.114	0.626	-0.018	0.667	0.500	0.392	0.061	0.112	0.070	0.193	0.088	0.395	0.218	0.206	0.151
COMET-2R	0.138	0.116	0.614	0.030	0.778	0.667	0.413	0.093	0.090	0.047	0.227	0.136	0.404	0.232	0.214	0.158
COMET-HTER	0.160	0.133	0.638	0.042	0.556	0.333	0.415	0.138	0.083	0.040	0.169	0.084	0.354	0.191	0.150	0.105
COMET-MQM	0.140	0.114	0.645	0.075	0.611	0.417	0.410	0.119	0.080	0.043	0.163	0.081	0.386	0.241	0.161	0.117
COMET-RANK	0.139	0.131	0.615	-0.026	0.667	0.500	0.365	0.035	0.112	0.096	0.185	0.074	0.325	0.154	0.199	0.147
COMET-QE	0.091	0.060	0.636	0.042	0.389	0.250	0.329	0.023	-0.002	-0.028	0.153	0.060	0.301	0.127	0.169	0.118
OPENKIWI-BERT	0.087	0.064	0.628	0.046	0.444	0.250	0.322	0.113	0.096	0.077	0.137	0.050	0.281	0.145	0.113	0.079
OPENKIWI-XLMR	0.133	0.114	0.613	0.010	0.556	0.500	0.418	0.145	0.055	0.017	0.155	0.060	0.389	0.227	0.187	0.135
YISI-2	0.083	0.072	0.547	-0.075	0.278	0.250	0.385	0.055	0.118	0.153	0.248	0.195	0.383	0.196	0.199	0.139
PRISM	0.169	0.156	0.636	-0.002	0.667	0.500	0.420	0.119	0.102	0.059	0.211	0.102	0.406	0.236	0.195	0.138
BAQ_DYN	–		–		–		–		–		–		–		0.223	0.172
BAQ_STATIC	–		–		–		–		–		–		–		0.214	0.160

Table 11: Document-level metric results for to-English language pairs: Kendall's Tau formulation of segment-level metric scores with DAdocument-level metric scores with DOC-DARR judgements. For language pairs that contain outlier systems, we also show correlation after discarding documents translated by outlier systems.

	en-cs		en-de		en-iu		en-ja	en-pl		en-ru	en-ta		en-zh
	all	all-out	all	all-out	all	all-out	all	all	all-out	all	all	all-out	all
	1442	572	729	312	203	48	469	677	254	387	389	99	651
SENTBLEU	0.680	0.273	0.550	0.359	0.596	-0.25	0.808	0.510	0.150	0.287	0.799	0.596	0.598
TER	0.691	0.294	0.517	0.308	0.567	-0.292	-0.07	0.439	0.094	0.178	0.748	0.616	0.118
CHRF++	0.692	0.294	0.583	0.372	0.547	-0.417	0.829	0.536	0.165	0.339	0.866	0.596	0.579
CHRF	0.688	0.290	0.597	0.397	0.576	-0.333	0.838	0.524	0.165	0.307	0.872	0.616	0.591
PARBLEU	0.727	0.381	0.528	0.269	0.576	-0.208	0.565	0.569	0.236	0.266	0.805	0.596	0.625
PARCHRF++	0.717	0.364	0.575	0.340	–		0.825	0.560	0.205	0.307	–		0.650
CHARACTER	0.656	0.224	0.520	0.263	0.547	-0.292	0.842	0.448	0.047	0.328	0.872	0.657	0.613
EED	0.678	0.280	0.569	0.340	0.596	-0.292	0.834	0.554	0.228	0.277	0.856	0.556	0.588
YISI-0	0.653	0.245	0.553	0.327	0.645	-0.125	0.821	0.554	0.228	0.318	0.830	0.515	0.441
MEE	0.714	0.357	0.558	0.314	0.527	-0.375	–	0.489	0.071	0.307	0.805	0.495	–
YISI-1	0.763	0.462	0.605	0.359	0.616	-0.083	0.855	0.663	0.339	0.349	0.882	0.657	0.690
YISI-COMBI	–		0.594	0.353	–		–	–		–	–		–
BLEURT-COMBI	–		0.594	0.353	–		–	–		–	–		–
MBERT-L2	0.781	0.517	0.580	0.308	–		0.842	0.648	0.299	0.431	0.861	0.596	0.693
BLEURT-EXTENDED	0.847	0.664	0.635	0.378	0.635	-0.167	0.851	0.740	0.488	0.437	0.856	0.636	0.708
ESIM	0.720	0.392	0.534	0.237	0.507	-0.25	0.855	0.616	0.315	0.354	0.836	0.596	0.674
PARESIM-1	0.741	0.441	0.528	0.237	0.507	-0.25	0.829	0.628	0.331	0.364	0.836	0.596	0.662
COMET	0.845	0.664	0.632	0.404	0.547	-0.125	0.847	0.687	0.346	0.359	0.897	0.758	0.561
COMET-2R	0.859	0.699	0.613	0.391	0.596	-0.25	0.868	0.725	0.409	0.375	0.866	0.636	0.558
COMET-HTER	0.849	0.675	0.616	0.410	0.606	-0.042	0.855	0.669	0.307	0.287	0.856	0.657	0.502
COMET-MQM	0.849	0.675	0.594	0.372	0.567	-0.125	0.817	0.678	0.291	0.364	0.836	0.657	0.472
COMET-RANK	0.803	0.573	0.572	0.321	0.586	-0.375	0.812	0.607	0.165	0.307	0.841	0.596	0.472
COMET-QE	0.839	0.650	0.514	0.250	0.212	-0.167	0.812	0.619	0.142	0.297	0.820	0.657	0.469
OPENKIWI-BERT	0.655	0.399	0.443	0.147	0.488	0.083	0.139	0.427	0.024	0.344	0.584	0.111	0.459
OPENKIWI-XLMR	0.821	0.622	0.589	0.314	0.527	0.167	0.859	0.592	0.094	0.328	0.856	0.596	0.524
YISI-2	0.255	0.105	0.416	0.224	0.527	0.333	0.680	0.117	-0.118	0.209	0.805	0.434	0.223
PRISM	0.792	0.545	0.594	0.378	0.665	-0.042	0.829	0.634	0.283	0.328	0.733	0.374	0.511
BAQ_DYN	–		–		–		–	–		–	–		0.567
BAQ_STATIC	–		–		–		–	–		–	–		0.619
EQ_DYN	–		–		–		–	–		–	–		0.613
EQ_STATIC	–		–		–		–	–		–	–		0.644

Table 12: Document-level metric results for out-of-English language pairs: Kendall's Tau formulation of document-level metric scores with DOC-DARR judgements. For language pairs that contain outlier systems, we also show correlation after discarding documents translated by outlier systems

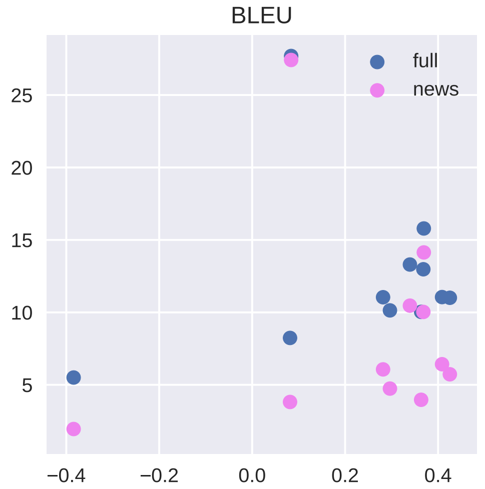


Figure 1: English → Inuktitut: Human vs. BLEU scores on the full dataset vs. the news subset. Only the news subset was included in the human evaluation. Each dot corresponds to an MT system, the outlier on the top-right is UQAM_TANLE.

only the news subset that was evaluated. The main exception is UQAM_TANLE; BLEU scores are really high on the out-of-domain data, and increase very little when computed on the full dataset. When looking at correlations with human scores (Table 7), we expected correlations to increase when computed over the news subset. This is true for most metrics such as COMET-QE, but the correlation stays the same or actually decreases for other metrics like PARBLEU.

5.1.2 Scoring Human Translations

The alternate reference was included in the manual evaluation for German → English, Russian → English and Chinese → English. All human references were included in the out-of-English manual evaluation.[7]

German → English: HUMAN-B was ranked third in the manual evaluation. The lexical metrics (BLEU, CHRF, CHARACTER, EED, MEE, YISI-0, CHRF++, PARBLEU, PARCHRF) give extremely low scores to the HUMAN-B reference. This is also true for PRISM and all the reference-free metrics except COMET-QE. The neural metrics also give low scores to the human reference, however, the margin of error is much smaller. COMET-QE is the only metric that gives high scores to the HUMAN-B reference.

Appendix B also shows the scatterplot "newstestB2020" where HUMAN-B served as the reference for the metrics. We see some differences in the vertical axis but the general picture remains the same even with this fairly different human translation.

Russian → English: The HUMAN-B reference was ranked after 6 MT systems in the manual evaluation but still within the same cluster, so not significantly distinguishable. Lexical metrics give relatively low scores to HUMAN-B. The neural metrics give relatively higher scores, but score it above *Online-A* and below *ariel197197*, i.e. differently than DA judgements.

Chinese → English: The Human translation is ranked 12th in the manual evaluation (in a giant cluster which puts together all but one top and one bottom system), and most metrics place it more or less correctly. Many metrics, including lexical metrics, still have correlations above 0.9 even after including the Human translation.

English → German: According to the WMT human evaluation, the HUMAN-B reference receives the highest scores, the HUMAN-A reference is ranked fourth and Human-P, which was generated by linguists paraphrasing the WMT references, is ranked lower at 10th place. Each human reference falls into a separate cluster of significance.

Lexical metrics score around 10 MT systems above each WMT reference (using the other WMT human translation as reference). COMET-QE and some neural metrics (BLEURT, COMET-MQM, COMET-HTER and MBERT-L2) score HUMAN-A and HUMAN-B as better than all MT systems.

When using either of the WMT references, most metrics, including all the lexical metrics, score the paraphrased reference much lower than the rest of the systems. The COMET family of metrics and BLEURT-EXTENDED are the only metrics that are able to recognise the merit of the paraphrased references.

