Results of the WMT17 Metrics Shared Task

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

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

Amir Kamran
University of Amsterdam
ILLC
a.kamran@uva.nl

Abstract

This paper presents the results of the WMT17 Metrics Shared Task. We asked participants of this task to score the outputs of the MT systems involved in the WMT17 news translation task and Neural MT training task. We collected scores of 14 metrics from 8 research groups. In addition to that, we computed scores of 7 standard metrics (BLEU, SentBLEU, 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 WMT17 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).

This year, we build upon two types of manual judgements: direct assessment (DA) and HUME manual semantic judgements.

1 Introduction

Evaluating the quality of machine translation (MT) is critical for developers of MT systems to monitor progress as well as for MT users to select among available MT engines for their language pair of interest. Manual evaluation is however costly and difficult to reproduce. Automatic MT evaluation can resolve these issues, if it matches manual evaluation. The Metrics Shared Task[1] of WMT annually evaluates the performance of automatic machine translation metrics in their ability to provide a substitute for human assessment of translation quality.

In contrast to *MT quality estimation*, the metrics task provides participating metrics with reference translations with which MT outputs are compared. The metrics task itself then needs manual judgements of translation quality in order to check the extent to which the automatic metrics can approximate the judgement. For situations where the reference translation is not available, please consult the results of Quality Estimation Task (Bojar et al., 2017a).

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 has to be assigned to every individual sentence.

The underlying texts and MT systems come from two other WMT tasks, namely News Translation Task (Bojar et al., 2017a, denoted as Findings 2017 in the following) and Neural MT training task (Bojar et al., 2017b), and from the EU project HimL, aiming at translation of health-related documents. The texts were drawn mainly from the news domain and, to a limited extent, from the medical domain and involve translations to/from Chinese (zh), Czech (cs), Finnish (fi), German (de), Latvian (lv), Russian (ru), and Turkish (tr), each paired with English, and additionally English into Romanian and Polish, making a total of 16 language pairs.

Two sources of golden truth of translation quality judgement are used this year:

- In *Direct Assessment* (DA) (Graham et al., 2015), humans assess the quality of a given MT output translation by comparison with a reference translation (but not the source). DA is the new standard used in WMT news translation task evaluation, requiring only monolingual evaluators. The added benefit for the metrics task is that the manual and automatic

[1] http://www.statmt.org/wmt17/metrics-task.html, starting with Koehn and Monz (2006) up to Bojar et al. (2016b)

Proceedings of the Conference on Machine Translation (WMT), Volume 2: Shared Task Papers, pages 489–513

evaluations are now a little closer: both humans and metrics compare the MT output with the reference.

- The *HUME* score (Birch et al., 2016) is a segment-level score aggregated over manual judgements of translation quality of semantic units of the source sentence.

In contrast to previous years, the official method of evaluation changes, moving from "relative ranking" (RR, evaluating up to five system outputs on an annotation screen relative to each other) to DA and employing the Pearson correlation r in most cases. Due to difficulties in obtaining sufficient number of judgements for segment-level evaluation of some language pairs, we re-interpret DA judgements for these language pairs 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 two types of test sets 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 sets, i.e. sets of source sentences and reference translations:

newstest2017 is the main test set. It is the test set used in WMT17 News translation task (see Findings 2017), with approximately 3,000 sentences for each translation direction (except Chinese and Latvian which only have 2,001 sentences). The set includes a single reference translation for each direction, except English→Finnish with two reference translations.

himltest2017 is a subset of HUME Test Set Round 2 as released by the EU project HimL. More details about the original dataset are available in Deliverable D5.4 of the project.[2] Out selection contains approximately 300 sentences for each of the four language pairs (from English into Czech, German, Polish and Romanian) coming from both WMT16 news translation task as well as from HimL test sets 2015,[3] which are sentences from health-related texts by Cochrane and NHS 24. The reference translations are the standard WMT16 references for the news domain and post-edits of phrase-based MT for the Cochrane and NHS 24 sentences. No document structure has been preserved in this dataset.

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 the humans, to distinguish between the quality of translations. If the task includes a wide range of systems of varying quality, however, or systems 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, we relied on the following underlying MT systems:

News Task Systems are all machine translation systems participating in the WMT17 News translation task (see Findings 2017). The best among these systems were neural MT systems (both token- and character-based) but a good number of standard phrase-based systems and also some transfer-based and rule-based systems participated. The exact set of systems and system types depends on the language pair.

NMT Training Task systems are all instances of Neural Monkey (Helcl and Libovický, 2017) implementing the Bahdanau et al. (2014) sequence-to-sequence model with attention. Participants of the NMT training task trained a fixed NMT model using fixed training data (a subset of the news translation task training data) and these submitted models were

[2] http://www.himl.eu/files/D5.4_Second_Evaluation_Report.pdf

[3] http://www.himl.eu/test-sets

then run by training task organizers on newstest2017, see Bojar et al. (2017b) for more details. All training task systems can be thus seen as regular submissions to the news translation task, with additional constraints in place. While one would expect these systems to produce outputs more similar to each other than the remaining news task systems, this is not the case, see Table 3 in Findings 2017. Based on the manual evaluation, training task systems however perform similarly, occupying the lower half of the ranking.

HUME Test Set Round 2 Systems are the MT systems translating himltest2017. For each language pair, three different MT systems are provided. The translations were run by the EU project HimL and the systems cover major MT system types for each language pair (phrase-based, neural and also syntax-based or combined systems). More details are provided in Table 3 of Deliverable 5.4 of the HimL project.[4]

To match the format of the newstest where all MT systems translate all sentences, we selected such subsets of sentences from HUME Test Set Round 2. The availability of MT systems for Romanian sentences was more varied than for other languages and we thus decided to split Romanian into two test sets, himltest2017a and himltest2017b, the first fully translated by three systems and the second fully translated only by two systems.

Important note: Due to the construction of himltest2017 for Polish, the outputs of one of the MT system were to a large part included in the HUME track last year and thus leaked to the training data we provided to metrics task participants this year. The affected test set file is `himltest2017a.Year1.en-pl` with 324 sentences out of 340 included in the training data. The file `himltest2017a.PBMT.en-pl` also contains 16 known sentences, probably due to identical translation. The performance of trained metrics for en-pl evaluation have the potential to be inflated therefore.

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 separately for newstest2017 and himltest2017 by randomly alternating sentences from the outputs of pairs of systems of the given dataset. In short, we create 10K hybrid MT systems for each language pair.

Excluding the hybrid systems, we ended up with 166 system outputs across 16 language pairs and 3 test sets.

2.3 Manual MT Quality Judgments

There are two distinct "golden truths" employed to evaluate metrics this year: Direct Assessment (DA) and HUME, a semantic-based manual metric.

The details of both of the methods are provided in this section, separately 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 and crowd-sourced workers on Amazon's Mechanical Turk.[5] 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.

2.3.1 System-level Manual Quality Judgments

In system-level evaluation, the goal is to assess the quality of translation of an MT system for the whole test set. Our manual scoring methods DA and HUME nevertheless proceed sentence by sentence, aggregating the final score in some way.

Direct Assessment (DA) This year the translation task employed monolingual direct assessment (DA) of translation adequacy (Graham et al., 2013; Graham et al., 2014; 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. Furthermore, DA avoids bias that has been problematic in previous evaluations introduced by assessment of several alternate translations on one screen, where

[4] `http://www.himl.eu/files/D5.4_Second_Evaluation_Report.pdf`

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

scores for translations were 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.[6]

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

In order to iron out differences in scoring strategies attributed to distinct workers, human assessment scores for translations were standardized according to an individual worker's overall mean and standard deviation score. Mean standardized scores for translation task participating 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 2017 (Bojar et al., 2017a).

Finally, although it is common to apply a sentence length restriction in WMT human evaluation, the simplified DA setup does not require restriction of the evaluation in this respect and no sentence length restriction was applied in DA WMT17.

HUME is a human evaluation measure that decomposes over the UCCA semantic units (Birch et al., 2016). UCCA (Abend and Rappoport, 2013) is an appealing candidate for semantic analysis, due to its cross-linguistic applicability, support for rapid annotation, and coverage of many fundamental semantic phenomena, such as verbal, nominal and adjectival argument structures and their inter-relations. HUME operates by aggregating human assessments of the translation quality of individual semantic units in the source sentence. HUME thus avoids the semantic annotation of machine-generated text, which can often be garbled or semantically unclear. This also allows the re-use of the source semantic annotation for measuring the quality of different translations of the same source sentence, and avoids reliance on possibly suboptimal reference translations. HUME shows good inter-annotator agreement, and reasonable correlation with Direct Assessment (Birch et al., 2016).

