Results of the WMT18 Metrics Shared Task: Both characters and embeddings achieve good performance

Qingsong Ma
Tencent-MIG
AI Evaluation & Test Lab
qingsong.mqs@gmail.com

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

Yvette Graham
Dublin City University
ADAPT
graham.yvette@gmail.com

Abstract

This paper presents the results of the WMT18 Metrics Shared Task. We asked participants of this task to score the outputs of the MT systems involved in the WMT18 News Translation Task with automatic metrics. We collected scores of 10 metrics and 8 research groups. In addition to that, we computed scores of 8 standard metrics (BLEU, SentBLEU, chrF, NIST, WER, PER, TER and CDER) as baselines. The collected scores were evaluated in terms of system-level correlation (how well each metric's scores correlate with WMT18 official manual ranking of systems) and in terms of segment-level correlation (how often a metric agrees with humans in judging the quality of a particular sentence relative to alternate outputs). This year, we employ a single kind of manual evaluation: direct assessment (DA).

1 Introduction

Accurate evaluation of machine translation (MT) is important for measuring improvements in system performance. Human evaluation can be costly and time consuming, and it is not always available for the language pair of interest. Automatic metrics can be employed as a substitute for human evaluation in such cases, metrics that aim to measure improvements to systems quickly and at no cost to developers. In the usual set-up, an automatic metric carries out a comparison of MT system output translations and human-produced reference translations to produce a single overall score for the system.[1] Since there exists a large number of possible approaches to producing quality scores for translations, it is sensible to carry out a meta-evaluation of metrics with the aim to estimate their accuracy as a substitute for human assessment of translation quality. The Metrics Shared Task[2] of WMT annually evaluates the performance of automatic machine translation metrics in their ability to provide a substitute for human assessment of translation quality.

Again, we keep the two main types of metric evaluation unchanged from the previous years. In *system-level* evaluation, each metric provides a quality score for the whole translated test set (usually a set of documents, in fact). In *segment-level* evaluation, a score is assigned by a given metric to every individual sentence.

The underlying texts and MT systems come from the News Translation Task (Bojar et al., 2018, denoted as Findings 2018 in the following). The texts were drawn from the news domain and involve translations to/from Chinese (zh), Czech (cs), German (de), Estonian (et), Finnish (fi), Russian (ru), and Turkish (tr), each paired with English, making a total of 14 language pairs.

A single form of golden truth of translation quality judgement is used this year:

- In *Direct Assessment* (DA) (Graham et al., 2016), humans assess the quality of a given MT output translation by comparison with a reference translation (as opposed to the source, or source and reference). DA is the new standard used in

[1] The availability of a reference translation is the key difference between our task and *MT quality estimation*, where no reference is assumed.

[2] `http://www.statmt.org/wmt18/metrics-task.html`, starting with Koehn and Monz (2006) up to Bojar et al. (2017)

Proceedings of the Third Conference on Machine Translation (WMT), Volume 2: Shared Task Papers, pages 682–701
Brussels, Belgium, October 31 - November 1, 2018. ©2018 Association for Computational Linguistics
https://doi.org/10.18653/v1/W18-6451

WMT News Translation Task evaluation, requiring only monolingual evaluators.

As in last year's evaluation, the official method of manual evaluation of MT outputs is no longer "relative ranking" (RR, evaluating up to five system outputs on an annotation screen relative to each other) as this was changed in 2017 to DA. For system-level evaluation, we thus use the Pearson correlation r of automatic metrics with DA scores. For segment-level evaluation, we re-interpret DA judgements as relative comparisons and use Kendall's τ as a substitute, see below for details and references.

Section 2 describes our datasets, i.e. the sets of underlying sentences, system outputs, human judgements of translation quality and also participating metrics. Sections 3.1 and 3.2 then provide the results of system and segment-level metric evaluation, respectively. We discuss the results in Section 4.

2 Data

This year, we provided the task participants with one test set along with reference translations and outputs of MT systems. Participants were free to choose which language pairs they wanted to participate and whether they reported system-level, segment-level scores or both.

2.1 Test Sets

We use the following test set, i.e. a set of source sentences and reference translations:

newstest2018 is the test set used in WMT18 News Translation Task (see Findings 2018), with approximately 3,000 sentences for each translation direction (except Chinese and Estonian which have 3,981 and 2,000 sentences, resp.). newstest2018 includes a single reference translation for each direction.

2.2 Translation Systems

The results of the Metrics Task are likely affected by the actual set of MT systems participating in a given translation direction. For instance, if all of the systems perform similarly, it will be more difficult, even for humans, to distinguish between the quality of translations. If the task includes a wide range of systems of varying quality, however, or systems are quite different in nature, this could in some way make the task easier for metrics, with metrics that are more sensitive to certain aspects of MT output performing better.

This year, the MT systems included in the Metrics Task were:

News Task Systems are machine translation systems participating in the WMT18 News Translation Task (see Findings 2018).[3]

Hybrid Systems are created automatically with the aim of providing a larger set of systems against which to evaluate metrics, as in Graham and Liu (2016). Hybrid systems were created for newstest2018 by randomly selecting a pair of MT systems from all systems taking part in that language pair and producing a single output document by randomly selecting sentences from either of the two systems. In short, we create 10K hybrid MT systems for each language pair.

Excluding the hybrid systems, we ended up with 149 systems across 14 language pairs.

2.3 Manual MT Quality Judgments

Direct Assessment (DA) was employed as the "golden truth" to evaluate metrics again this year. The details of this method of human evaluation is provided in two sections for system-level evaluation (Section 2.3.1) and segment-level evaluation (Section 2.3.2).

The DA manual judgements were provided by MT researchers taking part in WMT tasks, a number of in-house human evaluators at Amazon and crowd-sourced workers on Amazon Mechanical Turk.[4] Only judgements from workers who passed DA's quality control mechanism were included in the final datasets used to compute system and segment-level scores employed as a gold standard in the Metrics Task.

[3] One system for tr-en was unfortunately omitted from the first run of human evaluation in the News Translation Task and due to time constraints was subsequently omitted from the Metrics Task evaluation, Alibaba-Ensemble.

[4] `https://www.mturk.com`

2.3.1 System-level Manual Quality Judgments

In the system-level evaluation, the goal is to assess the quality of translation of an MT system for the whole test set. Our manual scoring method, DA, nevertheless proceeds sentence by sentence, aggregating the final score as described below.

Direct Assessment (DA) This year the translation task employed monolingual direct assessment (DA) of translation adequacy (Graham et al., 2013; Graham et al., 2014a; Graham et al., 2016). Since sufficient levels of agreement in human assessment of translation quality are difficult to achieve, the DA setup simplifies the task of translation assessment (conventionally a bilingual task) into a simpler monolingual assessment. In addition, DA avoids bias that has been problematic in previous evaluations introduced by assessment of several alternate translations on a single screen, where scores for translations had been unfairly penalized if often compared to high quality translations (Bojar et al., 2011). DA therefore employs assessment of individual translations in isolation from other outputs.

Translation adequacy is structured as a monolingual assessment of similarity of meaning where the target language reference translation and the MT output are displayed to the human assessor. Assessors rate a given translation by how adequately it expresses the meaning of the reference translation on an analogue scale corresponding to an underlying 0-100 rating scale.[5]

Large numbers of DA human assessments of translations for all 14 language pairs included in the News Translation Task were collected from researchers and from workers on Amazon's Mechanical Turk, via sets of 100-translation hits to ensure sufficient repeat assessments per worker, before application of strict quality control measures to filter out assessments from poor performers.

In order to iron out differences in scoring strategies attributed to distinct human assessors, human assessment scores for translations were standardized according to an individual judge's overall mean and standard deviation score. Final scores for MT systems were computed by firstly taking the average of scores for individual translations in the test set (since some were assessed more than once), before combining all scores for translations attributed to a given MT system into its overall adequacy score. The gold standard for system-level DA evaluation is thus what is denoted "Ave z" in Findings 2018 (Bojar et al., 2018).