When using the paraphrased references, all reference-based metrics score the two human translations above all MT systems, often by a large margin. PRISM is the sole exception; it scores the HUMAN-B reference about half way between the MT systems. Interestingly, most of these metrics score HUMAN-A above HUMAN-B, i.e. dis-

[7] Findings 2020 in the official tables label the alternate reference in into-English direction simply as HUMAN. The "first" reference, which serves as the primary reference for us, was not scored manually in DA into English. Out of English, the primary reference for us is labelled HUMAN-A in Findings.

agreeing with DA judgements. Metric correlations when including HUMAN-A system drop dramatically when using the alternate WMT reference, but the correlations are higher with the paraphrased reference. This also holds when scoring HUMAN-B using the paraphrased vs the main WMT reference (Table 7).

Of the reference-free metrics, COMET-QE scores the two WMT references above all MT systems, and ranks the paraphrased reference similar to its rank in the manual evaluation. OPENKIWI-BERT and OPENKIWI-XLMR are a little biased against these human translations, and YISI-2 scores all human translations below all MT systems.

English → Chinese The manual evaluation ranks the two Human translations above all MT systems, but most metrics give these much low scores.

To summarize, we see that the current MT metrics generally struggle to score human translations against machine translations reliably. Rare exceptions include primarily trained neural metrics and reference-less COMET-QE. While the metrics are not really prepared to score human translations, we find this type of test relevant as more and more language pairs are getting closer to the human translation benchmark. A general-enough metric should be thus able to score human translation comparably and not rely on some idiosyncratic properties of MT outputs. We hope that human translations will be included in WMT DA scoring in the upcoming years, too.

5.1.3 Influence of Outliers

There are no outlier systems for some language-pairs like Khmer→ English and English → Russian. For others, we have systems whose score is far away from the scores of the rest of the systems. As these outliers have a large influence on Pearson correlation, computing the correlation without outliers typically makes the task harder for metrics and results in a decrease in correlation.

For example, we identify three outliers in the German → English set; the quality of the last system is extremely low compared to rest. All reference-based metrics have high correlations when including all systems, but correlations drop when discarding outliers. In particular, CHRF and PARESIM both had a correlation of 0.95 when computed over all systems, but this drops to 0.69 and 0.83 respectively after removing outliers, revealing that PARESIM is more reliable with this language pair. An even larger drop is observed for CHRF and CHRF++ in English → Czech, from 0.8 to 0.3. We find this particularly surprising because CHRF has always performed well on this language pair, including in the evaluation on the gradually reducing set of top N systems, i.e. in harder and harder conditions, see SACREBLEU-CHRF in Appendix A.4 of Ma et al. (2019).

In some cases, metrics can be inaccurate when scoring outliers, resulting in an increased correlation when correlation is recomputed over non-outlier systems. For example, with Chinese → English, the score of WMTBIOMEDBASELINE score is much lower than the next system. Most metrics correctly rank it last as well, but COMET-HTER, COMET-MQM, COMET-QE and OPENKIWI-BERT give it a higher score than the next system(s). Note that the other metrics all have a correlation of above 0.9 even after removing the outlier.

In other cases, removing outliers decreases the correlation of a metric and yet it helps its final outcome. For instance SENTBLEU averaged over all sentences becomes one of the "winners" in the system-level evaluation of translation into English (Table 5). If we trust the results without outliers more, using *averaged sentBLEU* seems better than using plain old BLEU and not significantly worse than any other metric going from English into several target languages.

For some language-pairs, we override the decisions made by the outlier detection algorithm, based on whether we believe including or removing these systems from consideration would have an impact on the correlations: For example, with Tamil → English, the last two systems are not classified as outliers by the algorithm, but their human scores is some distance away from the rest of the systems. CHRF, CHRF++ and PARCHRF++ are the only metrics that correctly order these two systems. OPENKIWI-BERT and OPENKIWI-XLMR both get these two systems wrong with a large margin. But for all metrics, removing these systems leads to a significant drop in correlation. Thus we count these two systems as outliers.

Another example is Japanese → English. For this language-pair, we have two clusters of 7 and 3 systems. Metrics have high correlations when considering all systems, but when looking at MT systems within individual clusters, there are discrepancies between the metric scores compared to human scores. The outlier detection algorithm flags only the last two systems as outliers, but the presence of the third system has a disproportionate impact on the correlation. We include all three systems in the set of outliers.

The influence of references For all language pairs where multiple references were available, the correlations are typically very close whether using the primary reference or the alternate reference. For metrics where we do see a difference, there is no consistent pattern whether metrics prefer one reference or the other. We note that although the change in correlations is small when comparing across reference sets, the set of "winners" according to the William's test for statistical significance is not stable, particularly for English → German. When combining references, in most cases, the correlation with multiple references lies between the correlation of the individual references. For example, with English → German, BLEU correlates best with the secondary reference with a correlation of 0.844. But with multiple references, the correlation is 0.825, just above the correlation with the primary reference with is 0.822 (Table 8).

There are a few exceptions where there is a small increase in metric correlation above both individual references. For example, the correlation of CHARACTER with German → English increases from 0.687 and 0.696 with a single reference to 0.713 with both references (Table 8). But there are no metrics which consistently show an improvement with multiple references across multiple language pairs.

5.1.4 Neural vs. Lexical Metrics

For many language pairs, when we look at correlation clustering of the reference-based metrics based on their system-level scores, we end up with two major clusters: neural metrics and lexical metrics. We have seen that lexical and neural metrics differ in how they score the human translations. For English → German, all lexical metrics have a slightly higher correlation than any neural metric when evaluating MT systems. However, these metrics make major errors evaluating the HUMAN-A translations with the HUMAN-B reference.

We also see such differences with some MT systems. Selected examples:

- English → Czech: All lexical metrics including BLEU and CHRF are very biased towards ONLINE-B, with metric scores indicating that this system is better than all others by a large margin. It is ranked 7th in the human evaluation. Neural metrics and reference-free metrics are more or less correct when scoring this system. Surprisingly, ESIM is an exception to this, and also ranks ONLINE-B on top.
- Polish → English: Lexical metrics like BLEU give very low scores to ONLINE-G.
- Tamil → English: Lexical metrics consistently score ONLINE-Z above MICROSOFT_STC_INDIA, but the remaining metrics including the reference-free metrics rank them in the opposite order. The human evaluation agrees with the lexical metrics.
- Khmer → English: lexical metrics score the best system lower than the next two, whereas most neural metrics get the order of the top systems right.

5.1.5 Other Discrepancies between Metric and Human Scores

Here we briefly draw attention to particularities we spotted when manually examining the results.

- German → English: All metrics score Tohoku-AIP-NTT higher than OPPO, and UEDIN higher than PROMT_NMT.
- Russian → English: ONLINE-A, which is ranked 2nd in the human evaluation, receives low metric scores. In contrast, some metrics including BLEU and PARBLEU choose ARIEL197197, which is ranked 6th in the human evaluation, as the best system.
- Tamil → English: The highest ranked system according to human scores, GTCOM, receives lower metric scores than the next three to six systems. Metrics are biased towards ONLINE-A and against ONLINE-Z.
- Chinese → English: HUOSHAN_TRANSLATE is a clear winner according to human evaluation, but BLEU ranks it lower than the next 3 systems. The different between human scores for the next 8 systems is not statistically significant where metric ordering of the systems differently than human scores and these discrepancies aren't penalised harshly by Pearson correlation.

- English → Chinese: HUOSHAN_TRANSLATE is a clear winner according to human evaluation, but BLEU ranks it lower than the next 3 systems. The different between human scores for the next 8 systems is not statistically significant where metric ordering of the systems differently than human scores and these discrepancies aren't penalised harshly by Pearson correlation. While many metrics including BLEU have high correlations, others make major errors scoring the NIUTRANS. OPENKIWI-BERT assigns really low scores to

Overall, we note that these metric-human discrepancies often feature online systems which are probably more diverse that the MT system submissions to the WMT shared tasks.

5.1.6 Pearson vs. Kendall Tau

Overall, we found that Pearson correlation doesn't always give us the complete picture. In particular, outliers have a large influence on the correlation and can mask the presence of discrepancies between metric and human scores with the rest of the systems. But making a decision on which systems to discard is not easy.

In this paper, we also explore Kendall's Tau as an alternative to Pearson correlation. Tables 16 and 17 in the Appendix show Kendall Tau correlation of metrics over all MT systems (not including human translations).

Kendall's Tau is less sensitive to outliers, and directly measures whether metrics agree with humans when comparing pairs of systems. However, Kendall's Tau doesn't consider the differences in scores, and two metrics whose errors differ in magnitude can have the same Kendall's Tau correlation (Figure 2).

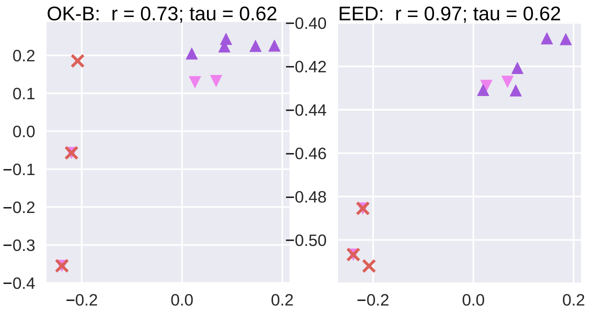


Figure 2: Scatterplots of human scores against two metrics that have the same Kendall Tau correlation with human scores, though OPENKIWI-BERT has bigger errors.

5.2 Segment and Document-Level Results

On the more fine-grained evaluation scales, PRISM and the trained neural metrics (the COMET and BLEURT family of metrics) have a better agreement with human judgements than lexical metrics

The correlations of the to-English language pairs are consistently much lower, on average, compared to that of the out-of-English language pairs. The difference could be due to the differing set of annotators: the to-English human evaluation was crowdsourced and therefore is likely to be noisier.

Finally, we find that correlations drop markedly for most language pairs if we consider only the segment/document pairs that do not contain outlier systems. We suspect that as the quality of outlier system translations is typically low, and most of the generated better-worse pairs that contain outliers can be easy for metrics. Removing these pairs would make the task a lot harder. It is also very likely that the remaining pairs of translations are noisier, which decreases metric agreement with these pairwise judgements.

The document-level correlations are typically higher than segment-level correlations. This could be due to reduced noise in human scores when averaging the scores of multiple segments. Computing metric scores over documents that contain multiple segments also helps reduce metric noise.