Since some translations in the HUME Test Set round 2 were annotated with HUME by more than one annotator, individual HUME scores for the same translation were combined into a single score for evaluation of metrics by taking the average of all HUME scores attributed to that translation. These segment-level HUME scores were then combined into an average score for each system.

2.3.2 Segment-level Manual Quality Judgments

Segment-level metrics have been evaluated against DA and HUME annotations for the newstest2017 and himl test sets, respectively. This year, since insufficient repeat judgements were collected for most of out-of-English language pairs to run a standard segment-level DA evaluation of metrics for the news task data, DA judgements for those language pairs were converted to relative ranking judgements to produce results similar to previous WMT metrics tasks.

Segment-level DA Adequacy assessments were collected for translations sampled from the output of systems participating in WMT17 translation task for 14 language pairs of the news translation task and 4 language pairs of the himl test set. Since the actual MT system is not important for segment-level assessment, we sampled 560 translations per language pair at random avoiding selection of identical ones.

Segment-level DA adequacy scores were collected as in system-level DA, described in Section 2.3.1, again with strict quality control and score standardization applied. To achieve accurate segment-level scores for translations, 15 distinct DA assessments were collected and com-

[6] 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
en-cs	2,960	6.9	67,404	32,810
en-de	2,053	3.1	8,140	3,227
en-fi	2,071	2.9	6,952	3,270
en-lv	1,616	3.4	8,047	3,456
en-tr	460	2.1	597	247

Table 1: Number of judgements for the five out-of-English language pairs employing DA converted to DARR data (DA produced by volunteer researchers in the news task manual evaluation); "DA>1" is the number of source input sentences in the manual evaluation where at least two translations of that same input sentence both received at least one 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" if the number of DA pairs with an absolute difference in DA scores greater than the 25 percentage point margin.

bined into a single mean adequacy score for each individual translation. Although in general agreement in human assessment of MT has been difficult to achieve, segment-level DA scores employing a minimum of 15 repeat assessments have been shown to be almost completely repeatable (Graham et al., 2015) and therefore provide a reliable gold standard for evaluating segment-level metrics.

HUME HUME annotations were taken from the HUME Test Set round 2 as described already in Section 2.3.1. Again, where an individual translation received more than one annotation its final segment-level score was arrived at by taking the average of all scores attributed to it.

DARR For five out-of-English language pairs (en-cs, en-de, en-fi, en-lv and en-tr) belonging to the news task, insufficient DA judgements were collected to provide reliable segment-level DA scores. 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 us to conclude 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, the difference in DA scores we employ to distinguish translations that are better/worse than one another is 25 points. In addition, DA judgements for these language pairs were only collected from known-reliable volunteers, and therefore avoid any inconsistency that could arise from reliance on individual DA judgements collected via crowd-sourcing, for example.

From the complete set of human assessments collected from researchers for the News task for these five language pairs, 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 four of the five 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. Only Turkish thus remains poorly covered.

Kendall's Tau-like Formulation for DARR 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 we use to evaluate metrics, it is not possible to apply conventional Kendall's Tau given the current DARR human evaluation setup (Graham et al., 2015). Vazquez-Alvarez and Huckvale (2002) also note that a genuine pairwise comparison is likely to lead to more stable results for segment-level metric evaluation.

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

$$\tau = \frac{|Concordant| - |Discordant|}{|Concordant| + |Discordant|} \quad (1)$$

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 from WMT14 and WMT15. For a detailed discussion on other options, see Macháček and Bojar (2014).

The method is formally described using the following matrix:

		Metric		
		$<$	=	$>$
Human	$<$	1	0	-1
	=	X	X	X
	$>$	-1	0	1

Given such a matrix $C_{h,m}$ where $h, m \in \{<, =, >\}$[7] and a metric, we compute the Kendall's τ for the metric the following way:

$$\tau = \frac{\sum_{\substack{h,m \in \{<,=,>\} \\ C_{h,m} \neq \text{X}}} C_{h,m} |S_{h,m}|}{\sum_{\substack{h,m \in \{<,=,>\} \\ C_{h,m} \neq \text{X}}} |S_{h,m}|} \quad (2)$$

We insert each extracted human pairwise comparison into exactly one of the nine sets $S_{h,m}$ according to human and metric ranks. For example the set $S_{<,>}$ contains all comparisons where the left-hand system was ranked better than right-hand system by humans and it was ranked the other way round by the metric in question.

To compute the numerator of our Kendall's τ formulation, we take the coefficients from the matrix $C_{h,m}$, use them to multiply the sizes of the corresponding sets $S_{h,m}$ and then sum them up. We do not include sets for which the value of $C_{h,m}$ is X. To compute the denominator, we simply sum the sizes of all the sets $S_{h,m}$ except those where $C_{h,m} = \text{X}$.

To summarize, the WMT17 matrix specifies to:

- exclude all human ties (this is already implied by the construction of DARR from DA judgements),
- count metric's ties only for the denominator (thus giving no credit for giving a tie),
- all cases of disagreement between human and metric judgements are counted as $Discordant$,
- all cases of agreement between human and metric judgements are counted as $Concordant$.

We employ bootstrap resampling 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.

2.4 Participants of the Metrics Shared Task

Table 2 lists the participants of the WMT17 Shared Metrics Task, along with their metrics. We have collected 14 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 2.4.10.

In this year's task, we asked participants whose metrics are publicly available to provide links to where the code can be accessed. Table 3 provides links for metrics that participated in WMT17 that are publicly available for download.

2.4.1 AUTODA, AUTODA.TECTO

AUTODA (Mareček et al., 2017) is a sentence-level metric trainable on any direct assessment scores. The metric is based on a simple linear regressor combining several features extracted from the automatically aligned an parsed translation-reference pair. The language-universal AUTODA uses seven features based on word-aligned parse trees in Universal Dependencies style (Nivre et al., 2016). All the features are some kind of similarity measures between two aligned nodes, e.g. lemma similarity, tag similarity, or morphosyntactic features similarity. The eighth feature used is the CHRF3 score (Popović, 2015). For the newstest2017 data, AUTODA was trained on Direct Assessment scores from newstest2015, which were available only for English. Nevertheless the same model was used for all the language pairs. For himltest2017, the metrics were trained on the provided HUMEseg2016.

The AUTODA.TECTO metric is similar to AUTODA but uses tectogrammatical trees (Hajič, 2004) instead of the Universal Dependencies. This

[7] Here the relation $<$ always means "is better than" even for metrics where the better system receives a higher score.

Metric	Seg-level	Sys-level	Hybrids	Participant
AUTODA	•	⊘	⊘	Charles University (Mareček et al., 2017)
AUTODA.TECTO	•	⊘	⊘	Charles University (Mareček et al., 2017)
BEER	•	⊘	⊘	ILLC – University of Amsterdam (Stanojević and Sima'an, 2015)
BLEND	•	⊘	⊘	ICTCAS-DCU (Ma et al., 2017)
BLEU2VEC_SEP	•	•	–	University of Tartu (Tättar and Fishel, 2017)
CHARACTER	–	•	•	RWTH Aachen University (Wang et al., 2016)
CHRF	•	⊘	⊘	(Popović, 2015)
CHRF+	•	⊘	⊘	(Popović, 2017)
CHRF++	•	⊘	⊘	(Popović, 2017)
MEANT_2.0	•	⊘	⊘	NRC (Lo, 2017)
MEANT_2.0-NOSRL	•	⊘	⊘	NRC (Lo, 2017)
NGRAM2VEC	•	•	–	University of Tartu (Tättar and Fishel, 2017)
TREEAGGREG	•	⊘	⊘	Charles University (Mareček et al., 2017)
UHH_TSKM	•	⊘	⊘	(Duma and Menzel, 2017)

Table 2: Participants of WMT17 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 average of segment-level scores.