Finally, although it was necessary to apply a sentence length restriction in WMT human evaluation prior to the introduction of DA, the simplified DA setup does not require restriction of the evaluation in this respect and no sentence length restriction was applied in DA WMT18.

2.3.2 Segment-level Manual Quality Judgments

Segment-level metrics have been evaluated against DA annotations for the newstest2018 test set. This year, a standard segment-level DA evaluation of metrics, where each translation is assessed a minimum of 15 times, was unfortunately not possible due to insufficient number of judgements collected. DA judgements from the system-level evaluation were therefore converted to relative ranking judgements (daRR) to produce results. This is the same strategy as carried out for some out-of-English language pairs in last year's evaluation.

daRR 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 will denote these re-interpreted DA judgements as "DARR", to distinguish it clearly from the "RR" golden truth used in the past years.

Since the analogue rating scale employed by DA is marked at the 0-25-50-75-100 points, we decided to 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.

[5]The only numbering displayed on the rating scale are extreme points 0 and 100%, and three ticks indicate the levels of 25, 50 and 75 %.

	DA>1	Ave	DA pairs	daRR
cs-en	2,491	3.6	13,223	5,110
de-en	2,995	11.4	192,702	77,811
et-en	2,000	11.2	118,066	56,721
fi-en	2,972	5.4	39,127	15,648
ru-en	2,916	4.9	31,361	10,404
tr-en	2,991	4.5	24,325	8,525
zh-en	3,952	7.2	97,474	33,357
en-cs	1,586	4.9	15,311	5,413
en-de	2,150	5.3	47,041	19,711
en-et	1,035	13.6	90,755	32,202
en-fi	1,481	5.3	30,613	9,809
en-ru	2,954	6.2	54,260	22,181
en-tr	707	3.4	4,750	1,358
en-zh	3,915	6.5	86,286	28,602
newstest2018				

Table 1: Number of judgements for DA converted to daRR data; "DA>1" is the number of source input sentences 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 sentence; "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.

Any inconsistency that could arise from reliance on DA judgements collected from low quality crowd-sourcing is thus prevented.

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 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 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 1 due to combinatorial advantage of extracting daRR judgements from all possible pairs of translations of the same source input.

The daRR judgements serve as the golden standard for segment-level evaluation in WMT18.

2.4 Participants of the Metrics Shared Task

Table 2 lists the participants of the WMT18 Shared Metrics Task, along with their metrics. We have collected 10 metrics from a total of 8 research groups.

The following subsections provide a brief summary of all the metrics that participated. The list is concluded by our baseline metrics in Section 2.4.9.

As in last year's task, we asked participants whose metrics are publicly available to provide links to where the code can be accessed. Table 3 summarizes links for metrics that participated in WMT18 that are publicly available for download.

We again distinguish metrics that are a combination of other metric scores, denoting them as "ensemble metrics".

2.4.1 BEER

BEER (Stanojević and Sima'an, 2015) is a trained evaluation metric with a linear model that combines features sub-word feature indicators (character n-grams) and global word order features (skip bigrams) to get language agnostic and fast to compute evaluation metric. BEER has participated in previous years of the evaluation task.

2.4.2 Blend

Blend incorporates existing metrics to form an effective combined metric, employing SVM regression for training and DA scores as the gold standard. For to-English language pairs, incorporated metrics include 25 lexical based metrics and 4 other metrics. Since some lexical based metrics are simply different variants of the same metric, there are only 9 kinds of lexical based metrics, namely BLEU, NIST, GTM, METEOR, ROUGE, Ol, WER, TER and PER. The 4 other metrics are CharacTER, BEER, DPMF and ENTF.

Blend has participated in the Metrics Task in WMT17. This year, Blend follows its setup in WMT17, but enlarges the training data since there are some data available in

Metric	Seg-level	Sys-level	Hybrids	Participant
BEER	•	⊘	⊘	ILLC – University of Amsterdam (Stanojević and Sima'an, 2015)
BLEND	•	⊘	⊘	Tencent-MIG-AI Evaluation & Test Lab (Ma et al., 2017)
CharacTer	•	•	•	RWTH Aachen University (Wang et al., 2016a)
ITER	•	•	⋆	Jadavpur University (Panja and Naskar, 2018)
meteor++	•	⊘	⊘	Peking University (Guo et al., 2018)
RUSE	•	⊘	⊘	Tokyo Metropolitan University (Shimanaka et al., 2018)
UHH_TSKM	•	⊘	⊘	(Duma and Menzel, 2017)
YiSi-*	•	⊘	⊘	NRC (Lo, 2018)

Table 2: Participants of WMT18 Metrics Shared Task. "•" denotes that the metric took part in (some of the language pairs) of the segment- and/or system-level evaluation and whether hybrid systems were also scored. "⊘" indicates that the system-level and hybrids are implied, simply taking arithmetic (macro-)average of segment-level scores. "⋆" indicates that the original ITER system-level scores should be calculated as the *micro-average* of segment-level scores but we calculate them as simple macro-averaged for the hybrid systems. See the ITER paper for more details.

BEER	`http://github.com/stanojevic/beer`
BLEND	`http://github.com/qingsongma/blend`
CharacTER	`http://github.com/rwth-i6/CharacTER`
RUSE	`http://github.com/Shi-ma/RUSE`
YiSi-0, incl. -1 and -1_srl	`http://chikiu-jackie-lo.org/home/index.php/yisi`
Baselines:	`http://github.com/moses-smt/mosesdecoder`
BLEU, NIST	`scripts/generic/mteval-v13a.pl`
CDER, PER, TER, WER	`mert/evaluator` ("Moses scorer")
sentBLEU	`mert/sentence-bleu`
chrF, chrF+	`http://github.com/m-popovic/chrF`

Table 3: Metrics available for public download that participated in WMT18. Most of the baseline metrics are available with Moses, relative paths are listed.

WMT17. For to-English language pairs, there are 9280 sentences as training data. 1620 sentences are used for English-Russian (en-ru). Experiments show the performance of Blend can be improved if the training data increases.

Blend is flexible to be applied to any language pairs if incorporated metrics support the specific language pair and DA scores are available.

2.4.3 CharacTER

CharacTER (Wang et al., 2016b; Wang et al., 2016a), 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 non-tokenized 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 reference 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.

A python library was used for calculating the Levenshtein distance, so that the metric is now about 7 times faster than before.

2.4.4 ITER

ITER (Panja and Naskar, 2018) is an improved Translation Edit/Error Rate (TER) metric. In addition to the basic edit operations in TER (insertion, deletion, substitution and shift), ITER also allows stem matching and uses optimizable edit costs and better normalization.

Note that for segment-level evaluation, we reverse the sign of the score, so that better translations get higher scores. For system-level confidence, we calculate the system-level scores for hybrids systems slightly differently than the original ITER definition would require. We use the unweighted arithmetic average of segment-level scores (macro-average) whereas ITER would use the micro-average.

2.4.5 meteor++

METEOR++ (Guo et al., 2018) is metric based on Meteor (Denkowski and Lavie, 2014), adding explicing treatment of "copy-words", i.e. words that are likely to be preserved across all paraphrases of a sentence in a given language.

2.4.6 RUSE

RUSE (Shimanaka et al., 2018) is a perceptron regressor based on three types of sentence embeddings: Infersent, Quick-Thought and Universal Sentence Encoder, designed with the aim to utilize global sentence information that cannot be captured by local features based on character or word n-grams. The sentence embeddings come from pre-trained models and the regression itself is trained on past manual judgements in WMT shared tasks.

2.4.7 UHH_TSKM

UHH_TSKM (Duma and Menzel, 2017) is a non-trained metric utilizing kernel functions, i.e. methods for efficient calculation of overlap of substructures between the candidate and the reference translations. The metric uses both sequence kernels, applied on the tokenized input data, together with tree kernels, that exploit the syntactic structure of the sentences. Optionally, the match can also be performed for the candidate and a pseudo-reference (i.e. a translation by another MT system) or for the source sentence and the candidate back-translated into the source language.