5.3 Reference-Based Metrics vs. Reference-Free Metrics

We have four submissions of metrics that directly compare MT outputs with the source segment: COMET-QE, OPENKIWI-BERT, OPENKIWI-XLMR, and YISI-2. Other members of the COMET family of metrics use information from both the source and reference. The remaining metrics compute scores by comparing the MT output with the reference.

While the task of comparing segments in different languages is harder than comparing segments in the same language, reference-free metrics have one advantage: they are not encumbered by reference-bias. COMET-QE is the only metric that correctly gives a high score to the human translation in German → English , and one of the few metrics that does so for English → Chinese.

This year, the reference-free metrics are highly competitive with reference-based metrics for all language-pairs. For English → Tamil, COMET-

QE which has a near perfect correlation of 0.97 even after discarding outliers. In contrast, many reference-based metrics including BLEU and chrF give really high scores to ONLINE-B, which results in low correlations.

6 Use Automatic Metrics to Detect Incorrect Human Preference

It has been argued that non-expert translators lack knowledge of translation and so might not notice subtle differences that make one translation better than another. Castilho et al. (2017) compared the evaluation of MT output of professional translators against crowd workers. Results showed that for all language pairs, the crowd workers tend to be more accepting of the MT output by giving higher fluency and adequacy scores. Toral et al. (2018) showed that the ratings acquired by professional translators show a wider gap between human and machine translations compared to judgments by non-experts. They recommend using professional linguists for MT evaluation going forward. Läubli et al. (2020) show that non-experts assess parity between human and machine translation where professional translators do not, indicating that the former neglect more subtle differences between different translation outputs. Given the previous work and the fact that the WMT human evaluation has been conducted with a mix of researchers and crowd workers, we rerun human evaluation for a subset of the submissions with professional linguists. In particular, we want to investigate if we can use the quality scores obtained by the automatic metrics to detect incorrect human ratings. We filtered out all pairs of systems where the human evaluation results disagree with all automatic metrics. Taking the metric scores as a signal, we rerun human evaluation for a subset of submissions for 2 language pairs: German→English and English→German. We hired 10 professional linguists, who rerun the source-based direct assessment human evaluation with the same document-based template that has been used for the original WMT ratings.

6.1 German→English

For German→English, we found that all automatic metrics disagree with the human evaluation results for OPPO and TOHOKU. OPPO yields a higher human rating, while all automatic metrics gave TOHOKU a higher score. To investigate which of the results to trust, we rerun the source-based direct assessment for these 2 systems with professional linguists. The results in Table 13 show that professional linguists in fact prefer the output of TOHOKU as predicted by all automatic metrics.

Evaluation	OPPO	TOHOKU
avg metric (HUMAN-A ref)	8.85	**8.95**
avg metric (Human-B ref)	10.15	**10.26**
WMT	**84.6**	81.5
z-score	**0.220**	0.179
prof. linguist	81.0	**81.7**
z-score	-0.005	**0.010**

Table 13: WMT 2020 German→English comparing the reference-based ratings acquired with crowd workers/researcher (WMT) against source-based ratings acquired with professional linguists.

6.2 English→German

For English→German, we rerun human evaluation for the top 2 ranked MT systems (based on human evaluation): OPPO, TOHOKU and the human translation HUMAN-A. The quality of human translations is usually underestimated by automatic metrics when computed with standard references. This is also visible in this year's evaluation campaign where the average metric scores of all submission for the human translation HUMAN-A is much lower when compared to the top MT submissions. To overcome this problem, Freitag et al. (2020) introduced paraphrased references that also value the translation quality of human translations and alternative (less simple/monotonic) MT output. As we can see in Table 14, the average metric scores of all submissions when computed with the paraphrased references HUMAN-P yield a much higher score for the human translation HUMAN-A when compared to all MT outputs.

The official WMT human evaluation ranked the human translation third, right behind the two MT outputs from OPPO and TOHOKU. Interestingly, based on the z-scores, WMT predicts OPPO to be of higher quality than TOHOKU which is in disagreement with most of the metric scores when calculated against both types of reference translations. Overall, the automatic metrics come to a very different ranking than the human evaluation for the top performing submissions.

Evaluation	OPPO	TOHOKU	HUMAN-A
avg metric (Human-B ref)	10.05	**10.09**	9.14
avg metric (Human-P ref)	11.93	12.07	**15.74**
WMT	87.39	**88.62**	85.10
z-score	**0.495**	0.468	0.379
prof. linguist	73.66	74.70	**84.09**
z-score	-0.051	-0.037	**0.088**

Table 14: WMT 2020 English→German comparing the source-based ratings acquired with crowd workers/researcher (WMT) against source-based ratings acquired with professional linguists.

We rerun the human evaluation with the same template, but with professional linguists. Interestingly, the human translation has been ranked first by a large margin. Furthermore, the MT output of TOHOKU has been rated as higher quality when compared to the MT output from OPPO. The results of the human evaluation with professional linguists yield a perfect correlation to the metric scores calculated with the paraphrased reference. This indicates not only the advantages of paraphrased references when scoring human translations, but also that automatic metrics can be used to identify incorrect human ratings.

7 Conclusion

This paper summarizes the results of WMT20 shared task in machine translation evaluation, the Metrics Shared Task. Participating metrics were evaluated in terms of their correlation with human judgement at the level of the whole test set (system-level evaluation), as well as at a more fine-grained level (document-level evaluation and sentences or paragraphs for segment-level evaluation). We reported scores for standard metrics requiring the reference as well metrics that compare MT output directly with the source text. For system-level, best metrics reach over 0.95 Pearson correlation or better across several language pairs. In many cases, this correlation drops considerably when the correlation is recomputed after discarding outlier systems.

Computing Pearson correlation without outliers can change the rankings of metrics, and selecting these outlier systems is not an exact science. We report results both with all systems and after discarding outliers as together, and also include Kendall Tau correlation, and hope that together, they give a more complete picture than just reporting only one of these numbers. In the end, we believe that it is impossible to adequately describe data with summary numbers, and that it's best to visualise data to understand patterns.

The results confirm the trends from previous years, namely metrics based on word or sentence-level embeddings, achieve the highest performance (Ma et al., 2018, 2019).

For some language pairs, we had two references available. On these test sets, we found that computing scores with two references rarely helped metrics achieve a higher correlation than using either reference individually. This contradicts earlier research that shows that multiple references improve correlation (Bojar et al., 2013), but is in line with more recent papers that show additional independent references might not be helpful (Freitag et al., 2020). We believe that the utility of additional independent references is dependent on the MT systems evaluated, that perhaps they are not as helpful when scoring high quality MT systems as with low/mid quality MT.

In addition to scoring MT systems, this year, we also requested scores for human reference translations. This highlighted the difference between lexical and embedding-based metrics, as lexical metrics consistently gave low scores to human translations. However, when using the English-German paraphrased references, all metrics scored the other human references above all MT systems, highlighting the advantages of using paraphrased references when scoring human translations.

In addition to human references, there are some MT systems where metrics (either the majority of metrics, or only the lexical metrics) make major errors. It remains an open question as to what it is about these systems that metrics struggle with scoring them correctly.

Compared to last year, the performance of the reference-free metrics has improved, and the correlations this year are competitive with the reference-based metrics, and in many cases, outperform BLEU. In particular, COMET-QE is good at recognising the high quality of human translations where BLEU falls short.

In terms of segment-level Kendall's τ results, the standard metrics correlations was very low for the to-English language pairs, particularly after discarding translations by outlier systems. The correlations of the out-of-English language pair are more in line with recent years, reaching a maximum of above 0.6.

It has been shown that context is really important when humans are rating MT outputs (Toral et al., 2018), and the WMT human evaluation is moving towards evaluating segments with the document context (Barrault et al., 2019). This creates a mismatch with automatic metrics, all of which, this year, score each segment independently. This year, we introduce document-level evaluation of metrics. When computing document-level scores, some metrics from the COMET family include document context when computing segment scores within the document. All other metrics included in this year's evaluation either use the average of the segment scores or compute the document score based on statistics computed independently for each segment. In the future, we hope to see more metrics that consider broader context when evaluating translations at all three levels.

For this year, we are unable to draw any meaningful conclusions from the document-level evaluation task, as it is hard to tease apart the influence of noise in the ground truth, inadequate segment-level translations and inadequate translation in context of the document.

We believe that the noise in the DARR judgements is a big factor in the low correlations in the to-English language pairs. We need further research into understanding the factors that contribute to the Kendall Tau scores and how much we can trust these results.

There are shortcomings in the methods used to evaluate metrics at the system-, document-, and segment-level, and we believe that improving methods for evaluating and analysing automatic metrics is a rich area for future research.

Finally, we assume that any discrepancies between metrics and WMT manual evaluation is a metric error, and we acknowledge that this might not be true in all cases. There is always scope for improvement in human evaluation methodology, and the best practice recommendations for human evaluation are always evolving.

We rerun human evaluation by using the same template as the WMT evaluation, but switching the rater pool from non-experts to professional linguists for a subset of translations where all metrics disagree with the WMT human evaluation. This experiment revealed a new use case of automatic metrics and demonstrated that automatic metrics can be used to identify bad ratings in human evaluations. The new obtained ratings were in line with the scores suggested by the automatic metrics and also confirmed the higher translation quality of human translations when compared to MT output.

In this paper, we looked at how outliers influence metric evaluation, and we wonder how the presence of these systems influence DA annotations. In a perfect world, annotators score each translation on its own merits without being influenced by previous instances. In this world, given the presence of much worse translations, do annotators assign high scores to the remaining translations that look relatively better? Does an MT system receive an unfair advantage if it is consistently scored alongside a low-scoring outlier? And does standardising the scores of individual annotators exacerbate this issue? These and other research questions remain open this year, keeping the WMT tasks increasingly interesting as MT systems are getting closer to human performance.

Acknowledgments

Results in this shared task would not be possible without tight collaboration with organizers of the WMT News Translation Task.

We are grateful to Google for sponsoring the evaluation of selected MT systems by professional linguists. We thank all participants of the task, particularly Weiyu Wang for pointing out some inconsistencies with the original release of inputs, Jackie Lo, Ricardo Rei, and Craig Stewart for some helpful feedback and Thibault Sellam for also submitting the scores of finetuned BERT on our request.