AUTODA incl. TECTO	`http://github.com/ufal/auto-hume`
BEER	`http://github.com/stanojevic/beer`
BLEND	`http://github.com/qingsongma/blend`
BLEU2VEC_SEP	`http://github.com/TartuNLP/bleu2vec`
CHARACTER	`http://github.com/rwth-i6/CharacTER`
CHRF, incl. + and ++	`http://github.com/m-popovic/chrF`
MEANT_2.0 incl. NOSRL	`http://chikiu-jackie-lo.org/home/index.php/meant`
NGRAM2VEC	`http://github.com/TartuNLP/bleu2vec`
TREEAGGREG	`http://github.com/ufal/auto-hume/tree/rudolf`
Baselines:	`http://github.com/moses-smt/mosesdecoder`
BLEU, NIST	`scripts/generic/mteval-v13a.pl`
CDER, PER, TER, WER	`mert/evaluator`
SENTBLEU	`mert/sentence-bleu`

Table 3: Metrics available for public download that participated in WMT17. The baseline metrics scripts are all available with Moses, relative paths are listed.

very rich annotation allowed to use also the deep-syntactic features. It uses 18 features based on aligned tectogrammatical nodes similarity and two additional measures: CHRF3 and BLEU. The AUTODA.TECTO metric was applied only to the Czech outputs and it was trained on HUME-seg2016 en-cs data.

The AUTODA metrics are labelled as ensemble metrics because they include the scores of CHRF3 and BLEU.

2.4.2 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. The metric is identical to the 2016 run, including the training, so no 2016 data were used to train BEER in 2017.

2.4.3 BLEND

BLEND (Ma et al., 2017) is a novel combined metric that takes good advantage of merits of existing metrics. Contrary to another combined metric DPMFcomb (Yu et al., 2015), BLEND employs SVM regression for training, with DA scores as the gold standard in order to adapt to the new development of human evaluation. Experiments on WMT16 to-English language pairs show that, with a vast reduction in required training data,

BLEND still achieves improved performance over DPMFcomb when incorporated the same metrics. BLEND also finds a trade-off between its performance and efficiency by exploring the contribution of incorporated metrics. Besides, BLEND is flexible to be applied to any language pairs if incorporated metrics support the specific language pair.

BLEND is an ensemble metric, building upon scores provided by 25 lexical based metrics and 4 other metrics for to-English language pairs. 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. 4 other metrics include CharacTer, BEER, DPMF and ENTF.

BLEND for en-ru incorporates 20 lexical based metrics (the same 9 kinds of metrics mentioned above), and 2 other metrics, namely CharacTer and BEER.

2.4.4 BLEU2VEC_SEP, NGRAM2VEC

The metrics BLEU2VEC_SEP and NGRAM2VEC (Tättar and Fishel, 2017) are token-level metrics, which are trained on raw monolingual corpora. They are a direct modification of the original BLEU metric (Papineni et al., 2002) with fuzzy matches added to strict matches. The fuzzy match score is implemented via token and n-gram embedding similarities and applied to same-length n-grams in the hypothesis and reference(s).

2.4.5 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 applied to non-tokenized outputs and references, which also holds for this year's submission.

2.4.6 CHRF, CHRF+, and CHRF++

CHRF (Popović, 2015) is an evaluation metric which compares character n-grams in the hypothesis with those in the reference. Previous experiments have shown that the optimal set-up is to use maximal character n-gram length of 6 with uniform n-gram weights, arithmetic n-gram averaging and beta parameter set to 2. It has participated in previous two years of the evaluation task. This year's CHRF is identical to the CHRF2 from the 2016 metric task.

CHRF+ and CHRF++ (Popović, 2017) are extended CHRF metrics which, in addition to character n-grams, also compare word unigrams (CHRF+) and bigrams (CHRF++).

2.4.7 MEANT_2.0, MEANT_2.0-NOSRL

MEANT_2.0 is a non-trained evaluation metric that uses distributional word vector model to evaluate lexical semantic similarity and shallow semantic parses to evaluate structural semantic similarity between the reference and the MT output. It is a new version of MEANT (Lo et al., 2015) with improved evaluation of semantic role fillers phrasal similarity using idf-weighted n-gram similarity. Another improvement in MEANT_2.0 is its no-srl variant, MEANT_2.0-NOSRL. It provides accurate semantic evaluation of machine translation in any output language, even if no shallow semantic parser is available in that language. It considers the whole sentences as one long phrase for computing the phrasal similarity and the evaluation score.

2.4.8 TREEAGGREG

TREEAGGREG (Mareček et al., 2017) is an n-gram based metric computed over aligned syntactic structures instead of the linear representation of the translated sentences. Sentences are segmented into phrases based on their dependency parse trees, evaluating each of these phrases independently using CHRF3 metric (Popović, 2015). The resulting scores are then aggregated into a final sentence-level score using a simple weighted average.

TREEAGGREG is labelled as an ensemble metric, because it builds upon CHRF. It is however not trained at all, it only follows the dependency structure of the reference and candidate translation.

2.4.9 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 sub-structures 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.

2.4.10 Baseline Metrics

As mentioned by Bojar et al. (2016a), 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":

- **Mteval.** The metrics BLEU (Papineni et al., 2002) and NIST (Doddington, 2002) were computed using the script `mteval-v13a.pl`[8] 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).
- **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. 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 included in future tasks.

As for segment-level baselines, we employ the following modified version of BLEU:

- **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.

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[9] 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. New scripts have been added for generation of hybrid systems from the given hybrid descriptions.

3 Results

We discuss system-level results for news task systems (including NMT training task systems) in Section 3.1. The segment-level results are in Section 3.2.

3.1 System-Level Results

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

$$r = \frac{\sum_{i=1}^{n}(H_i - \overline{H})(M_i - \overline{M})}{\sqrt{\sum_{i=1}^{n}(H_i - \overline{H})^2}\sqrt{\sum_{i=1}^{n}(M_i - \overline{M})^2}} \quad (3)$$

where H_i are human assessment scores of all systems in a given translation direction, M_i are 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.

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

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

	cs-en	de-en	fi-en	lv-en	ru-en	tr-en	zh-en
n	4	11	6	9	9	10	16
Correlation	$\|r\|$	$\|r\|$	$\|r\|$	$\|r\|$	$\|r\|$	$\|r\|$	$\|r\|$
AUTODA	**0.438**	0.959	0.925	0.973	0.907	0.916	0.734
BEER	0.972	**0.960**	**0.955**	**0.978**	0.936	0.972	**0.902**
BLEND	0.968	**0.976**	**0.958**	**0.979**	**0.964**	**0.984**	**0.894**
BLEU	0.971	0.923	0.903	**0.979**	0.912	**0.976**	0.864
BLEU2VEC_SEP	**0.989**	**0.936**	**0.888**	0.966	0.907	**0.961**	**0.886**
CDER	0.989	0.930	**0.927**	**0.985**	0.922	**0.973**	**0.904**
CHARACTER	0.972	**0.974**	**0.946**	0.932	**0.958**	0.949	**0.799**
CHRF	0.939	**0.968**	**0.938**	0.968	0.952	0.944	0.859
CHRF++	0.940	**0.965**	0.927	**0.973**	0.945	0.960	0.880
MEANT_2.0	0.926	0.950	**0.941**	**0.970**	**0.962**	0.932	0.838
MEANT_2.0-NOSRL	0.902	0.936	0.933	0.963	**0.960**	0.896	0.800
NGRAM2VEC	0.984	0.935	0.890	0.963	0.907	0.955	**0.880**
NIST	**1.000**	0.931	0.931	0.960	0.912	0.971	0.849
PER	0.968	**0.951**	**0.896**	0.962	0.911	0.932	**0.877**
TER	0.989	0.906	**0.952**	**0.971**	0.912	0.954	0.847
TREEAGGREG	0.983	0.920	**0.977**	**0.986**	0.918	**0.987**	0.861
UHH_TSKM	**0.996**	0.937	**0.921**	**0.990**	0.914	**0.987**	**0.902**
WER	0.987	0.896	**0.948**	**0.969**	0.907	0.925	**0.839**
	newstest2017						