2.4.8 YiSi-0, YiSi-1 and YiSi-1_srl

The YISI metrics (Lo, 2018) are recently proposed semantic MT evaluation metrics inspired by MEANT_2.0 (Lo, 2017). Specifically, YISI-1 is identical to MEANT_2.0-NOSRL from the WMT17 Metrics Task.

YISI-1 also successfully served in the parallel corpus filtering task. Some details are provided in the system description paper (Lo et al., 2018).

YISI-1 measures the relative lexical semantic similarity (weighted word embeddings cosine similarity aggregated into n-grams similarity) of the candidate and reference translations, optionally taking the shallow semantic structure (semantic role labelling, "srl") into account. YISI-0 is a degenerate resource-free version using the longest common character substring, instead of word embeddings cosine similarity, to measure the word similarity of the candidate and reference translations.

2.4.9 Baseline Metrics

As mentioned by Bojar et al. (2016), Metrics Task occasionally suffers from "loss of knowledge" when successful metrics participate only in one year.

We attempt to avoid this by regularly evaluating also a range of "baseline metrics" as implemented in the following tools:

- **Mteval.** The metrics BLEU (Papineni et al., 2002) and NIST (Doddington, 2002) were computed using the script `mteval-v13a.pl`[6] that is used in the OpenMT Evaluation Campaign and includes its own tokenization. We run `mteval` with the flag `--international-tokenization` since it performs slightly better (Macháček and Bojar, 2013).

[6] `http://www.itl.nist.gov/iad/mig/tools/`

- **Moses Scorer.** The metrics TER (Snover et al., 2006), WER, PER and CDER (Leusch et al., 2006) were produced by the Moses scorer, which is used in Moses model optimization. To tokenize the sentences, we used the standard tokenizer script as available in Moses toolkit. When tokenizing, we also convert all outputs to lowercase.

  Since Moses scorer is versioned on Github, we strongly encourage authors of high-performing metrics to add them to Moses scorer, as this will ensure that their metric can be easily included in future tasks.

- **SentBLEU.** The metric sentBLEU is computed using the script sentence-bleu, a part of the Moses toolkit. It is a smoothed version of BLEU that correlates better with human judgements for segment-level. Standard Moses tokenizer is used for tokenization.

- **chrF** The metrics chrF and chrF+ (Popović, 2015; Popović, 2017) are computed using their original Python implementation, see Table 3.

  We run `chrF++.py` with the parameters `-nw 0 -b 3` to obtain the chrF score and with `-nw 0 -b 1` to obtain the chrF+ score. Note that chrF intentionally removes all spaces before matching the n-grams, detokenizing the segments but also concatenating words.

  We originally planned to use the chrF implementation which was recently made available in Moses Scorer but it mishandles Unicode characters for now.

The baselines serve in system and segment-level evaluations as customary: BLEU, TER, WER, PER and CDER for system-level only; sentBLEU for segment-level only and chrF for both.

Chinese word segmentation is unfortunately not supported by the tokenization scripts mentioned above. For scoring Chinese with baseline metrics, we thus pre-processed MT outputs and reference translations with the script tokenizeChinese.py[7] by Shujian Huang, which separates Chinese characters from each other and also from non-Chinese parts.

For computing system-level and segment-level scores, the same scripts were employed as in last year's Metrics Task as well as for generation of hybrid systems from the given hybrid descriptions.

3 Results

We discuss system-level results for news task systems in Section 3.1. The segment-level results are in Section 3.2.

3.1 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:

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

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.

Since some metrics, such as BLEU, for example, aim to achieve a strong positive correlation with human assessment, while error metrics, such as TER aim for a strong negative correlation, after computation of r for metrics, we compare metrics via the absolute value of a given metric's correlation with human assessment.

3.1.1 System-Level Results

Table 4 provides the system-level correlations of metrics evaluating translation of newstest2018 into English while Table 5 provides the same for out-of-English language pairs. The underlying texts are part of the WMT18 News Translation test set (newstest2018) and the underlying MT systems are all MT systems participating in the WMT18 News Translation Task with the exception of a single tr-en system not included in the initial human evaluation run.

As recommended by Graham and Baldwin (2014), we employ Williams significance test (Williams, 1959) to identify differences

[7] `http://hdl.handle.net/11346/WMT17-TVXH`

	cs-en	**de-en**	**et-en**	**fi-en**	**ru-en**	**tr-en**	**zh-en**
n	5	16	14	9	8	5	14
Correlation	$\|r\|$	$\|r\|$	$\|r\|$	$\|r\|$	$\|r\|$	$\|r\|$	$\|r\|$
BEER	**0.958**	0.994	**0.985**	**0.991**	0.982	0.870	**0.976**
BLEND	**0.973**	0.991	0.985	**0.994**	**0.993**	**0.801**	**0.976**
BLEU	**0.970**	0.971	**0.986**	0.973	0.979	**0.657**	**0.978**
CDER	**0.972**	0.980	**0.990**	0.984	0.980	**0.664**	**0.982**
CharacTER	**0.970**	**0.993**	0.979	0.989	**0.991**	**0.782**	0.950
chrF	0.966	0.994	0.981	0.987	**0.990**	0.452	0.960
chrF+	0.966	0.993	0.981	0.989	**0.990**	0.174	0.964
ITER	**0.975**	0.990	0.975	**0.996**	0.937	**0.861**	**0.980**
meteor++	**0.945**	0.991	0.978	0.971	**0.995**	0.864	0.962
NIST	**0.954**	0.984	0.983	0.975	0.973	**0.970**	0.968
PER	**0.970**	0.985	**0.983**	**0.993**	0.967	0.159	0.931
RUSE	**0.981**	**0.997**	**0.990**	**0.991**	**0.988**	**0.853**	**0.981**
TER	**0.950**	0.970	**0.990**	0.968	0.970	0.533	**0.975**
UHH_TSKM	0.952	0.980	**0.989**	0.982	0.980	0.547	**0.981**
WER	**0.951**	0.961	**0.991**	0.961	0.968	0.041	**0.975**
YiSi-0	0.956	**0.994**	0.975	0.978	**0.988**	**0.954**	0.957
YiSi-1	**0.950**	0.992	0.979	0.973	**0.991**	**0.958**	0.951
YiSi-1_srl	**0.965**	**0.995**	0.981	0.977	**0.992**	**0.869**	0.962
				newstest2018			

Table 4: Absolute Pearson correlation of to-English system-level metrics with DA human assessment in newstest2018; correlations of metrics not significantly outperformed by any other for that language pair are highlighted in bold; ensemble metrics are highlighted in gray.

	en-cs	**en-de**	**en-et**	**en-fi**	**en-ru**	**en-tr**	**en-zh**
n	5	16	14	12	9	8	14
Correlation	$\|r\|$	$\|r\|$	$\|r\|$	$\|r\|$	$\|r\|$	$\|r\|$	$\|r\|$
BEER	**0.992**	**0.991**	**0.980**	**0.961**	**0.988**	**0.965**	0.928
BLEND	–	–	–	–	**0.988**	–	–
BLEU	0.995	0.981	0.975	**0.962**	0.983	0.826	0.947
CDER	0.997	0.986	**0.984**	**0.964**	**0.984**	0.861	0.961
CharacTER	**0.993**	**0.989**	0.956	**0.974**	0.983	0.833	**0.983**
chrF	**0.990**	**0.990**	**0.981**	**0.969**	**0.989**	0.948	0.944
chrF+	**0.990**	0.989	**0.982**	**0.970**	0.989	0.943	0.943
ITER	0.915	**0.984**	**0.981**	**0.973**	0.975	0.865	–
NIST	**0.999**	0.986	**0.983**	0.949	**0.990**	0.902	0.950
PER	0.991	0.981	0.958	0.906	**0.988**	0.859	0.964
TER	**0.997**	**0.988**	**0.981**	**0.942**	**0.987**	0.867	**0.963**
WER	**0.997**	**0.986**	**0.981**	**0.945**	0.985	0.853	0.957
YiSi-0	0.973	0.985	0.968	0.944	**0.990**	**0.990**	0.957
YiSi-1	**0.987**	0.985	**0.979**	0.940	**0.992**	**0.976**	0.963
YiSi-1_srl	–	**0.990**	–	–	–	–	0.952
				newstest2018			

Table 5: Absolute Pearson correlation of out-of-English system-level metrics with DA human assessment in newstest2018; correlations of metrics not significantly outperformed by any other for that language pair are highlighted in bold; ensemble metrics are highlighted in gray.