Nitika Mathur is supported by the Australian Research Council. Ondřej Bojar would like to acknowledge the support from the Czech Science Foundation (grant n. 19-26934X, NEUREM3).

References

Manish Shrivastava Ananya Mukherjee, Hema Ala and Dipti Misra Sharma. 2020. Mee: An automatic metric for evaluation using embeddings for machine translation. (in press).

Loïc Barrault, Ondřej Bojar, Marta R Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, et al. 2019. Findings of the 2019 Conference on Machine Translation (WMT19). In

Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1), pages 1–61.

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 *Proceedings of the Fifth Conference on Machine Translation*, Online. Association for Computational Linguistics.

Rachel Bawden, Biao Zhang, Andre Tättar, and Matt Post. 2020. ParBLEU: Augmenting metrics with automatic paraphrases for the WMT'20 metrics shared task. In *Proceedings of the Fifth Conference on Machine Translation (Volume 2: Shared Task Papers)*, Online. Association for Computational Linguistics.

Ondřej Bojar, Matouš Macháček, Aleš Tamchyna, and Daniel Zeman. 2013. Scratching the surface of possible translations. In *Text, Speech, and Dialogue*, pages 465–474, Berlin, Heidelberg. Springer Berlin Heidelberg.

Ondřej Bojar, Yvette Graham, and Amir Kamran. 2017. Results of the wmt17 metrics shared task. In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 489–513, Copenhagen, Denmark.

Ondřej Bojar, Yvette Graham, Amir Kamran, and Miloš Stanojević. 2016. Results of the WMT16 Metrics Shared Task. In *Proceedings of the First Conference on Machine Translation*, pages 199–231, Berlin, Germany.

Sheila Castilho, Joss Moorkens, Federico Gaspari, Andy Way, Panayota Georgakopoulou, Maria Gialama, Vilelmini Sosoni, and Rico Sennrich. 2017. Crowdsourcing for nmt evaluation: Professional translators versus the crowd. *Translating and the Computer*, 39.

Qian Chen, Xiaodan Zhu, Zhenhua Ling, Si Wei, Hui Jiang, and Diana Inkpen. 2017. Enhanced LSTM for natural language inference. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, Vancouver, Canada.

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

Markus Dreyer and Daniel Marcu. 2012. HyTER: Meaning-equivalent semantics for translation evaluation. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 162–171, Montréal, Canada. Association for Computational Linguistics.

Markus Freitag, David Grangier, and Isaac Caswell. 2020. BLEU might be guilty but references are not innocent. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 61–71, Online. Association for Computational Linguistics.

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

Yvette Graham, Timothy Baldwin, and Nitika Mathur. 2015. Accurate evaluation of segment-level machine translation metrics. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1183–1191, Denver, Colorado. Association for Computational Linguistics.

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.

Yvette Graham, Qingsong Ma, Timothy Baldwin, Qun Liu, Carla Parra, and Carolina Scarton. 2017. Improving evaluation of document-level machine translation quality estimation. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 356–361, Valencia, Spain. Association for Computational Linguistics.

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

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

Fabio Kepler, Jonay Trénous, Marcos Treviso, Miguel Vera, and André F. T. Martins. 2019. OpenKiwi: An open source framework for quality estimation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 117–122, Florence, Italy. Association for Computational Linguistics.

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

Philipp Koehn, Vishrav Chaudhary, Ahmed El-Kishky, Naman Goyal, Peng-Jen Chen, and Francisco Guzmán. 2020. Findings of the wmt 2020 shared task on parallel corpus filtering and alignment. In *Proceedings of the Fifth Conference on Machine Translation*, Online. Association for Computational Linguistics.

Samuel Läubli, Sheila Castilho, Graham Neubig, Rico Sennrich, Qinlan Shen, and Antonio Toral. 2020. A set of recommendations for assessing human–machine parity in language translation. *Journal of Artificial Intelligence Research*, 67:653–672.

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

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

Chi-kiu Lo. 2020. Extended study on using pretrained language models and YiSi-1 for machine translation evaluation. In *Proceedings of the Fifth Conference on Machine Translation: Shared Task Papers*.

Chi-kiu Lo and Samuel Larkin. 2020. Machine translation reference-less evaluation using YiSi-2 with bilingual mappings of massive multilingual language model. In *Proceedings of the Fifth Conference on Machine Translation: Shared Task Papers*.

Qingsong Ma, Ondřej Bojar, and Yvette Graham. 2018. Results of the WMT18 metrics shared task: Both characters and embeddings achieve good performance. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 671–688, Belgium, Brussels. Association for Computational Linguistics.

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

Matouš Macháček and Ondřej Bojar. 2014. Results of the WMT14 metrics shared task. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 293–301, Baltimore, Maryland, USA. Association for Computational Linguistics.

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

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, Online. Association for Computational Linguistics.

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

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

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

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

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

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020a. 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.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020b. Unbabel's participation in the wmt20 metrics shared task. In *Proceedings of the Fifth Conference on Machine Translation*, Online. Association for Computational Linguistics.

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

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

Thibault Sellam, Amy Pu, Hyung Won Chung, Sebastian Gehrmann, Qijun Tan, Markus Freitag, Dipanjan Das, and Ankur Parikh. 2020b. Learning to evaluate translation beyond english: Bleurt submissions to the wmt metrics 2020 shared task. In *Proceedings of the Fifth Conference on Machine Translation*, Online. Association for Computational Linguistics.

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

Peter Stanchev, Weiyue Wang, and Hermann Ney. 2019. Eed: Extended edit distance measure for machine translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 514–520, Florence, Italy. Association for Computational Linguistics.

Brian Thompson and Matt Post. 2020. Automatic machine translation evaluation in many languages via zero-shot paraphrasing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online. Association for Computational Linguistics.

Antonio Toral, Sheila Castilho, Ke Hu, and Andy Way. 2018. Attaining the Unattainable? Reassessing Claims of Human Parity in Neural Machine Translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 113–123, Belgium, Brussels. Association for Computational Linguistics.

Tereza Vojtěchová, Michal Novák, Miloš Klouček, and Ondřej Bojar. 2019. SAO WMT19 Test Suite: Machine Translation of Audit Reports. In *Proceedings of the Fourth Conference on Machine Translation*, Florence, Italy. Association for Computational Linguistics.

Weiyue Wang, Jan-Thorsten Peter, Hendrik Rosendahl, and Hermann Ney. 2016. CharacTer: Translation edit rate on character level. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 505–510, Berlin, Germany. Association for Computational Linguistics.

Evan James Williams. 1959. *Regression analysis*. wiley.

Jin Xu, Yinuo Guo, and Junfeng Hu. 2020. Incorporate semantic structures into machine translation evaluation via ucca. In *Proceedings of the Fifth Conference on Machine Translation*, Online. Association for Computational Linguistics.

A List of Outliers

lp	Outliers
cs-en	zlabs-nlp.1149, CUNI-DocTransformer.1457
de-en	yolo.1052, zlabs-nlp.1153, WMTBiomedBaseline.387
iu-en	NiuTrans.1206, Facebook_AI.729
ja-en	Online-G.1564, zlabs-nlp.66, Online-Z.1640
pl-en	zlabs-nlp.1162
ru-en	zlabs-nlp.1164
ta-en	Online-G.1568, TALP_UPC.192
zh-en	WMTBiomedBaseline.183
en-cs	zlabs-nlp.1151, Online-G.1555
en-de	zlabs-nlp.179, WMTBiomedBaseline.388, Online-G.1556
en-iu_news	UEDIN.1281, OPPO.722, UQAM_TanLe.521
en-iu_full	UEDIN.1281, OPPO.722, UQAM_TanLe.521
en-iu	UEDIN.1281, OPPO.722, UQAM_TanLe.521
en-pl	Online-Z.1634, zlabs-nlp.180, Online-A.1576
en-ta	TALP_UPC.1049, SJTU-NICT.386, Online-G.1561

Table 15: List of all MT systems that we consider as outliers

B Scatterplots

Here we show scatterplots of human and metric scores of selected metrics.

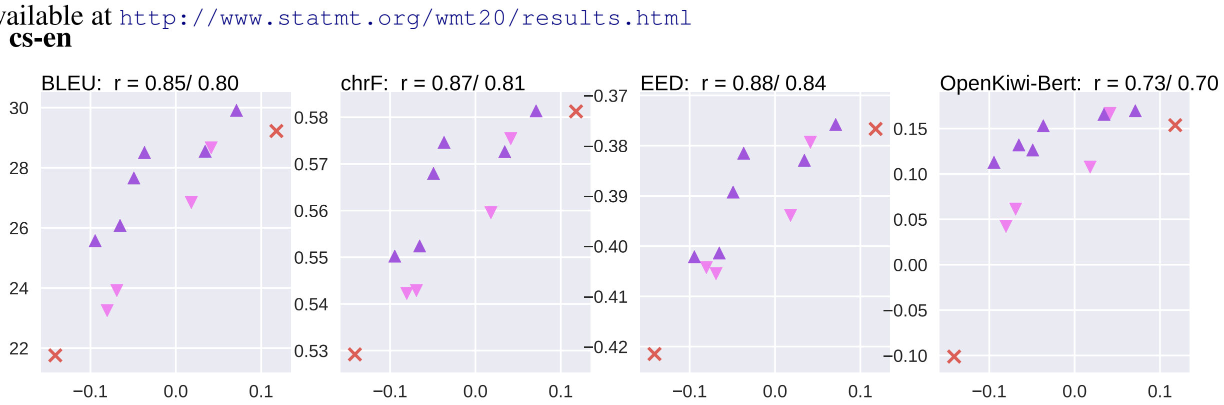
We report the correlation of each metric with human scores on all systems as well as all systems minus the outliers. Note that we do not exclude human translations when computing these correlations.

In the following scatterplots, the violet triangles indicate individual indicate MT system submissions by researchers and pink downward triangles are online systems. [8] The red crosses are outlier systems.

The black diamonds are human translations. For newstest2020 reference set, this is the Human-A translation, and for newstestB2020 reference set, this is the Human-B translation. The plots for English → German have two human translations included, and we annotate the label in the plot. In many cases, metric errors scoring these translations stand out.