Table 4: Absolute Pearson correlation of to-English system-level metrics with DA human assessment; 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-fi	en-lv	en-ru	en-tr	en-zh
n	14	16	12	17	9	8	11
Correlation	$\|r\|$	$\|r\|$	$\|r\|$	$\|r\|$	$\|r\|$	$\|r\|$	$\|r\|$
AUTODA	**0.975**	0.603	0.879	0.729	0.850	0.601	**0.976**
AUTODA-TECTO	**0.969**	–	–	–	–	–	–
BEER	**0.970**	0.842	0.976	**0.930**	0.944	**0.980**	0.914
BLEND	–	–	–	–	**0.953**	–	–
BLEU	0.956	0.804	0.920	0.866	0.898	0.924	**0.981**
BLEU2VEC_SEP	**0.963**	0.810	0.942	0.859	0.903	0.911	–
CDER	**0.968**	0.813	**0.965**	**0.930**	0.924	0.957	**0.983**
CHARACTER	**0.981**	**0.938**	**0.972**	0.897	**0.939**	**0.975**	0.933
CHRF	**0.976**	0.863	**0.981**	**0.955**	0.950	**0.991**	**0.976**
CHRF+	**0.976**	0.855	**0.980**	**0.956**	0.948	**0.988**	–
CHRF++	**0.974**	0.852	**0.979**	**0.956**	0.945	0.986	**0.976**
MEANT_2.0	–	0.858	–	–	–	–	**0.956**
MEANT_2.0-NOSRL	**0.976**	0.770	**0.972**	**0.959**	**0.957**	**0.991**	0.943
NGRAM2VEC	–	–	0.940	0.862	–	–	–
NIST	0.962	0.769	0.957	**0.935**	0.920	**0.986**	**0.976**
PER	**0.954**	0.687	0.949	0.851	0.887	**0.963**	0.934
TER	**0.955**	0.796	0.961	**0.909**	**0.933**	**0.967**	**0.970**
TREEAGGREG	0.947	0.773	0.965	0.927	0.921	**0.983**	0.938
WER	**0.954**	0.802	0.960	**0.906**	**0.934**	0.956	**0.954**
	newstest2017						

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

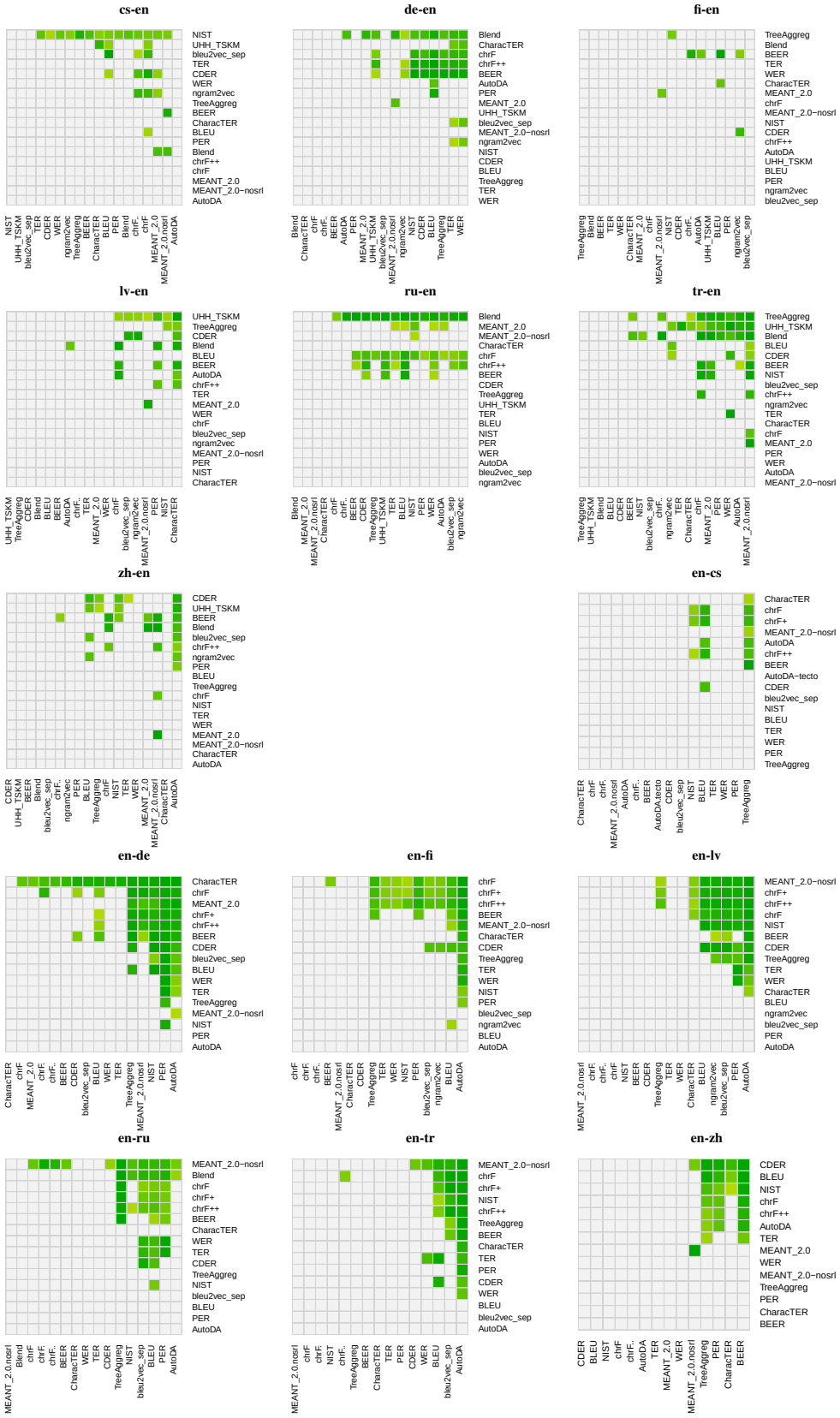


Figure 1: System-level metric significance test results for DA human assessment in newstest2017; green cells denote a statistically significant increase in correlation with human assessment for the metric in a given row over the metric in a given column according to Williams test.

	cs-en	de-en	fi-en	lv-en	ru-en	tr-en	zh-en
n	10K	10K	10K	10K	10K	10K	10K
Correlation	$\|r\|$	$\|r\|$	$\|r\|$	$\|r\|$	$\|r\|$	$\|r\|$	$\|r\|$
AUTODA	0.4395	0.9505	0.9220	0.9698	0.9015	0.9138	0.7341
BEER	0.9662	0.9524	0.9532	0.9740	0.9299	0.9692	**0.8970**
BLEND	0.9633	**0.9685**	0.9562	0.9761	**0.9569**	0.9809	0.8897
BLEU	0.9644	0.9136	0.9061	0.9741	0.9070	0.9688	0.8523
CDER	0.9833	0.9219	0.9247	0.9814	0.9160	0.9702	**0.8975**
CHARACTER	0.9628	0.9648	0.9438	0.9271	0.9484	0.9459	0.7398
CHRF	0.9330	0.9602	0.9352	0.9647	0.9456	0.9408	0.8551
CHRF++	0.9348	0.9572	0.9242	0.9696	0.9381	0.9568	0.8756
MEANT_2.0	0.9209	0.9418	0.9390	0.9668	0.9546	0.9307	0.8357
MEANT_2.0-NOSRL	0.8962	0.9275	0.9305	0.9599	0.9523	0.8951	0.7992
NIST	**0.9937**	0.9173	0.9284	0.9566	0.9035	0.9693	0.8309
PER	0.9673	0.9198	0.8917	0.9578	0.9040	0.8982	0.8659
TER	0.9830	0.8991	0.9503	0.9672	0.9051	0.9510	0.8366
TREEAGGREG	0.9769	0.9133	**0.9752**	0.9828	0.9115	**0.9834**	0.8535
UHH_TSKM	0.9896	0.9294	0.9183	**0.9857**	0.9077	0.9821	0.8955
WER	0.9814	0.8894	0.9458	0.9649	0.9004	0.9222	0.8281

newstest2017 Hybrids

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

	en-cs	en-de	en-fi	en-lv	en-ru	en-tr	en-zh
n	10K	10K	10K	10K	10K	10K	10K
Correlation	$\|r\|$	$\|r\|$	$\|r\|$	$\|r\|$	$\|r\|$	$\|r\|$	$\|r\|$
AUTODA	0.9670	0.6021	0.8789	0.7307	0.8501	0.5857	0.9676
AUTODA-TECTO	0.8572	–	–	–	–	–	–
BEER	0.9634	0.8285	0.9748	0.9233	0.9417	0.9684	0.9062
BLEND	–	–	–	–	0.9499	–	–
BLEU	0.9447	0.7925	0.9190	0.8385	0.8929	0.9157	0.9686
CDER	0.9582	0.8030	0.9620	0.9111	0.9215	0.9484	**0.9748**
CHARACTER	**0.9725**	**0.8931**	0.9698	0.8921	0.9292	0.9609	0.9140
CHRF	0.9683	0.8446	**0.9788**	0.9445	0.9474	**0.9801**	0.9686
CHRF+	0.9679	0.8375	0.9779	0.9455	0.9453	0.9779	–
CHRF++	0.9658	0.8354	0.9774	0.9441	0.9423	0.9752	0.9683
MEANT_2.0	–	0.8437	–	–	–	–	0.9444
MEANT_2.0-NOSRL	0.9682	0.7530	0.9704	**0.9470**	**0.9550**	0.9796	0.9310
NIST	0.9544	0.7607	0.9567	0.9140	0.9167	0.9760	0.9681
PER	0.9599	0.6803	0.9388	0.8169	0.8758	0.9546	0.8928
TER	0.9507	0.7899	0.9593	0.8881	0.9299	0.9582	0.9646
TREEAGGREG	0.9419	0.7648	0.9630	0.9149	0.9188	0.9712	0.9331
WER	0.9489	0.7967	0.9589	0.8841	0.9310	0.9466	0.9507