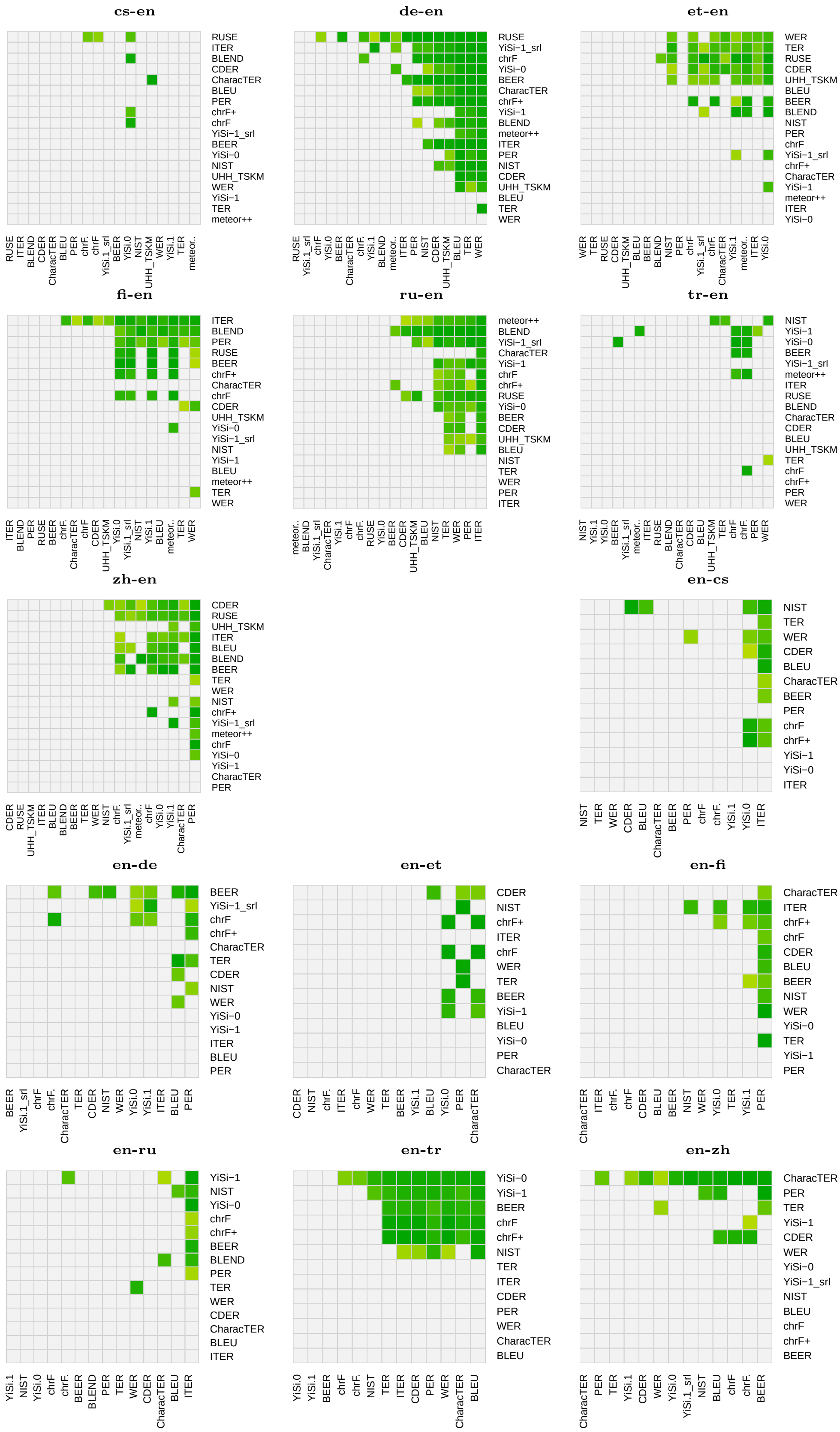

Figure 1: System-level metric significance test results for DA human assessment in newstest2018; green cells denote a statistically significant increase in correlation with human assessment for the metric in a given row over the metric in a given column according to Williams test.

	cs-en	de-en	et-en	fi-en	ru-en	tr-en	zh-en
n	10K	10K	10K	10K	10K	10K	10K
Correlation	$\|r\|$	$\|r\|$	$\|r\|$	$\|r\|$	$\|r\|$	$\|r\|$	$\|r\|$
BEER	0.9497	0.9927	0.9831	0.9824	0.9755	0.7234	0.9677
BLEND	0.9646	0.9904	0.9820	0.9853	0.9865	0.7243	0.9686
BLEU	0.9557	0.9690	0.9812	0.9618	0.9719	0.5862	0.9684
CDER	0.9642	0.9797	0.9876	0.9764	0.9739	0.5767	0.9733
CharacTER	0.9595	0.9919	0.9754	0.9791	0.9841	0.6798	0.9424
chrF	0.9565	0.9929	0.9787	0.9785	0.9836	0.3859	0.9524
chrF+	0.9575	0.9922	0.9784	0.9806	0.9832	0.1737	0.9561
ITER	0.9656	0.9904	0.9746	**0.9885**	0.9429	0.7420	**0.9780**
meteor++	0.9367	0.9898	0.9753	0.9621	**0.9892**	0.7871	0.9541
NIST	0.9419	0.9816	0.9804	0.9655	0.9650	0.8622	0.9589
PER	0.9369	0.9820	0.9782	0.9834	0.9550	0.0433	0.9233
RUSE	**0.9736**	**0.9959**	0.9879	0.9829	0.9820	0.7796	0.9734
TER	0.9419	0.9699	0.9882	0.9599	0.9635	0.4495	0.9670
UHH_TSKM	0.9429	0.9794	0.9869	0.9738	0.9734	0.4433	0.9717
WER	0.9420	0.9612	**0.9892**	0.9534	0.9618	0.0720	0.9667
YiSi-0	0.9465	0.9925	0.9719	0.9694	0.9817	0.8629	0.9495
YiSi-1	0.9425	0.9909	0.9758	0.9641	0.9846	**0.8810**	0.9429
YiSi-1_srl	0.9565	0.9940	0.9783	0.9682	0.9860	0.7850	0.9540
				newstest2018 Hybrids			

Table 6: Absolute Pearson correlation of to-English system-level metrics with DA human assessment for 10K hybrid super-sampled systems in newstest2018; ensemble metrics are highlighted in gray.