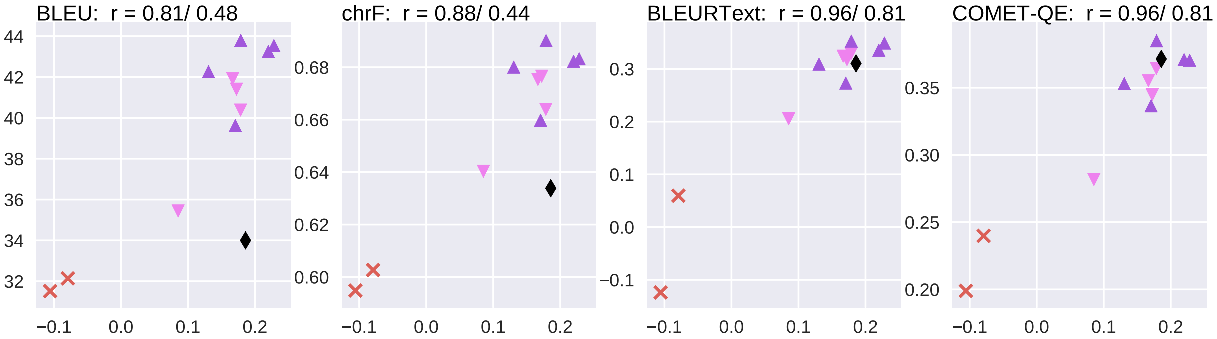
Metric scores of MT systems with multiple references does not deviate from the scores of either reference. So we do not include the scatterplots of the other reference sets unless a human translation is included (which is interesting).

We will have scatterplots for all metrics over all reference sets in the metrics package to be made available at http://www.statmt.org/wmt20/results.html



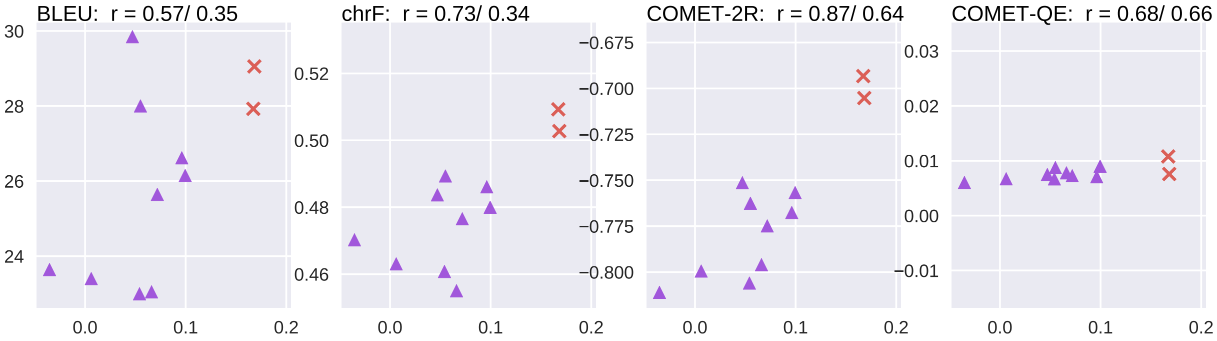
[8] We distinguish between the two in these scatterplots as we notice that metrics often make errors when scoring online systems.

de-en newstest2020[a]

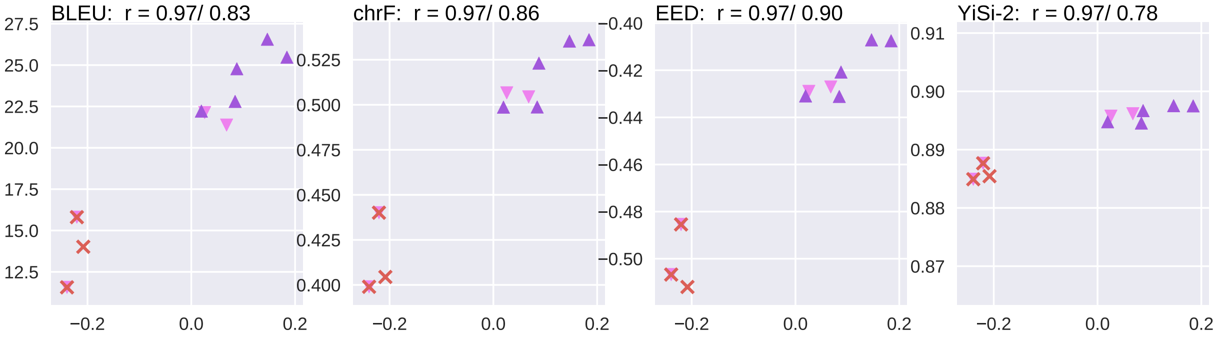


[a] Including the YOLO.1052 system, which has an extremely low quality, would make it hard to distinguish between the rest of the systems, so these plots exclude the system.