newstest2017 Hybrids

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

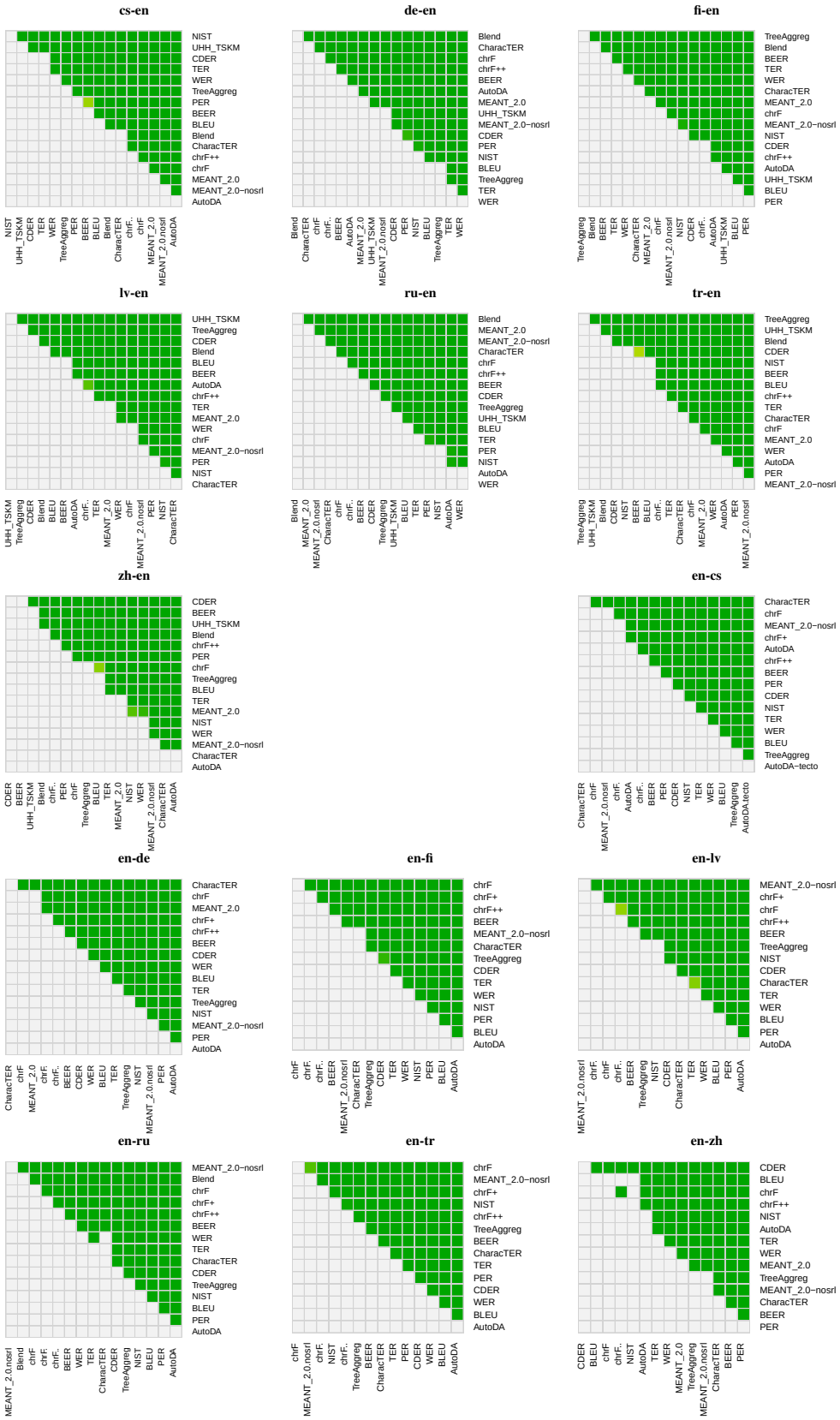


Figure 2: System-level metric significance test results for 10K hybrid systems (DA human evaluation) from newstest2017; green cells denote a statistically significant increase in correlation with human assessment for the metric in a given row over the metric in a given column according to Williams test.

3.1.1 System-Level Results for News Task

Table 4 provides the system-level correlations of metrics evaluating translation of newstest2017 into English while Table 5 provides the same for out-of-English language pairs. DA is the golden truth. The underlying texts are part of the WMT17 News Translation test set (newstest2017) and the underlying MT systems are all MT systems participating in the WMT17 news translation task. The en-cs translation direction also includes the translation systems participating in the NMT training task.

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 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.

For instance, we see that for en-cs (outputs of 14 MT systems), even the best-performing metric CHARACTER was not significantly better than any other metric except TREEAGGREG. CHRF+ and CHRF++ were significantly better than BLEU and TREEAGGREG, as were several other metrics.

The sample of systems we employ to evaluate metrics is often small, as few as four 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. In addition, the Williams test takes into account the correlation between each pair of metrics and the correlation between the metric scores themselves increases the likelihood of a significant difference being identified. For cs-en, this led to one counter-intuitive result: AUTODA achieved a substantially lower correlation with human assessment compared to other metrics (0.438 compared to $\sim$0.9 in Table 4) and yet it was not significantly outperformed by any other metric. The lack of significance here is due to the small sample size and lack of correlation of metric AUTODA metric scores with the scores of the other competing metrics, reducing the likelihood of identifying a significant difference. In short, AUTODA differed too much from others, underperforming, but the four underlying MT systems are too few for the statistical significance. Other metrics are more similar to each other and the differences are sufficient for confidence as to which metric performs better. The small sample size also explains the cs-en NIST correlation of 1.0.

The situation is also interesting for de-en, with BLEND significantly outperforming numerous metrics but the second CHARACTER not being better than any other metric, and this is in part again due to the varying correlations between the metric scores themselves, as the statistical power of Williams test increases with stronger metric scores correlations between each other.

We also 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 WMT17 translation task and NMT training 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 most metrics this year, so where ever possible, we provided scores for hybrids ourselves.

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 also includes significance test results for hybrid super-sampled correlations for all pairs of competing metrics for a given language pair.

3.1.2 System-Level Results for HUME

In addition to the WMT17 news task, we also assess the performance of metrics on the system-level for himltest datasets. Tables 8 and 9 show correlation with human assessment of system-level metrics with HUME scores on himltest2017 "a" and "b", respectively. Since there are only two or three systems in each dataset, the sample size is too small to test for statistical significance. In fact,

	en-cs	en-de	en-pl	en-ro
n	3	3	3	3
Correlation	$\|r\|$	$\|r\|$	$\|r\|$	$\|r\|$
AUTODA	0.932	0.593	0.161	0.594
AUTODA-TECTO	0.917	–	–	–
BEER	0.833	0.460	0.342	0.188
BLEU	0.815	0.537	0.675	0.064
CDER	0.751	0.461	0.211	0.285
CHARACTER	0.958	0.735	0.241	0.961
CHRF	0.855	0.631	0.131	0.119
CHRF+	0.840	0.616	0.006	0.168
CHRF++	0.836	0.573	0.119	0.172
MEANT_2.0	–	0.851	–	–
MEANT_2.0-NOSRL	0.812	0.805	0.555	0.331
NIST	0.730	0.484	0.427	0.283
PER	0.704	0.738	0.853	0.239
TER	0.778	0.127	0.838	0.253
TREEAGGREG	0.753	0.799	0.670	0.018
WER	0.784	0.011	0.839	0.151

himltest2017a

Table 8: Absolute Pearson correlation of system-level metrics with HUME human assessment; ensemble metrics are highlighted in gray.

	en-ro
n	2
Correlation	$\|r\|$
BEER	1.000
BLEU	1.000
CDER	1.000
CHARACTER	1.000
CHRF	1.000
CHRF+	1.000
CHRF++	1.000
MEANT_2.0-NOSRL	1.000
NIST	1.000
PER	1.000
TER	1.000
TREEAGGREG	1.000
WER	1.000

himltest2017b

Table 9: Absolute Pearson correlation of system-level metrics with HUME human assessment; ensemble metrics are highlighted in gray.

results in Table 9 are not very informative because two systems will always lie on a line, producing perfect absolute Pearson correlations. We include results nonetheless for demonstration purposes.