	en-cs	en-de	en-et	en-fi	en-ru	en-tr	en-zh
n	10K	10K	10K	10K	10K	10K	10K
Correlation	$\|r\|$	$\|r\|$	$\|r\|$	$\|r\|$	$\|r\|$	$\|r\|$	$\|r\|$
BEER	0.9903	**0.9891**	0.9775	0.9587	0.9864	0.9327	0.9251
BLEND	–	–	–	–	0.9861	–	–
BLEU	0.9931	0.9774	0.9706	0.9582	0.9767	0.7963	0.9414
CDER	0.9949	0.9842	0.9809	0.9605	0.9821	0.8322	0.9564
CharacTER	0.9902	0.9862	0.9495	0.9627	0.9814	0.7752	**0.9784**
chrF	0.9885	0.9876	0.9781	0.9656	0.9868	0.9158	0.9398
chrF+	0.9883	0.9866	0.9786	**0.9665**	0.9861	0.9116	0.9389
ITER	0.8649	0.9778	**0.9817**	**0.9664**	0.9650	0.8724	–
NIST	**0.9967**	0.9839	0.9797	0.9436	0.9877	0.8703	0.9442
PER	0.9865	0.9787	0.9545	0.9044	0.9862	0.8289	0.9500
TER	0.9948	0.9861	0.9770	0.9391	0.9845	0.8373	0.9591
WER	0.9944	0.9842	0.9772	0.9418	0.9829	0.8239	0.9537
YiSi-0	0.9713	0.9829	0.9648	0.9422	0.9879	**0.9530**	0.9513
YiSi-1	0.9851	0.9826	0.9761	0.9384	**0.9893**	0.9418	0.9572
YiSi-1_srl	–	0.9881	–	–	–	–	0.9479
				newstest2018 Hybrids			

Table 7: Absolute Pearson correlation of out-of-English system-level metrics with DA human assessment for 10K hybrid super-sampled systems in newstest2018; ensemble metrics are highlighted in gray.

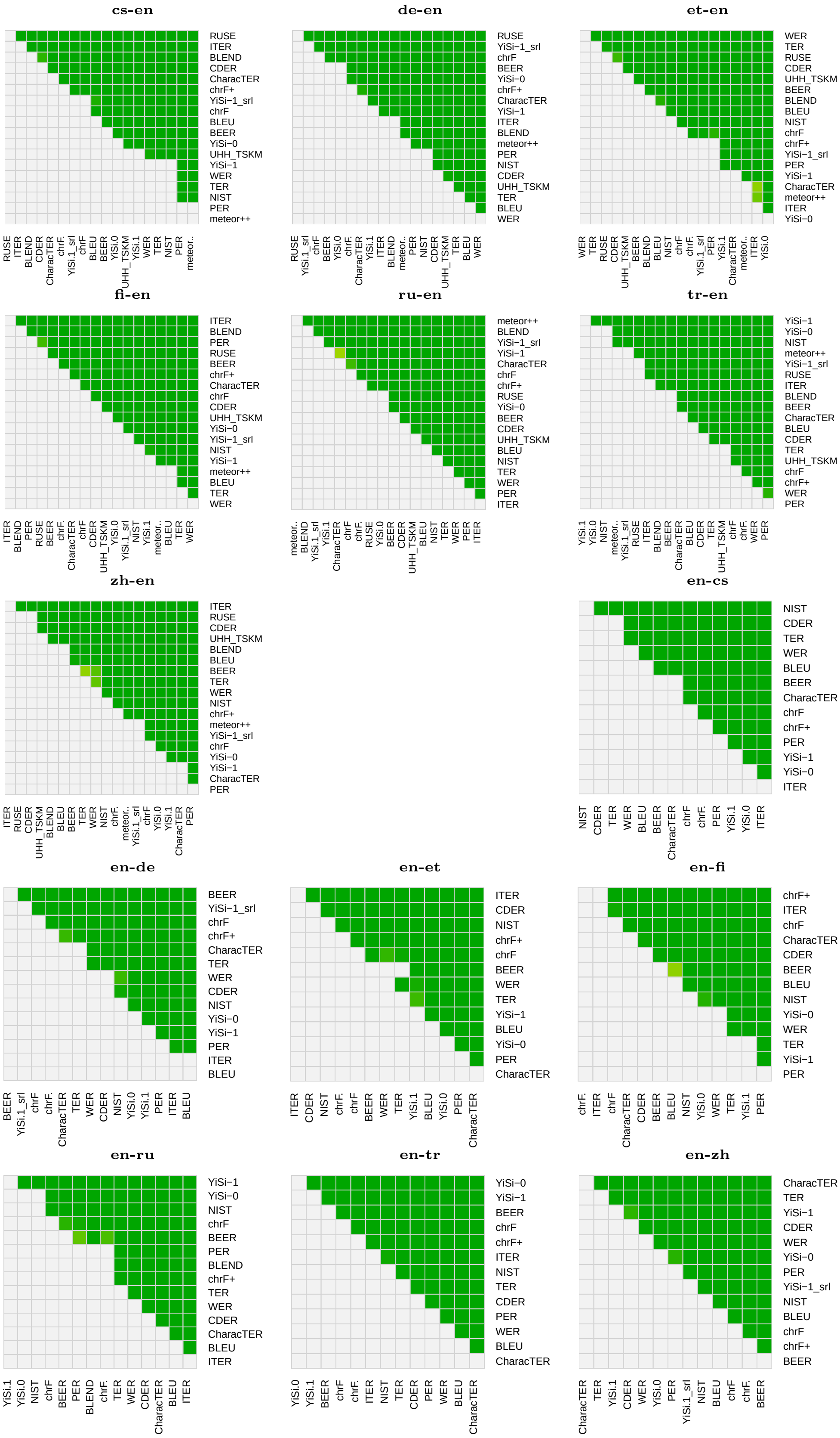

Figure 2: System-level metric significance test results for 10K hybrid systems (DA human evaluation) from newstest2018; green cells denote a statistically significant increase in correlation with human assessment for the metric in a given row over the metric in a given column according to Williams test.

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 Tables 4 and 5.

Since pairwise comparisons of metrics may be also of interest, e.g. to learn which metrics significantly outperform the most widely employed metric BLEU, we include significance test results for every competing pair of metrics including our baseline metrics in Figure 1.

The sample of systems we employ to evaluate metrics is often small, as few as five MT systems for cs-en, 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.

To cater for this, we include significance test results for large hybrid-super-samples of systems (Graham and Liu, 2016). 10K hybrid systems were created per language pair, with corresponding DA human assessment scores by sampling pairs of systems from WMT18 News Translation Task, creating hybrid systems by randomly selecting each candidate translation from one of the two selected systems. Similar to last year, not all metrics participating in the system-level evaluation submitted metric scores for the large set of hybrid systems. Fortunately, taking a simple average of segment-level scores is the proper aggregation method for almost all metrics this year, so where needed, we provided scores for hybrids ourselves, see Table 2.

Correlations of metric scores with human assessment of the large set of hybrid systems are shown in Tables 6 and 7, where again metrics not significantly outperformed by any other are highlighted in bold. Figure 2 then provides significance test results for hybrid super-sampled correlations for all pairs of competing metrics for a given language pair.

3.2 Segment-Level Evaluation

Segment-level evaluation relies on the manual judgements collected in the News Translation Task evaluation. This year, we were unable to follow the methodology outlined in Graham et al. (2015) for evaluation of segment-level metrics because the sampling of sentences 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 measure the quality of metrics' segment-level 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 (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 the last years' WMT17 daRR evaluation (but not earlier). 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:

		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

In the notation of Macháček and Bojar (2014), this corresponds to the setup used in WMT12 (with a different underlying method of manual judgements, RR):

	WMT12	Metric <	Metric =	Metric >
Human	<	1	-1	-1
Human	=	X	X	X
Human	>	-1	-1	1

The key differences between the evaluation used in WMT14–WMT16 and evaluation used in WMT17 and WMT18 are (1) the move from RR to daRR and (2) the treatment of ties.[8] In the years 2014-2016, ties in metrics scores were not penalized. With the move to daRR, where the quality of the two candidate translations is deemed substantially different and no ties in human judgements arise, it makes sense to penalize ties in metrics' predictions in order to promote discerning metrics.

Note that the penalization of ties makes our evaluation asymmetric, dependent on whether the metric predicted the tie for a pair where humans predicted <, or >. It is now important to interpret the meaning of the comparison identically for humans and metrics. For error metrics, we thus reverse the sign of the metric score prior to the comparison with human scores: higher scores have to indicate better translation quality. In WMT18, we did this for ITER and the original authors did this for CharacTER.