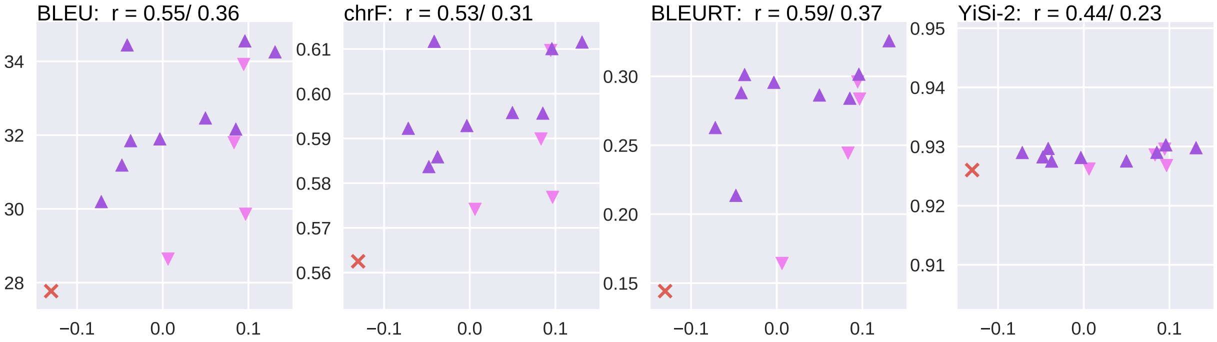
iu-en



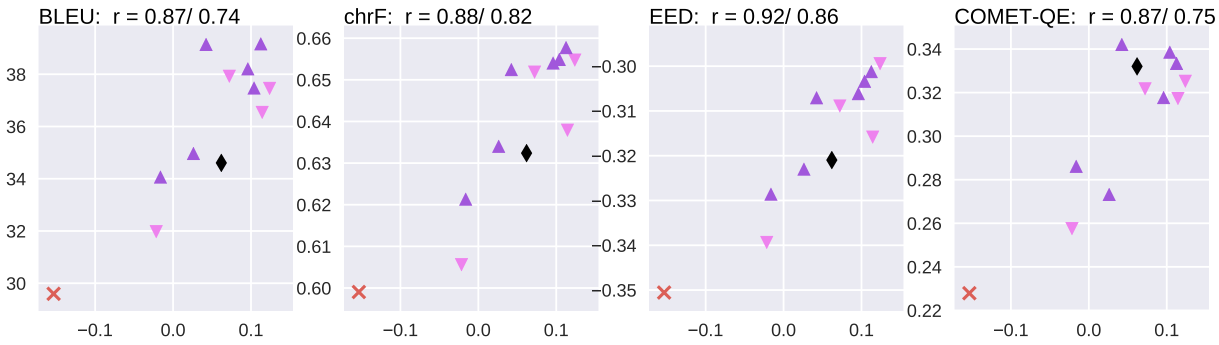
ja-en



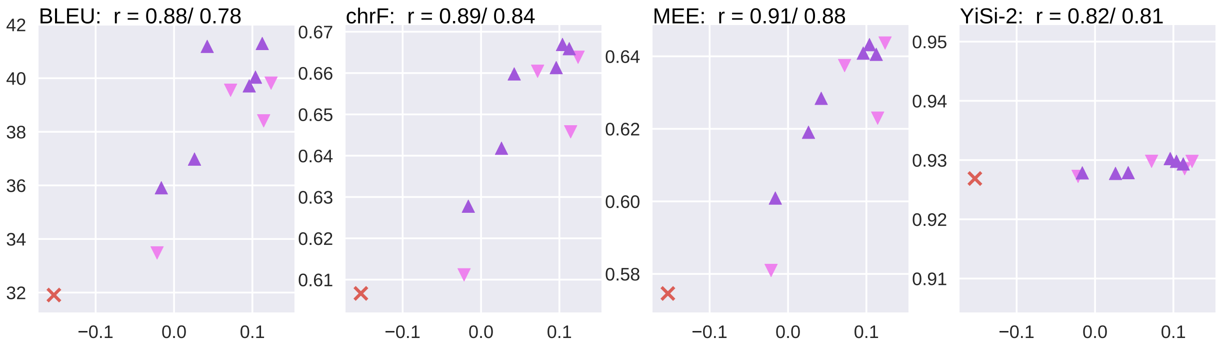
pl-en



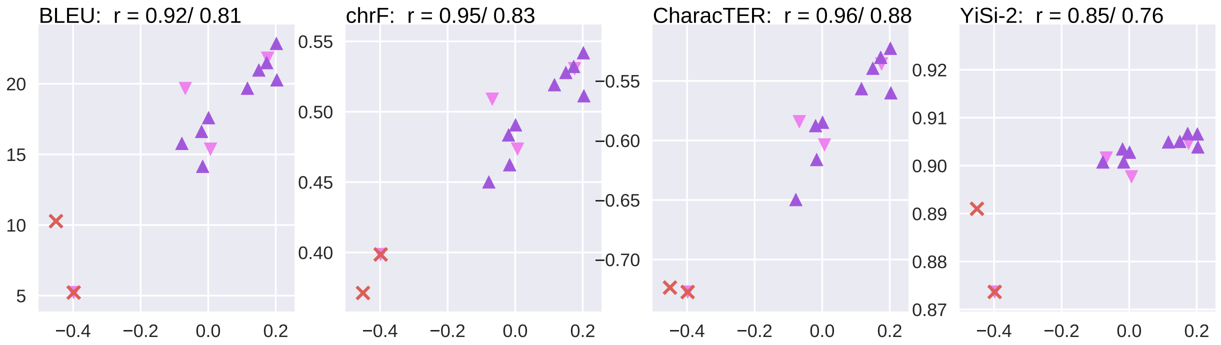
ru-en newstest2020



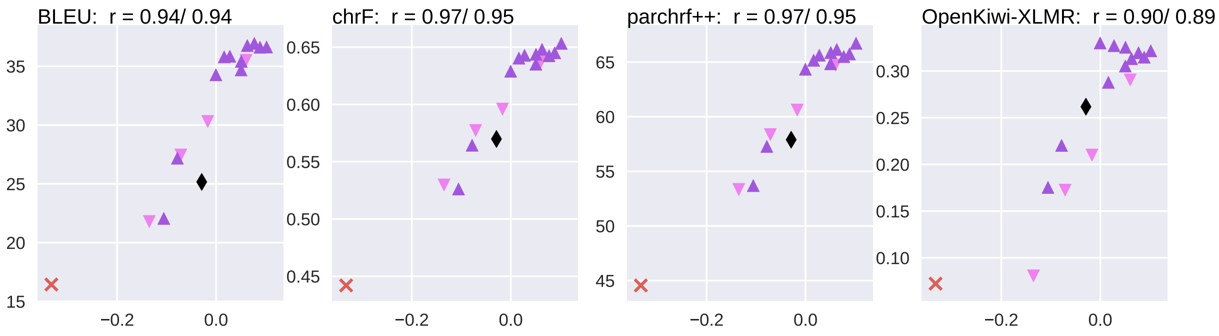
ru-en newstestB2020



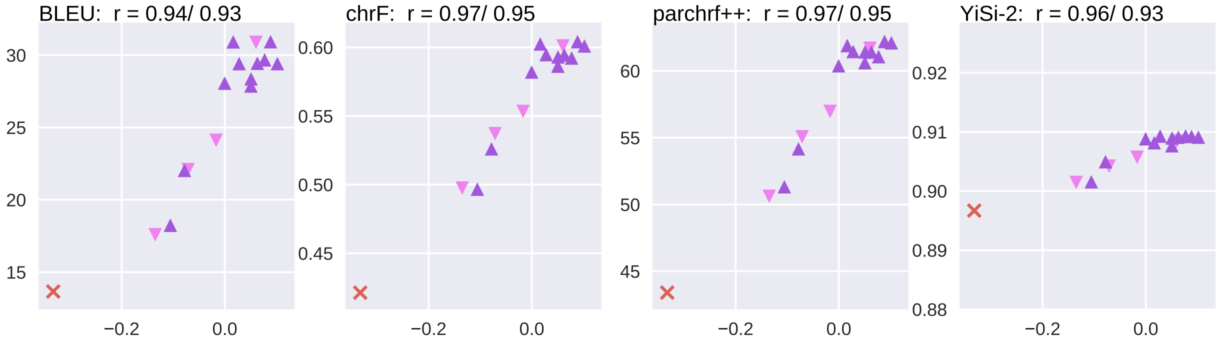
ta-en



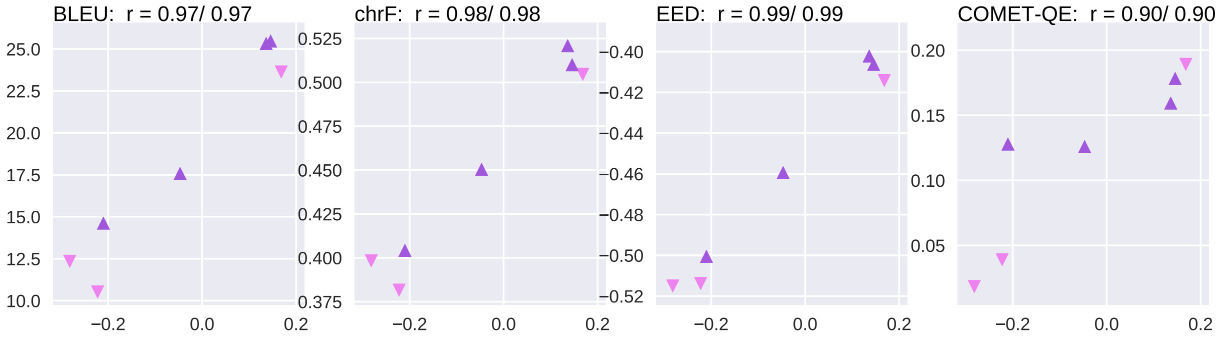
zh-en newstest2020



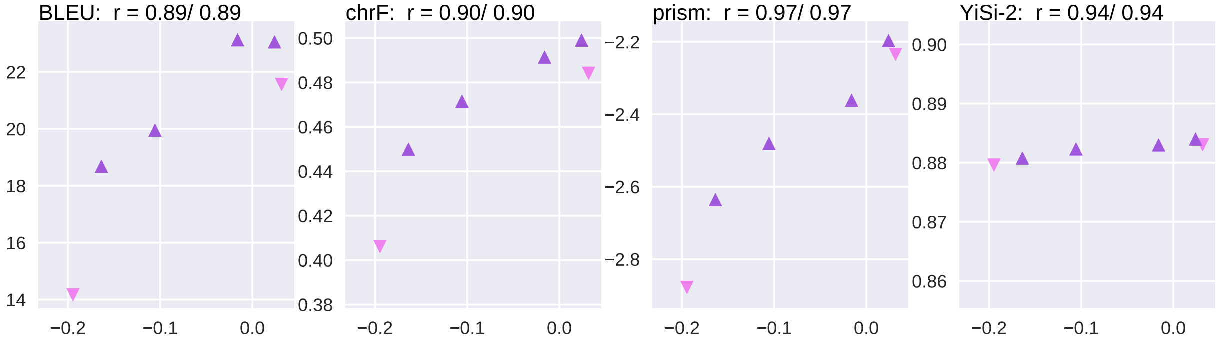
zh-en newstestB2020


BLEU: r = 0.94/ 0.93
chrF: r = 0.97/ 0.95
parchrf++: r = 0.97/ 0.95
YiSi-2: r = 0.96/ 0.93

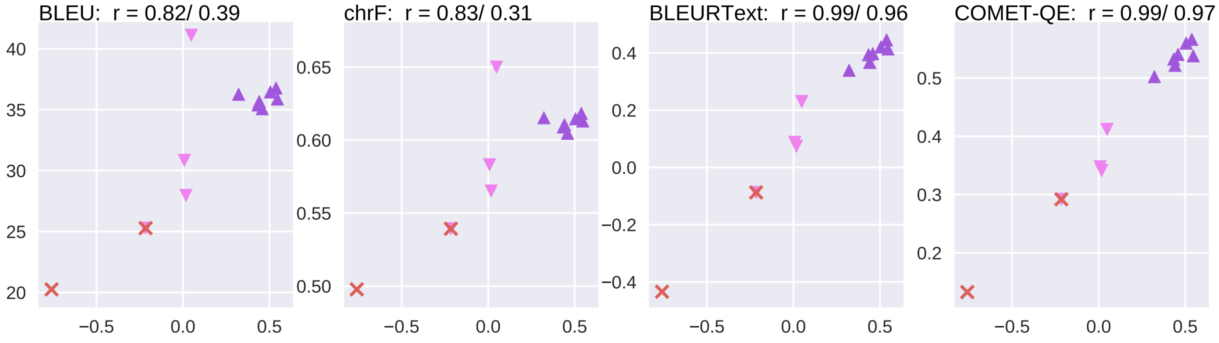
km-en


BLEU: r = 0.97/ 0.97
chrF: r = 0.98/ 0.98
EED: r = 0.99/ 0.99
COMET-QE: r = 0.90/ 0.90

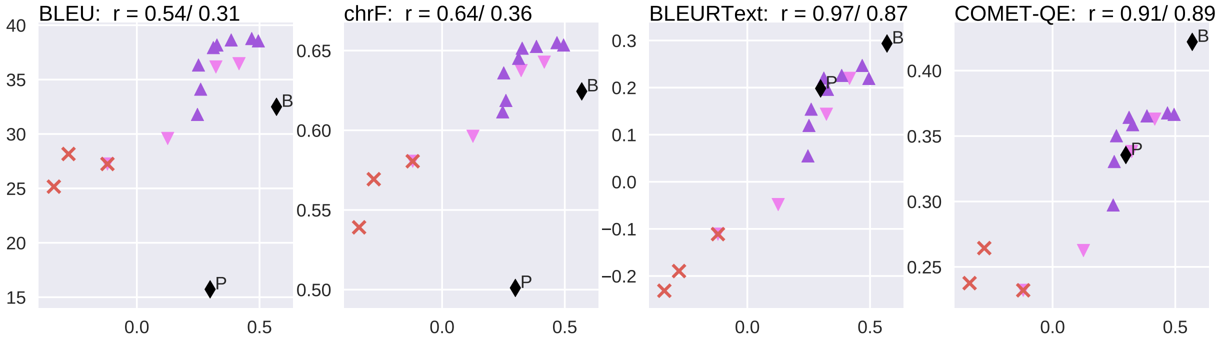
ps-en


BLEU: r = 0.89/ 0.89
chrF: r = 0.90/ 0.90
prism: r = 0.97/ 0.97
YiSi-2: r = 0.94/ 0.94

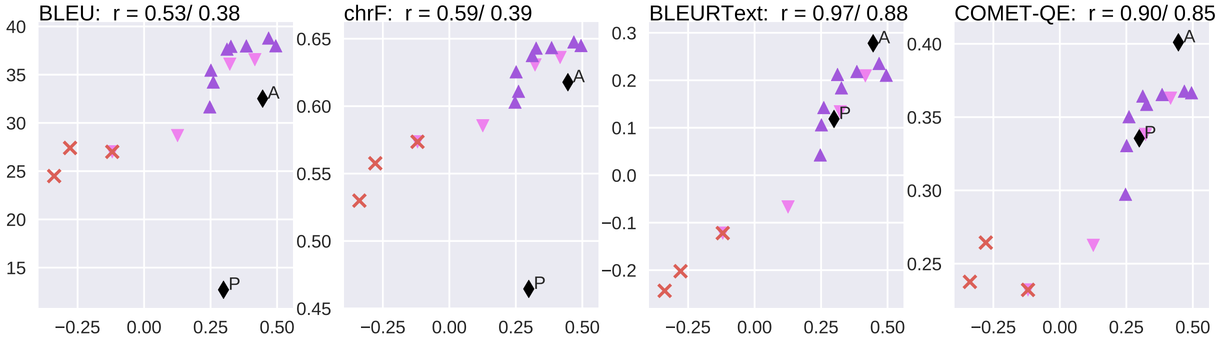
en-cs


BLEU: r = 0.82/ 0.39
chrF: r = 0.83/ 0.31
BLEURText: r = 0.99/ 0.96
COMET-QE: r = 0.99/ 0.97