To obtain more meaningful results, we compute correlations for 10K hybrid systems for himltest2017a. Table 10 shows metric correlation with human assessment for the large set of 10K hybrid systems for himltest2017a and Figure 3 shows significance test results. Since a minimum of three systems is required for hybrid super-sampling and only two systems were included in himltest2017b, no hybrid results are reported for that test set.

3.2 Segment-Level Results

3.2.1 Segment-Level Results for News Task

In WMT17, since manual evaluation in the news task now takes the form of Direct Assessment of translations, this forms the basis of our segment-level metrics task results for the newstest2017 data set. Note however, that the sampling of the sentences is different, as described in Section 2.3.2. We follow the methodology outlined in Graham et al. (2015) and combine a minimum of 15 individual DA scores for a given translation by taking its average score. We then compute the absolute Pearson correlation between segment-level metric scores and segment-level DA scores where a

	en-cs	en-de	en-pl	en-ro
n	10K	10K	10K	10K
Correlation	$\|r\|$	$\|r\|$	$\|r\|$	$\|r\|$
AUTODA	**0.8700**	0.2266	0.1781	0.3494
AUTODA-TECTO	0.8451	–	–	–
BEER	0.7803	0.0976	0.1859	0.0808
BLEU	0.7732	0.1546	0.4385	0.0020
CDER	0.7124	0.0911	0.2383	0.2025
CHARACTER	**0.8683**	0.3900	0.0527	**0.5881**
CHRF	0.8006	0.2712	0.0043	0.0405
CHRF+	0.7887	0.2564	0.0960	0.0763
CHRF++	0.7869	0.2131	0.1912	0.0794
MEANT2.0	–	0.5484	–	–
MEANT2.0-NOSRL	0.7697	0.4630	0.4447	0.1831
NIST	0.6987	0.0559	0.3276	0.1989
PER	0.6672	0.3897	0.2342	0.0366
TER	0.7252	0.2197	0.5812	0.1686
TREEAGGREG	0.7044	**0.7337**	0.4915	0.0524
WER	0.7287	0.3268	**0.5896**	0.0971

himltest2017a Hybrids

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

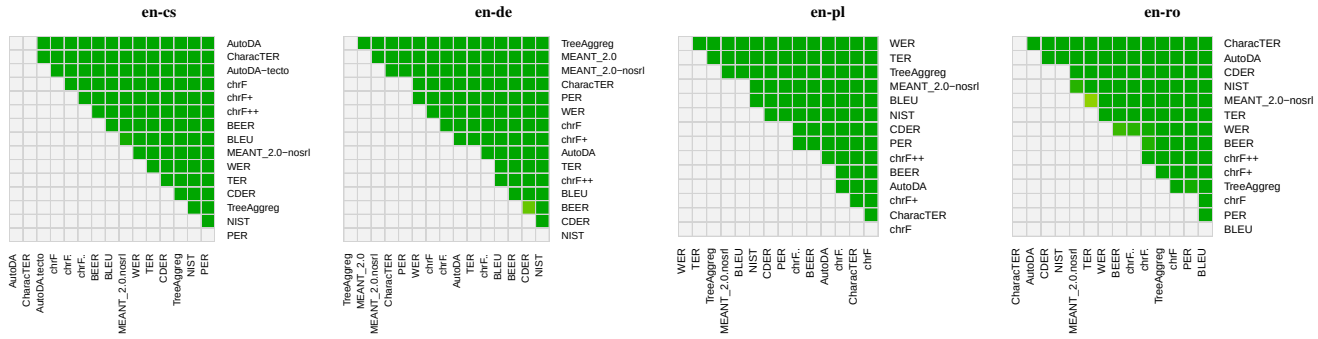


Figure 3: System-level metric significance test results for 10K hybrid systems (HUME human evaluation) from himltest2017a; green cells denote a statistically significant increase in correlation with human assessment for the metric in a given row over the metric in a given column according to Williams test.

stronger correlation indicates higher performance.

As described in Section 2.3.2, for some language pairs, insufficient human assessments were completed to provide accurate segment-level DA scores for segment-level evaluation. For those five language pairs, en-cs, en-de, en-fi, en-lv and en-tr, we therefore convert pairs of DA to DARR better/worse preferences and employ a Kendall's Tau formulation as in previous WMT metric evaluations.

Results of the segment-level human evaluation for translations sampled from the news task are shown in Tables 11 and 12, 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 4.

3.2.2 Segment-Level Results for HUME

For the himltest2017 datasets, we employ segment-level HUME scores also using absolute Pearson correlation.

Results of segment-level metrics task evaluated with HUME on the himltest datasets are shown in Tables 13 and 14 where metrics not significantly outperformed by any other in a given language pair are again highlighted in bold. Head-to-head significance test results for all metrics are shown in Figures 5 and 6.

4 Discussion

The major switch from RR to DA that happened this year in the main news task evaluation did not affect metrics task in any negative way, also because we trialed DA in metrics evaluation already last year.

We discuss various particular observations in the rest of this section.

4.1 Obtaining Human Judgements

The sentence sampling for segment-level evaluation is different from the sampling used to obtain system-level scores. We were aware of the difficulties in finding assessors for some language pairs on the crowdsourcing platforms, as mentioned e.g. by Birch et al. (2016), and we relied on researchers. We were indeed able to cover all the required target languages but for many of them, insufficient numbers of assessments were collected. Fortunately, DA allows to resort to a relative-ranking re-interpretation, DARR, and use a variation of Kendall's τ as in the previous years. This method proved effective and only English-Turkish segment-level evaluation suffers from having all metrics indistinguishable.

4.2 Hybrid Super-sampling vs. Document-level Evaluation

As in the previous year, hybrid super-sampling proved very effective and allowed to obtain conclusive results of system-level evaluation even for language pairs where as few as 4 MT systems participated.

We should however note that this style of aggregated evaluation may not be a substitute for truly document-level evaluation. Hybrid systems are constructed by randomly mixing sentence and they therefore may possibly break cross-sentence links in MT outputs (if such links are at all preserved by current MT systems). There is a good chance that document-level links are well represented in individual sentences of the reference, as these were created taking the whole document into

	cs-en	de-en	fi-en	lv-en	ru-en	tr-en	zh-en
Human Evaluation	DA	DA	DA	DA	DA	DA	DA
n	560	560	560	560	560	560	560
Correlation	$\|r\|$	$\|r\|$	$\|r\|$	$\|r\|$	$\|r\|$	$\|r\|$	$\|r\|$
AUTODA	0.499	**0.543**	0.673	0.533	0.584	0.625	0.583
BEER	0.511	0.530	0.681	0.515	0.577	0.600	0.582
BLEND	**0.594**	**0.571**	**0.733**	**0.577**	**0.622**	**0.671**	**0.661**
BLEU2VEC_SEP	0.439	0.429	0.590	0.386	0.489	0.529	0.526
CHRF	0.514	0.531	0.671	0.525	**0.599**	0.607	0.591
CHRF++	0.523	0.534	0.678	0.520	0.588	0.614	0.593
MEANT_2.0	**0.578**	**0.565**	0.687	**0.586**	**0.607**	0.596	**0.639**
MEANT_2.0-NOSRL	**0.566**	**0.564**	0.682	0.573	0.591	0.582	0.630
NGRAM2VEC	0.436	0.435	0.582	0.383	0.490	0.538	0.520
SENTBLEU	0.435	0.432	0.571	0.393	0.484	0.538	0.512
TREEAGGREG	0.486	0.526	0.638	0.446	0.555	0.571	0.535
UHH_TSKM	0.507	0.479	0.600	0.394	0.465	0.478	0.477
	newstest2017						