To summarize, the WMT18 Metrics Task for segment-level evaluation:

- excludes all human ties (this is already implied by the construction of DARR from DA judgements),
- counts metric's ties as a *Discordant* pairs,
- ensures that error metrics are first converted to the same orientation as the human judgements, i.e. higher score indicating higher translation quality.

We employ bootstrap resampling (Koehn, 2004; Graham et al., 2014b) to estimate confidence intervals for our Kendall's Tau formulation, and metrics with non-overlapping 95% confidence intervals are identified as having statistically significant difference in performance.

[8]Due to an error in the write-up for WMT17 (errata to follow), this second change was not properly reflected in the paper, only in the evaluation scripts.

3.2.1 Segment-Level Results

Results of the segment-level human evaluation for translations sampled from the News Translation Task are shown in Tables 8 and 9, where metric correlations not significantly outperformed by any other metric are highlighted in bold. Head-to-head significance test results for differences in metric performance are included in Figure 3.

4 Discussion

4.1 Obtaining Human Judgements

Human data was collected in the usual way, a portion via crowd-sourcing and the remaining from researchers who mainly committed their time contribution to the manual evaluation as they had submitted a system in that language pair. Evaluation of translations employed the DA set-up and it again successfully acquired sufficient judgments to evaluate systems. As in the previous years, hybrid supersampling proved very effective and allowed to obtain conclusive results of system-level evaluation even for language pairs where as few as 5 MT systems participated. We should however note that hybrid systems are constructed by randomly mixing sentences coming from different MT systems. As soon as document-level evaluation becomes relevant (which we anticipate in the next evaluation campaign already), this style of hybridization is susceptible to breaking cross-sentence references in MT outputs and may no longer be applicable.

In the case of segment-level evaluation, the optimal human evaluation data was unfortunately not available due to resource constraints. Conversion of document-level data held as a substitute for segment-level DA scores. These scores are however not optimal for evaluation of segment-level metrics and we would like to return to DA's standard segment-level evaluation in future, where a minimum of 15 human judgments of translation quality are collected per translation and combined to get highly accurate scores for translations.

	cs-en	de-en	et-en	fi-en	ru-en	tr-en	zh-en
Human Evaluation	daRR	daRR	daRR	daRR	daRR	daRR	daRR
n	5,110	77,811	56,721	15,648	10,404	8,525	33,357
Correlation	τ	τ	τ	τ	τ	τ	τ
BEER	0.295	0.481	0.341	0.232	**0.288**	**0.229**	**0.214**
BLEND	**0.322**	**0.492**	0.354	0.226	**0.290**	**0.232**	**0.217**
CharacTER	0.256	0.450	0.286	0.185	0.244	0.172	**0.202**
chrF	0.288	0.479	0.328	0.229	0.269	0.210	**0.208**
chrF+	0.288	0.479	0.332	0.234	0.279	0.218	**0.207**
ITER	0.198	0.396	0.235	0.128	0.139	-0.029	0.144
meteor++	0.270	0.457	0.329	0.207	0.253	0.204	0.179
RUSE	**0.347**	**0.498**	**0.368**	**0.273**	**0.311**	**0.259**	**0.218**
sentBLEU	0.233	0.415	0.285	0.154	0.228	0.145	0.178
UHH_TSKM	0.274	0.436	0.300	0.168	0.235	0.154	0.151
YiSi-0	0.301	0.474	0.330	0.225	**0.294**	0.215	**0.205**
YiSi-1	**0.319**	0.488	0.351	0.231	**0.300**	**0.234**	**0.211**
YiSi-1_srl	**0.317**	0.483	0.345	0.237	**0.306**	**0.233**	**0.209**
				newstest2018			

Table 8: Segment-level metric results for to-English language pairs in newstest2018: absolute Kendall's Tau formulation of segment-level metric scores with DA scores; correlations of metrics not significantly outperformed by any other for that language pair are highlighted in bold; ensemble metrics are highlighted in gray.

	en-cs	en-de	en-et	en-fi	en-ru	en-tr	en-zh
Human Evaluation	daRR	daRR	daRR	daRR	daRR	daRR	daRR
n	5,413	19,711	32,202	9,809	22,181	1,358	28,602
Correlation	τ	τ	τ	τ	τ	τ	τ
BEER	**0.518**	**0.686**	0.558	**0.511**	**0.403**	**0.374**	0.302
BLEND	–	–	–	–	**0.394**	–	–
CharacTER	0.414	0.604	0.464	0.403	0.352	**0.404**	**0.313**
chrF	**0.516**	0.677	**0.572**	**0.520**	0.383	**0.409**	**0.328**
chrF+	**0.513**	**0.680**	**0.573**	**0.525**	**0.392**	**0.405**	**0.328**
ITER	0.333	0.610	0.392	0.311	0.291	0.236	–
sentBLEU	0.389	0.620	0.414	0.355	0.330	0.261	**0.311**
YiSi-0	0.471	0.661	0.531	0.464	**0.394**	**0.376**	**0.318**
YiSi-1	**0.496**	**0.691**	0.546	**0.504**	**0.407**	**0.418**	**0.323**
YiSi-1_srl	–	**0.696**	–	–	–	–	**0.310**
				newstest2018			

Table 9: Segment-level metric results for out-of-English language pairs in newstest2018: absolute Kendall's Tau formulation of segment-level metric scores with DA scores; correlations of metrics not significantly outperformed by any other for that language pair are highlighted in bold; ensemble metrics are highlighted in gray.