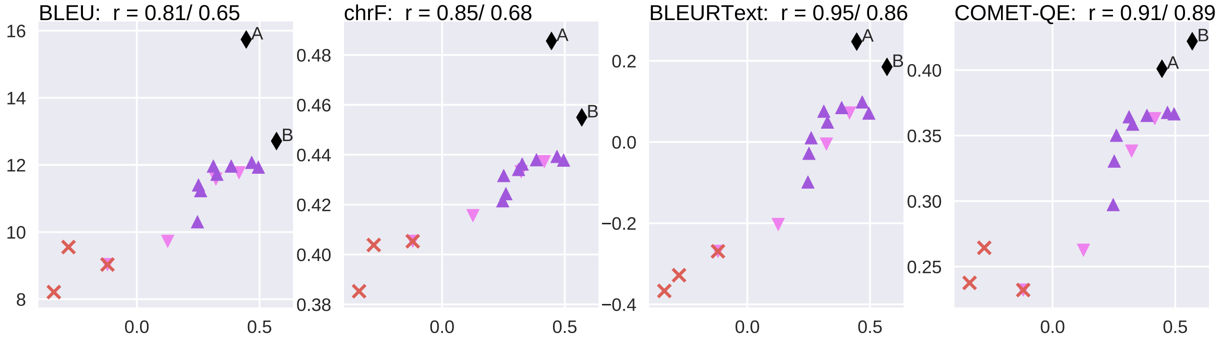
en-de newstest2020



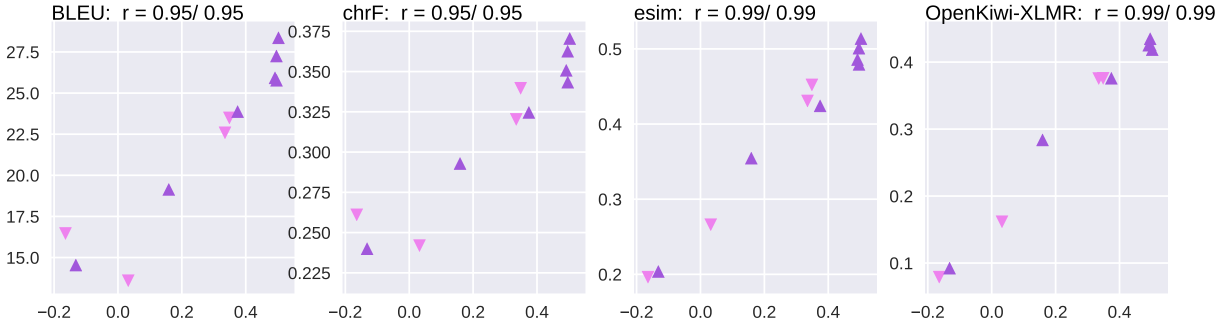
en-de newstestB2020



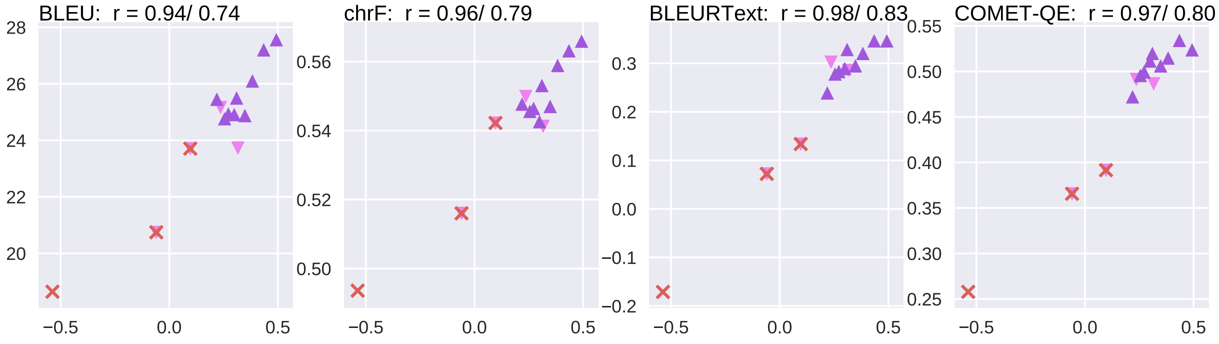
en-de newstestP2020



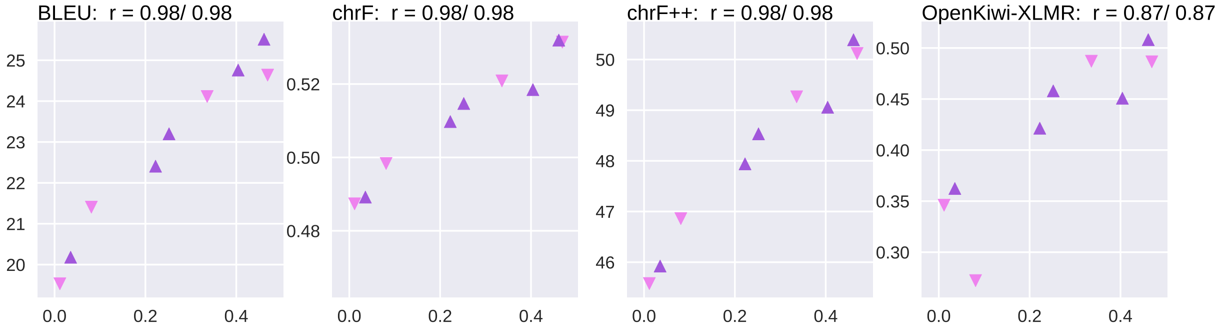
en-ja



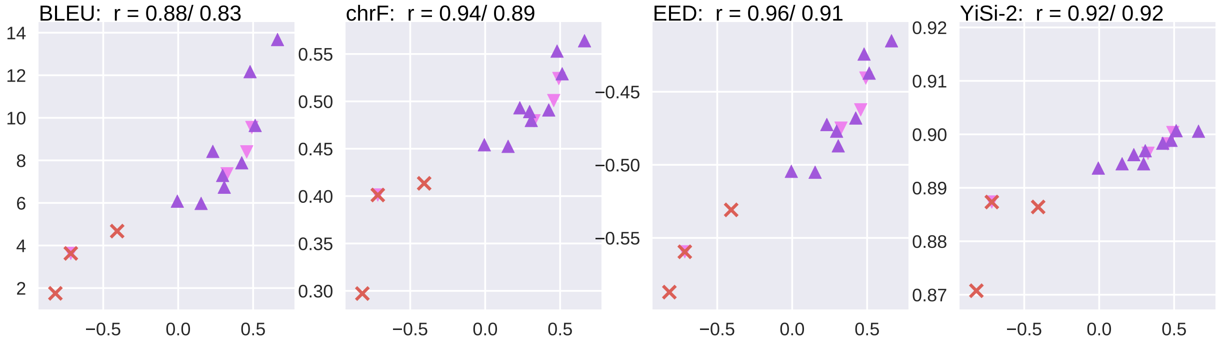
en-pl



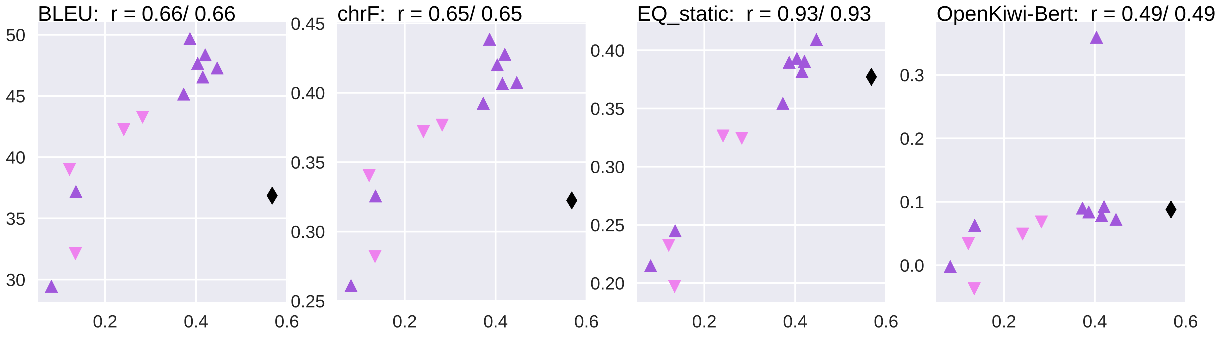
en-ru



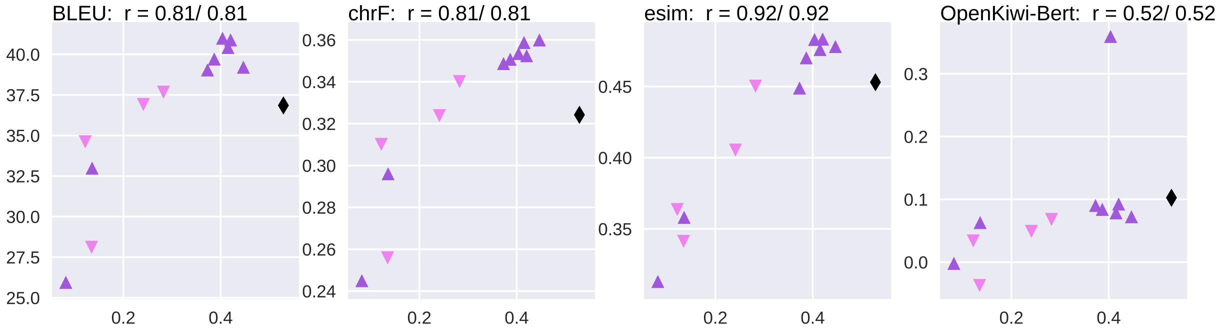
en-ta



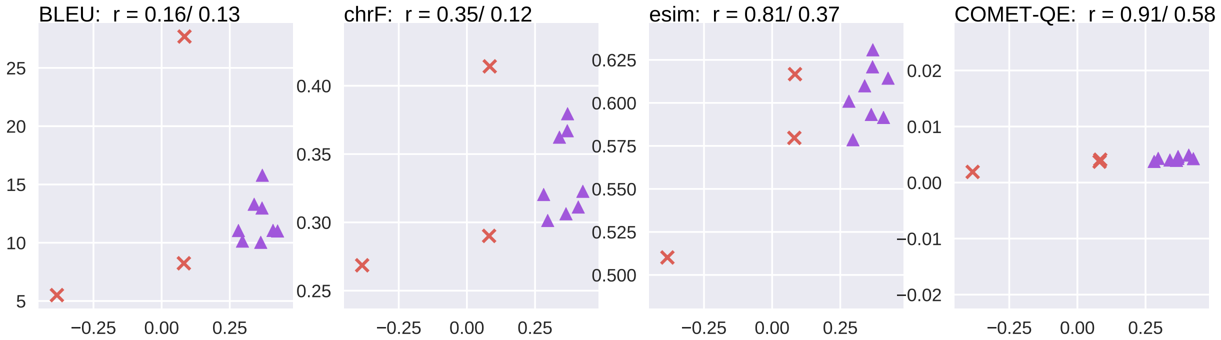
en-zh newstest2020



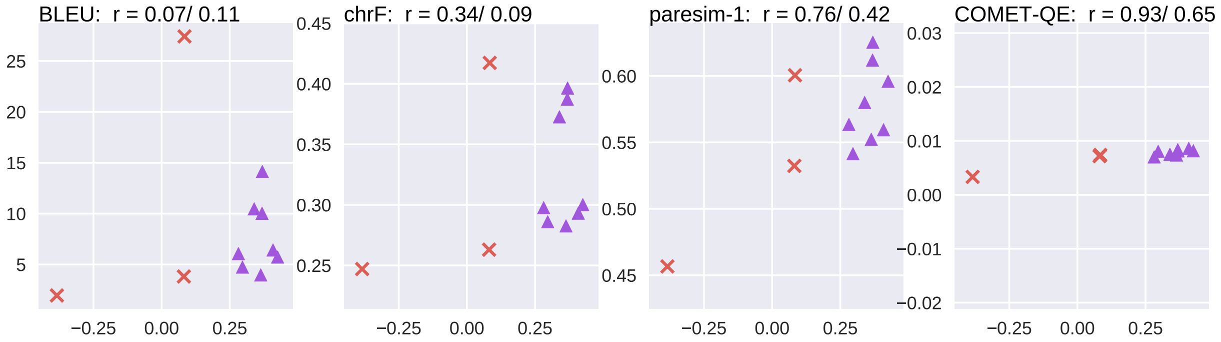
en-zh newstestB2020



en-iu Full test set



en-zh Out of domain (News) subset



C Additional System-level Results

We also report Kendall Tau correlation of metrics at the system level.

	cs-en	de-en	ja-en	pl-en	ru-en	ta-en	zh-en	iu-en	km-en	ps-en
	12	12	10	14	11	14	16	11	7	6
HUMAN_RAW	0.727	0.758	0.778	0.429	0.673	0.604	0.650	0.891	0.905	1.000
sentBLEU	0.788	0.758	0.733	0.297	0.564	0.692	0.850	0.455	0.619	0.600
BLEU	0.848	0.697	0.778	0.407	0.455	0.692	0.833	0.309	0.714	0.600
TER	0.758	0.788	0.689	0.287	0.600	0.780	0.800	0.514	0.878	0.867
chrF++	0.818	0.697	0.778	0.407	0.673	0.714	0.850	0.418	0.619	0.733
chrF	0.818	0.727	0.822	0.363	0.709	0.714	0.833	0.418	0.619	0.733
parbleu	0.809	0.779	0.778	0.420	0.491	0.685	0.807	0.404	0.714	0.867
parchrF++	0.818	0.727	0.822	0.407	0.709	0.714	0.817	0.491	0.619	0.733
CharacTER	0.758	0.758	0.822	0.341	0.745	0.692	0.800	0.527	0.810	0.733
EED	0.788	0.727	0.733	0.297	0.782	0.758	0.833	0.636	0.714	0.733
YiSi-0	0.758	0.758	0.689	0.231	0.782	0.802	0.833	0.600	0.714	0.733
SWSS+METEOR	–	–	0.822	0.341	0.818	0.736	0.817	0.491	0.714	0.733
MEE	0.758	0.697	0.867	0.363	0.709	0.692	0.783	0.636	0.714	0.733
prism	0.758	0.727	0.867	0.341	0.564	0.648	0.800	0.673	0.714	0.867
YiSi-1	0.758	0.758	0.778	0.451	0.564	0.692	0.817	0.673	1.000	0.867
BERT-base-L2	0.758	0.848	0.822	0.407	0.491	0.604	0.633	0.564	1.000	0.867
BERT-large-L2	0.758	0.848	0.867	0.341	0.564	0.626	0.700	0.527	1.000	0.867
mBERT-L2	0.758	0.818	0.822	0.429	0.564	0.604	0.750	0.673	1.000	0.867
BLEURT	0.758	0.788	0.822	0.407	0.600	0.604	0.650	0.527	1.000	0.867
BLEURT-extended	0.727	0.848	0.778	0.341	0.455	0.582	0.617	0.527	0.905	0.867
esim	0.727	0.848	0.822	0.451	0.491	0.670	0.717	0.636	1.000	0.867
paresim-1	0.727	0.879	0.822	0.451	0.491	0.670	0.700	0.636	1.000	0.867
COMET	0.727	0.758	0.778	0.407	0.564	0.626	0.733	0.636	1.000	0.867
COMET-2R	0.727	0.788	0.778	0.451	0.527	0.582	0.717	0.600	1.000	0.867
COMET-HTER	0.667	0.788	0.822	0.275	0.491	0.604	0.533	0.564	1.000	0.867
COMET-MQM	0.667	0.727	0.822	0.275	0.455	0.582	0.517	0.636	1.000	1.000
COMET-Rank	0.576	0.727	0.822	0.341	0.455	0.626	0.650	0.309	0.810	1.000
BAQ_dyn	–	–	–	–	–	–	0.817	–	–	–
BAQ_static	–	–	–	–	–	–	0.867	–	–	–