Table 11: Segment-level metric results for to-English language pairs: absolute correlation 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-fi	en-lv	en-ru	en-tr	en-zh
Human Evaluation	DARR	DARR	DARR	DARR	DA	DARR	DA
n	32,810	3,227	3,270	3,456	560	247	560
Correlation	τ	τ	τ	τ	$\|r\|$	τ	$\|r\|$
AUTODA	0.041	0.099	0.204	0.130	0.511	**0.409**	0.609
AUTODA-TECTO	0.336	–	–	–	–	–	–
BEER	**0.398**	**0.336**	**0.557**	**0.420**	0.569	**0.490**	0.622
BLEND	–	–	–	–	0.578	–	–
BLEU2VEC_SEP	0.305	**0.313**	0.503	0.315	0.472	**0.425**	–
CHRF	0.376	**0.336**	0.503	**0.420**	0.605	**0.466**	0.608
CHRF+	0.377	**0.325**	0.514	**0.421**	0.609	**0.474**	–
CHRF++	0.368	**0.328**	0.484	**0.417**	0.604	**0.466**	0.602
MEANT_2.0	–	**0.350**	–	–	–	–	**0.727**
MEANT_2.0-NOSRL	**0.395**	**0.324**	**0.565**	**0.425**	**0.636**	**0.482**	0.705
NGRAM2VEC	–	–	0.486	0.317	–	–	–
SENTBLEU	0.274	0.269	0.446	0.259	0.468	**0.377**	0.642
TREEAGGREG	0.361	**0.305**	0.509	**0.383**	0.535	**0.441**	0.566
	newstest2017						

Table 12: Segment-level metric results for out-of-English language pairs: absolute correlation of segment-level metric scores with human assessment variants, where τ are computed similar to Kendall's τ and over relative ranking (RR) human assessments (converted from DA scores); $|r|$ are absolute Pearson correlation coefficients of metric scores with DA scores; correlations of metrics not significantly outperformed by any other are highlighted in bold; ensemble metrics are highlighted in gray.

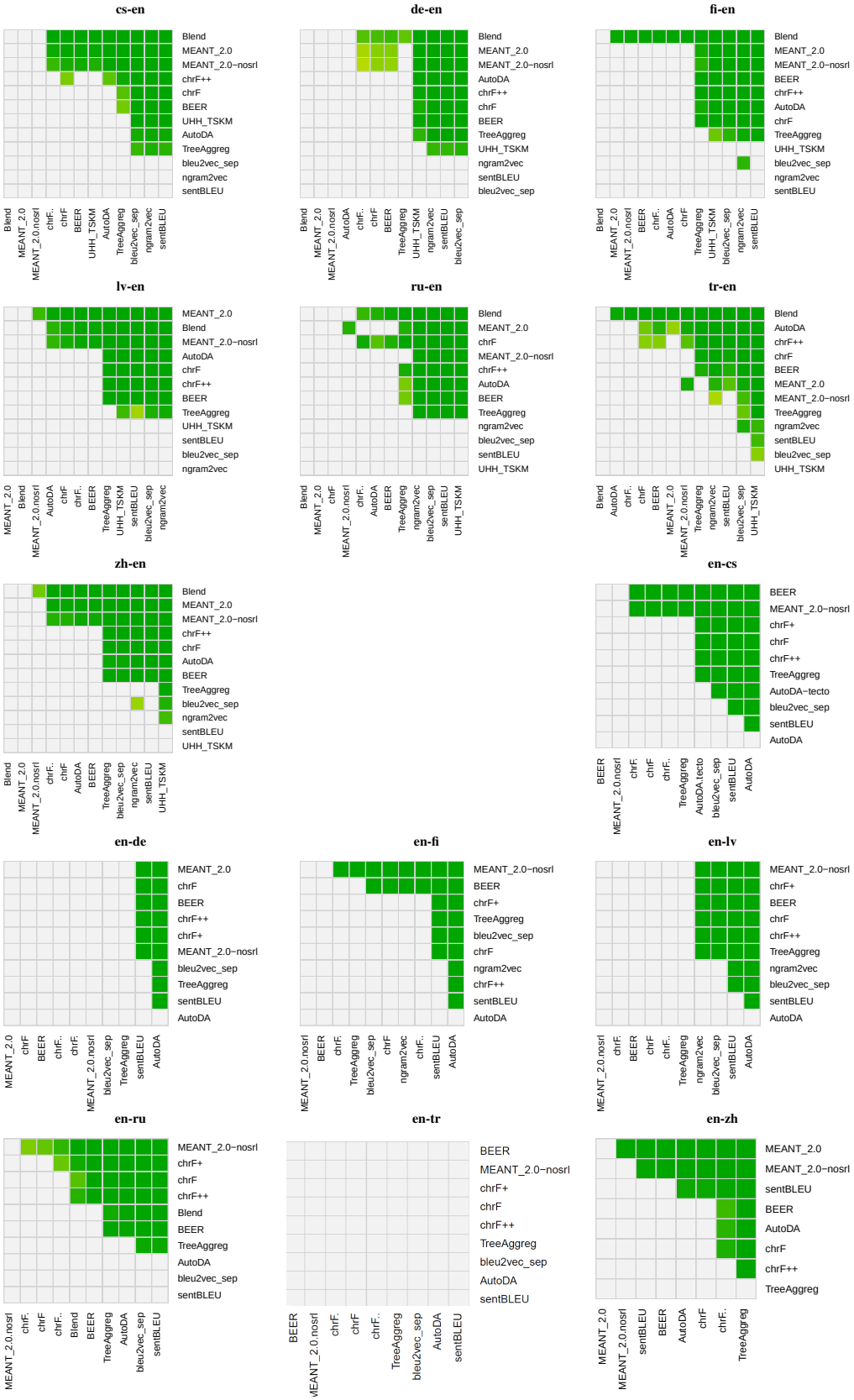


Figure 4: Direct Assessment (DA) and DARR segment-level metric significance test results for all language pairs (newstest2017): Green cells denote a significant win for the metric in a given row over the metric in a given column according to Williams test for DA (all to-English language pairs; en-ru; en-zh) and bootstrap resampling for DARR (en-cs; en-de; en-fi; en-ro; en-tr).

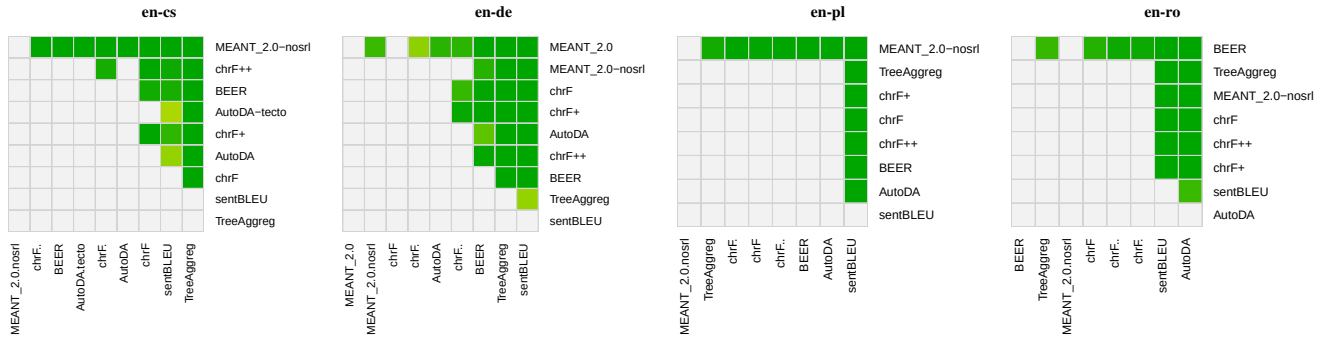


Figure 5: HUME segment-level metric significance test results (himltest2017a): Green cells denote a significant win for the metric in a given row over the metric in a given column according to Williams test for difference in dependent correlation.

	en-cs	en-de	en-pl	en-ro
n	879	891	1,020	354
Correlation	$\|r\|$	$\|r\|$	$\|r\|$	$\|r\|$
AUTODA	0.391	0.445	0.442	0.127
AUTODA-TECTO	0.400	–	–	–
BEER	0.400	0.428	0.442	**0.508**
CHRF	0.383	**0.454**	0.445	0.477
CHRF+	0.395	0.451	0.445	0.474
CHRF++	0.400	0.445	0.444	0.477
MEANT_2.0	–	**0.479**	–	–
MEANT_2.0-NOSRL	**0.473**	0.463	**0.489**	**0.479**
SENTBLEU	0.347	0.338	0.329	0.261
TREEAGGREG	0.323	0.374	0.450	0.481
himltest2017a				

Table 13: Absolute Pearson correlation of segment-level metric scores with HUME scores for himltest2017a; ensemble metrics are highlighted in gray.

	en-ro
n	350
Correlation	$\|r\|$
BEER	0.293
CHRF	0.305
CHRF+	0.314
CHRF++	0.310
MEANT_2.0-NOSRL	**0.370**
SENTBLEU	0.254
TREEAGGREG	0.244
himltest2017b	

Table 14: Absolute Pearson correlation of segment-level metric scores with HUME scores for himltest2017b; ensemble metrics are highlighted in gray.