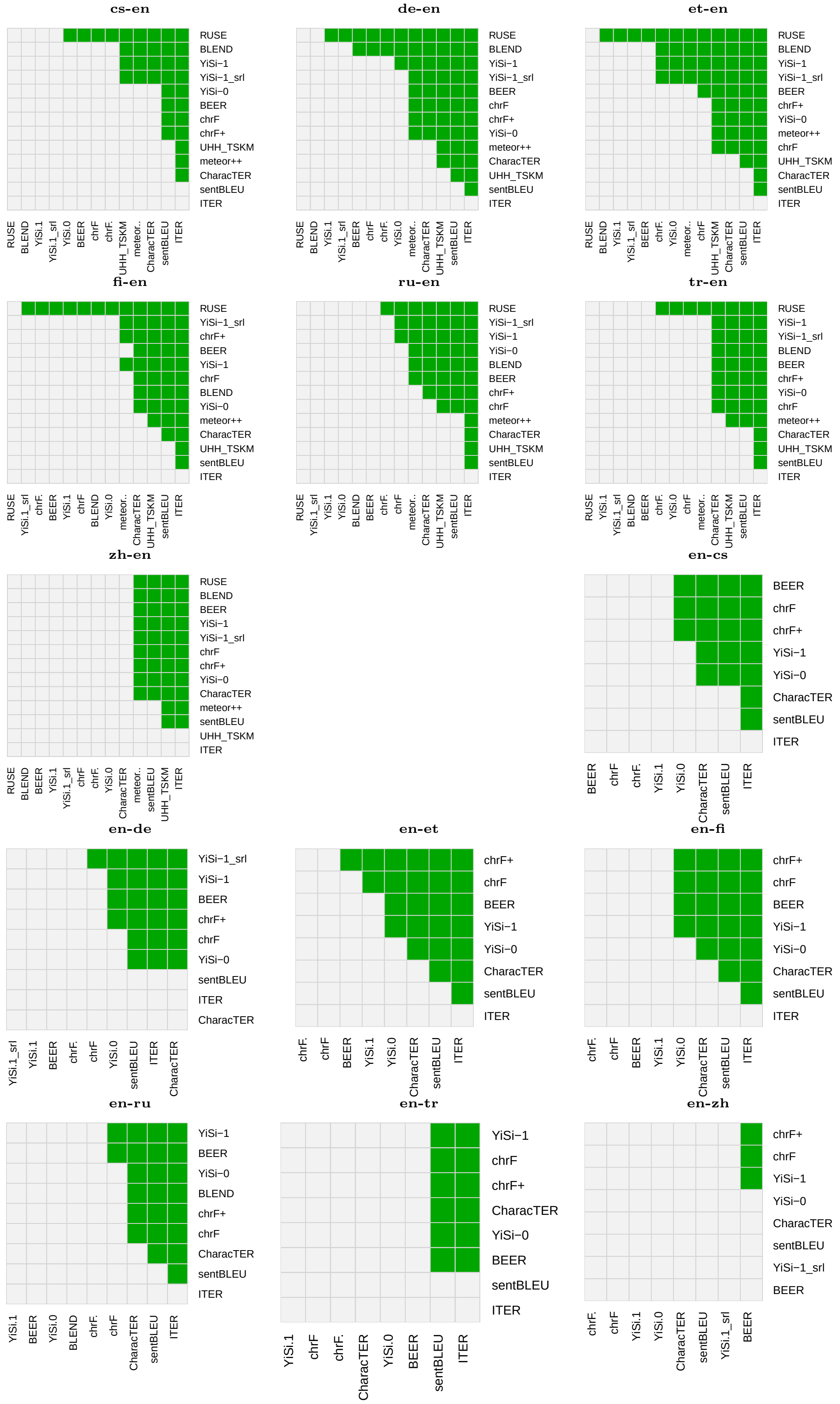

Figure 3: DARR segment-level metric significance test results for all language pairs (newstest2018): Green cells denote a significant win for the metric in a given row over the metric in a given column according bootstrap resampling.

Metric	LPs	Med Corr	Relative Wins
RUSE	7	0.9882	0.8318
BLEND	8	0.9864	0.7043
chrF+	14	0.9812	0.5364
chrF	14	0.9811	≀ 0.5705
BEER	14	0.9810	≀ 0.6041
CDER	14	0.9809	0.5853
CharacTER	14	0.9809	0.4680
UHH_TSKM	7	0.9801	0.4453
YiSi-1	14	0.9775	≀ 0.4631
YiSi-1_srl	9	0.9770	≀ 0.5972
ITER	13	0.9747	0.5181
YiSi-0	14	0.9740	0.4585
NIST	14	0.9739	≀ 0.4926
BLEU	14	0.9739	0.3317
meteor++	7	0.9711	≀ 0.3863
TER	14	0.9698	≀ 0.4046
PER	14	0.9685	0.2292
WER	14	0.9644	≀ 0.3364

Table 10: Summary of system-level evaluation across all language pairs. Metrics sorted by the median correlation; "≀" marks pairwise wins out of sequence.

4.2 Overall Metric Performance

As always, the observed performance of metrics depends on the underlying texts and systems that participate in the News Translation Task. To obtain at least a partial overall view, consider Table 10 for the system-level evaluation and Table 11 for the segment-level evaluation, where each table lists all participating metrics according to median correlation ("Med Corr", evaluated without hybrid systems) achieved across all language pairs they took part in. The number of language pairs is also provided under "LPs".

The final column "Relative Wins" indicates the (micro-averaged) proportion of significant pairwise wins achieved by each metric out of the total possible wins, as presented in Figures 2 and 3 (for this purpose, hybrids are most appropriate). For example, if a metric outperformed every competitor in all pairwise matches, the "Relative Wins" score would be 1.0. The rankings of metrics according to median correlation and relative wins do not always agree and the symbol "≀" indicates such entries. One such striking example is RUSE in segment-level evaluation. Its median performance is not outstanding but it significantly outperformed many competitors in many language pairs, see Figure 3.

Figure 4 shows box plots of system and segment-level correlations for metrics that participated in *all* 14 language pairs. Comparing metrics within this subset only is more meaningful because these metrics evaluate both easy and more difficult language pairs.

Overall, the reported figures confirm the observation from the past years that system-level metrics can achieve correlations above 0.9 but even the best ones can fall to 0.7 or 0.8 for some language pairs. Kendall's Tau achieved by segment-level metrics are generally lower, in the range of 0.25–0.4. The best metrics in their best language pairs can reach up to 0.69 of segment-level correlations with humans. This capping could be possibly in part attributed to the sub-optimal human evaluation data, DA judgements converted to relative ranking.

In system-level evaluation, the new metric RUSE stands out as a metric that achieve highest correlation in more than one language pair according to the hybrid evaluation. YiSi, on the other hand, performs best across all language pairs on average (but not in terms

Metric	LPs	Med Corr	Relative Wins
YiSi-1	14	0.3786	0.4223
chrF+	14	0.3620	0.3660
BEER	14	0.3577	0.3589
chrF	14	0.3553	0.3378
YiSi-0	14	0.3529	0.3167
CharacTER	14	0.3326	0.1548
RUSE	7	0.3110	≀ 0.7126
YiSi-1_srl	9	0.3100	0.4031
BLEND	8	0.3060	≀ 0.4337
sentBLEU	14	0.2979	0.0774
meteor++	7	0.2528	≀ 0.1923
ITER	13	0.2356	0
UHH_TSKM	7	0.2353	≀ 0.0905

Table 11: Summary of segment-level evaluation across all language pairs. Metrics sorted by the median correlation; "≀" marks pairwise wins out of sequence.

of the median). At the system-level, ITER also performs very well in en-et, en-fi, zh-en and several other languages but fails for en-ru and en-cs, which drags its overall performance down.

Both YiSi and RUSE are based on neural networks (YiSi via word and phrase embeddings, RUSE via sentence embeddings). This is a new trend compared to the last year evaluation where the best performance was reached by character-level (not deep) metrics BEER, chrF (and its variants) and CharacTER.

It is important to note that the results of performance agreggated over language pairs are not particularly stable across years. In the last year's evaluation, NIST seemed worse than TER and the reverse seems to happen this year. These differences however easily fall into the indicate boxplots quartiles.

All of the "winners" in this years campaign are publicly available, which is very good for their prospective wider adoption. If participants could put the additional effort of adding their code to Moses scorer, this would guarantee their long-term inclusion in the Metrics Task.

5 Conclusion

This paper summarizes the results of WMT18 shared task in machine translation evaluation, the Metrics Shared Task. Participating metrics were evaluated in terms of their correlation with human judgment at the level of the whole test set (system-level evaluation), as well as at the level of individual sentences (segment-level evaluation). For the former, best metrics reach over 0.95 Pearson correlation or better across several language pairs. Correlations varied more than usual between 0.2 and 0.7 in terms of segment-level metrics Kendall's τ results.

The results confirm the observation form the last year, namely that character-level metrics (chrF, BEER, CharacTER, etc.) generally perform very well. This year adds two new metrics based on word or sentence-level embeddings (RUSE and YiSi), and both join this high-performing group.

Acknowledgments

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

This study was supported in parts by the grants 18-24210S of the Czech Science Foundation, ADAPT Centre for Digital Content Technology (`www.adaptcentre.ie`) at Dublin City University funded under the SFI Research Centres Programme (Grant 13/RC/2106) co-funded under the European Regional Development Fund, and Charles University Research Programme "Progres" Q18+Q48.