COMET-QE	0.697	0.788	0.778	0.297	0.455	0.516	0.550	0.491	0.905	0.733
OpenKiwi-Bert	0.697	0.667	0.733	0.187	0.455	0.429	0.450	-0.055	0.714	0.467
OpenKiwi-XLMR	0.727	0.636	0.822	0.275	0.418	0.560	0.567	0.018	1.000	0.867
YiSi-2	0.576	0.515	0.778	0.319	0.527	0.582	0.750	0.491	0.810	0.867

Table 16: Kendall Tau correlation of system-level metrics with DA human assessment for all MT systems not including Human translations. In addition to the metrics, we also include raw human scores where annotator scores were not standardised.

	en-cs 12	en-de 14	en-ja 11	en-pl 14	en-ru 9	en-ta 15	en-zh 12	en-iu_full 11	en-iu_news 11
HUMAN_RAW	1.000	0.868	0.964	0.846	0.778	0.810	0.818	0.600	0.600
sentBLEU	0.515	0.802	0.855	0.604	0.944	0.867	0.727	0.236	0.273
BLEU	0.515	0.802	0.818	0.582	0.889	0.829	0.727	0.236	0.236
TER	0.515	0.824	0.018	0.641	0.556	0.752	0.242	0.309	0.309
chrF++	0.485	0.868	0.782	0.604	0.889	0.829	0.727	0.309	0.309
chrF	0.485	0.868	0.818	0.604	0.889	0.810	0.727	0.345	0.309
parbleu	0.504	0.736	0.611	0.633	0.761	0.842	0.718	0.404	0.345
parchrf++	0.515	0.846	0.818	0.670	0.889	–	0.727	–	–
CharacTER	0.515	0.890	0.782	0.560	0.944	0.771	0.697	0.236	0.345
EED	0.545	0.868	0.782	0.604	0.833	0.867	0.727	0.273	0.273
YiSi-0	0.545	0.846	0.818	0.604	0.944	0.790	0.515	0.236	0.345
MEE	0.576	0.802	–	0.582	0.667	0.829	–	0.273	0.382
prism	0.818	0.868	0.818	0.670	0.611	0.562	0.576	0.418	0.600
YiSi-1	0.606	0.868	0.782	0.626	0.833	0.810	0.758	0.091	0.273
Yisi-combi	–	0.824	–	–	–	–	–	–	–
bleurt-Yisi-combi	–	0.824	–	–	–	–	–	–	–
mBERT-L2	0.788	0.846	0.782	0.736	0.778	0.752	0.909	–	–
BLEURT-extended	0.879	0.802	0.782	0.780	0.833	0.771	0.848	0.382	0.345
esim	0.606	0.912	0.855	0.692	0.833	0.752	0.788	0.382	0.455
paresim-1	0.667	0.890	0.818	0.692	0.833	0.752	0.818	0.382	0.455
COMET	0.909	0.846	0.745	0.736	0.722	0.771	0.606	0.382	0.382
COMET-2R	0.909	0.890	0.891	0.714	0.611	0.790	0.606	0.309	0.418
COMET-HTER	0.909	0.802	0.818	0.736	0.667	0.619	0.576	0.491	0.491
COMET-MQM	0.909	0.802	0.818	0.736	0.667	0.619	0.545	0.527	0.455
COMET-Rank	0.848	0.780	0.782	0.692	0.556	0.524	0.515	0.127	0.345
BAQ_dyn	–	–	–	–	–	–	0.697	–	–
BAQ_static	–	–	–	–	–	–	0.788	–	–
EQ_dyn	–	–	–	–	–	–	0.727	–	–
EQ_static	–	–	–	–	–	–	0.818	–	–
COMET-QE	0.848	0.802	0.709	0.802	0.667	0.543	0.576	0.600	0.673
OpenKiwi-Bert	0.758	0.780	0.236	0.538	0.722	0.314	0.606	-0.273	0.200
OpenKiwi-XLMR	0.909	0.780	0.818	0.692	0.667	0.657	0.545	0.018	0.200
YiSi-2	0.485	0.582	0.527	0.077	0.444	0.886	0.121	0.309	0.455

Table 17: Kendall Tau correlation of out-of-English system-level metrics with DA human assessment for all MT systems not including Human translations. In addition to the metrics, we also include raw human scores where annotator scores were not standardised.