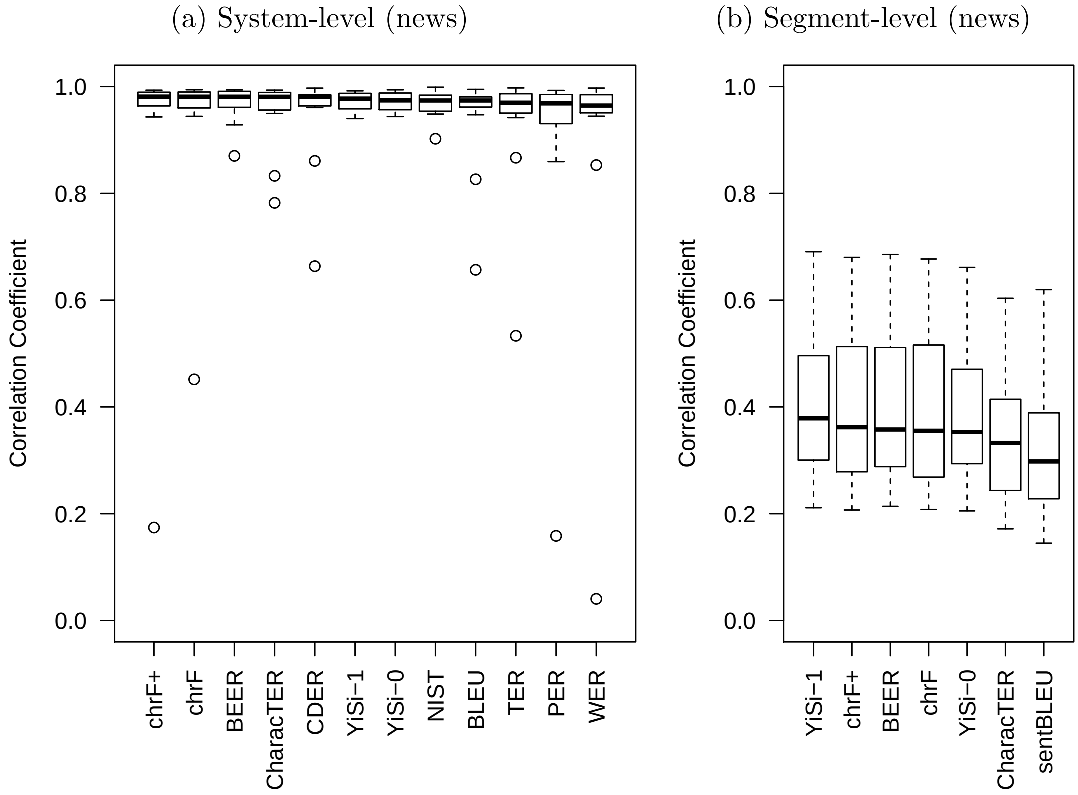

Figure 4: Plots of correlations achieved by metrics in (a) all language pairs for newstest2018 on the system level; (b) all language pairs for newstest2018 on the segment-level; all correlations are for non-hybrid correlations only.

References

Ondřej Bojar, Miloš Ercegovčević, Martin Popel, and Omar Zaidan. 2011. A Grain of Salt for the WMT Manual Evaluation. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 1–11, Edinburgh, Scotland, July. Association for Computational Linguistics.

Ondřej Bojar, Christian Federmann, Barry Haddow, Philipp Koehn, Matt Post, and Lucia Specia. 2016. Ten Years of WMT Evaluation Campaigns: Lessons Learnt. In *Proceedings of the LREC 2016 Workshop "Translation Evaluation – From Fragmented Tools and Data Sets to an Integrated Ecosystem"*, pages 27–34, Portorose, Slovenia, 5.

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 Tasks Papers*, Copenhagen, Denmark, September. Association for Computational Linguistics.

Ondřej Bojar, Jiří Mírovský, Kateřina Rysová, and Magdaléna Rysová. 2018. Evald reference-less discourse evaluation for wmt18. In *Proceedings of the Third Conference on Machine Translation*, Belgium, Brussels, October. Association for Computational Linguistics.

Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, Baltimore, Maryland, USA, June. Association for Computational Linguistics.

George Doddington. 2002. Automatic Evaluation of Machine Translation Quality Using N-gram Co-occurrence Statistics. In *Proceedings of the Second International Conference on Human Language Technology Research*, HLT '02, pages 138–145, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Melania Duma and Wolfgang Menzel. 2017. UHH submission to the WMT17 metrics shared task. In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Tasks Papers*, Copenhagen, Denmark, September. 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, October. Association for Computational Linguistics.

Yvette Graham and Qun Liu. 2016. Achieving Accurate Conclusions in Evaluation of Automatic Machine Translation Metrics. In *Proceedings of the 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, San Diego, CA. 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 & Interoperability with Discourse*, pages 33–41, Sofia, Bulgaria. Association for Computational Linguistics.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2014a. Is Machine Translation Getting Better over Time? In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 443–451, Gothenburg, Sweden, April. Association for Computational Linguistics.

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

Yvette Graham, Nitika Mathur, and Timothy Baldwin. 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*, Denver, Colorado.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2016. Can machine translation systems be evaluated by the crowd alone. *Natural Language Engineering*, FirstView:1–28, 1.

Yinuo Guo, Chong Ruan, and Junfeng Hu. 2018. Meteor++: Incorporating copy knowledge into machine translation evaluation. In *Proceedings of the Third Conference on Machine Translation*, Belgium, Brussels, October. Association for Computational Linguistics.

Philipp Koehn and Christof Monz. 2006. Manual and Automatic Evaluation of Machine Translation Between European Languages. In *Proceedings of the Workshop on Statistical Machine Translation*, StatMT '06, pages 102–121, Stroudsburg, PA, USA. Association for Computational Linguistics.

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

Gregor Leusch, Nicola Ueffing, and Hermann Ney. 2006. CDER: Efficient MT Evaluation Using Block Movements. In *In Proceedings of EACL*, pages 241–248.

Chi-kiu Lo, Michel Simard, Darlene Stewart, Samuel Larkin, Cyril Goutte, and Patrick Littell. 2018. Accurate semantic textual similarity for cleaning noisy parallel corpora using semantic machine translation evaluation metric: The nrc supervised submissions to the parallel corpus filtering task. In *Proceedings of the Third Conference on Machine Translation*, Belgium, Brussels, October. Association for Computational Linguistics.

Chi-kiu Lo. 2017. MEANT 2.0: Accurate semantic MT evaluation for any output language. In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Tasks Papers*, Copenhagen, Denmark, September. Association for Computational Linguistics.

Chi-kiu Lo. 2018. The NRC metric submission to the WMT18 metric and parallel corpus filtering shared task. In *Arxiv*.

Qingsong Ma, Yvette Graham, Shugen Wang, and Qun Liu. 2017. Blend: a novel combined MT metric based on direct assessment — casict-dcu submission to WMT17 metrics task. In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Tasks Papers*, Copenhagen, Denmark, September. 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, MD, USA. Association for Computational Linguistics.

Matouš Macháček and Ondřej Bojar. 2013. Results of the WMT13 Metrics Shared Task. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 45–51, Sofia, Bulgaria, August. Association for Computational Linguistics.

Joybrata Panja and Sudip Kumar Naskar. 2018. Iter: Improving translation edit rate through optimizable edit costs. In *Proceedings of the Third Conference on Machine Translation*, Belgium, Brussels, October. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 311–318.

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

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Tasks Papers*, Copenhagen, Denmark, September. Association for Computational Linguistics.

Hiroki Shimanaka, Tomoyuki Kajiwara, and Mamoru Komachi. 2018. Ruse: Regressor using sentence embeddings for automatic machine translation evaluation. In *Proceedings of the Third Conference on Machine Translation*, Belgium, Brussels, October. 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 *In Proceedings of Association for Machine Translation in the Americas*, pages 223–231.

Miloš Stanojević and Khalil Sima'an. 2015. BEER 1.1: ILLC UvA submission to metrics and tuning task. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, Lisboa, Portugal, September. Association for Computational Linguistics.

Weiyue Wang, Jan-Thorsten Peter, Hendrik Rosendahl, and Hermann Ney. 2016a. Character: Translation edit rate on character level. In *ACL 2016 First Conference on Machine Translation*, pages 505–510, Berlin, Germany, August.

Weiyue Wang, Jan-Thorsten Peter, Hendrik Rosendahl, and Hermann Ney. 2016b. CharacTer: Translation Edit Rate on Character Level. In *Proceedings of the First Conference on Machine Translation*, Berlin, Germany, August. Association for Computational Linguistics.

Evan James Williams. 1959. *Regression analysis*, volume 14. Wiley New York.