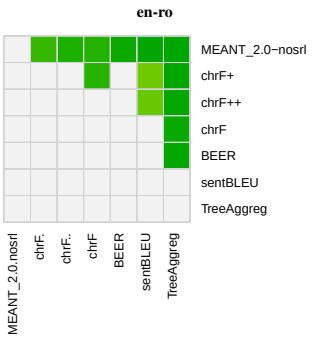


Figure 6: HUME segment-level metric significance test results (himltest2017b): Green cells denote a significant win for the metric in a given row over the metric in a given column according to Williams test for difference in dependent correlation.

account, but this would have to be empirically validated.

4.3 Overall Metric Performance

As mentioned above, the observed performance of metrics very much depends on the underlying texts and participating MT systems. We can nevertheless confirm the trend since 2014, with character-level metrics performing on average better: BEER, CHRF (and its variants) and CHARACTER.

In order to get an idea of the stability of metrics at achieving a high correlation with human assessment across all language pairs, Figure 7 shows box plots of correlations achieved by metrics.[10]

[10]We only include metrics that participated in all language pairs in each box plot, to provide a fair indication of metric performance, otherwise metrics not participating in difficult language pairs could (unfairly) appear to perform better when they did not participate in that language.

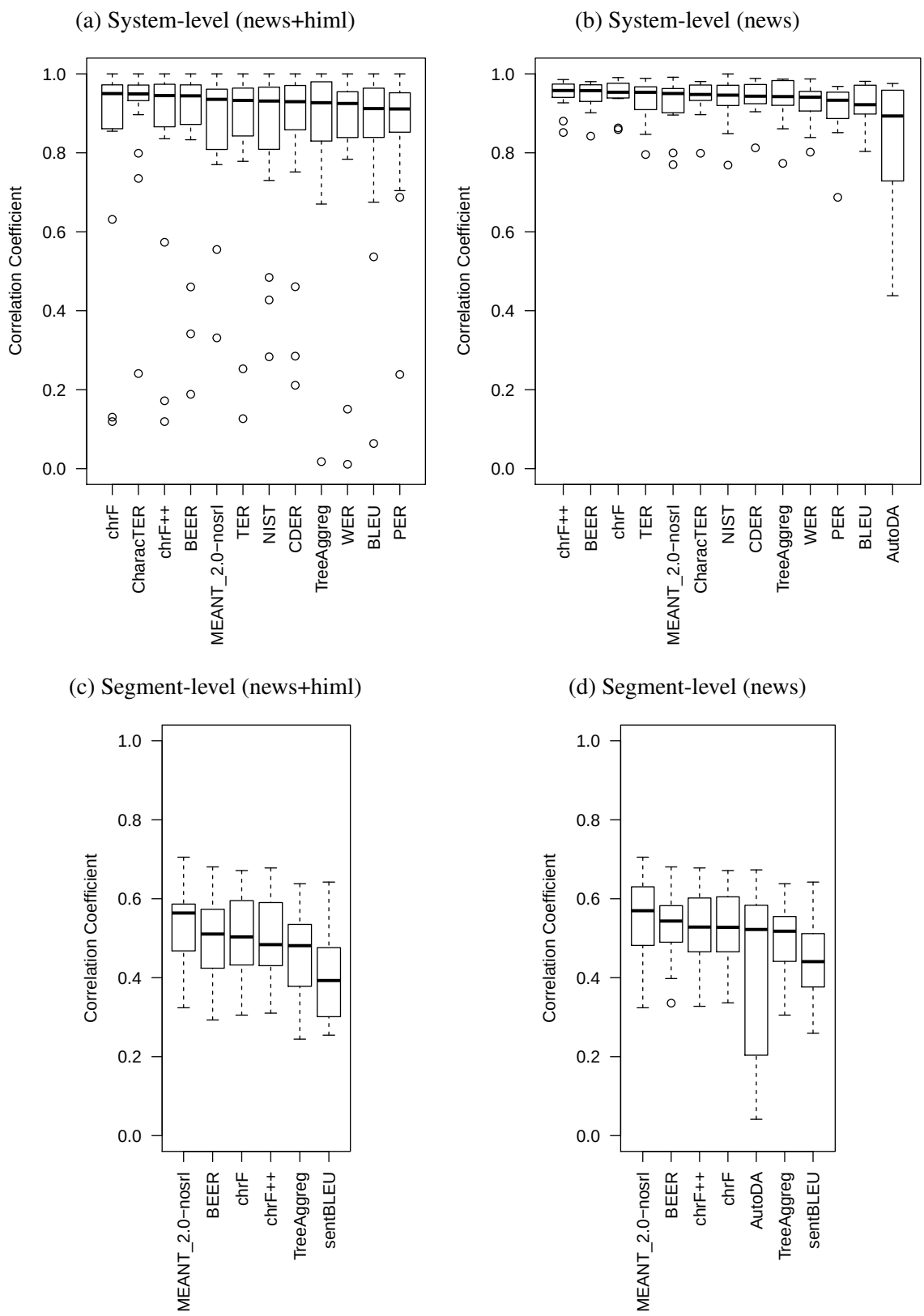


Figure 7: Plots of correlations achieved by metrics in (a) all language pairs and test sets on the system level; (b) all language pairs for newstest2017 on the system level; (c) all language pairs and test sets on the segment level; (d) all language pairs for newstest2017 on the segment-level; all correlations are for non-hybrid correlations only.

The figures confirm the observation from the past years that system-level metrics can achieve correlations above 0.9 while segment-level evaluation is only around 0.5 or slightly above. The variance in the achieved correlations across language pairs and test sets is generally acceptable, with only AUTODA getting very varied results. Comparing the plots (a) and (b) in Figure 7, we see that himl datasets allowed only for less stable results, possibly due to the smaller number of translations comprising test sets for himl. For system-level newstest, plot Figure 7(b), the variance of the majority of metrics is very low, indicating that their scores are reliable across language pairs.

The generally well-performing and stable metrics are CHRF or CHRF++, CHARACTER and BEER. MEANT_2.0-NOSRL is new this year and also performed very well, esp. in segment-level evaluation, although it is currently not yet quite as stable as others on the system-level. Traditional metrics like NIST or TER also reach relatively good results, clearly surpassing BLEU when applied in the common way with only 1 reference and not 4 as recommended by the original authors.

All of the "winners" in this years campaign are publicly available, which is very good for a 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.

4.4 Data Overlap for Polish HUME

As mentioned in Section 2.2, HUME evaluation of translation into Polish suffered from a large overlap of training and evaluation data. Fortunately, only AUTODA was actually affected by this, other trained metrics such as BEER, BLEND or NGRAM2VEC either did not evaluate himltest2017 or were not retrained this year.

4.5 HUME Results

The dataset used to evaluate metrics against HUME, himltest2017, is rather small. It contains only ∼300 sentences (and actually only 118 sentences for Romanian, himltest2017a) with three MT system outputs per sentence. The discriminative power of the experiment is correspondingly low.

The segment-level scores in Figures 5 and 6 however still indicate that MEANT_2.0 (in SRL and noSRL variant) performed well, significantly outperforming all others except for Romanian on himltest2017a but still outperforming it on himltest2017b. This result nicely corresponds with the design of the manual scores of HUME, aggregated over key semantic elements of the sentence.

4.6 Metric Efficiency

This year we asked participants to submit information about the speed of their metrics in order to analyze a possible relationship between metric efficiency and performance in terms of correlation with human assessment. Many participants submitted time durations for metrics to process system outputs for the system-level news task test set. Figures 8(a) and 8(b) show scatter-plots of average correlation coefficient achieved by a given metric versus self-reported times to process a single translation (on average).[11]

Based on these plots, we can conclude that the generally good metrics are not prohibitively slow, only MEANT_2.0 being more expensive, needing up to a second per sentence. The plots show all metrics for which times were submitted, regardless the number of language pairs they took part in.

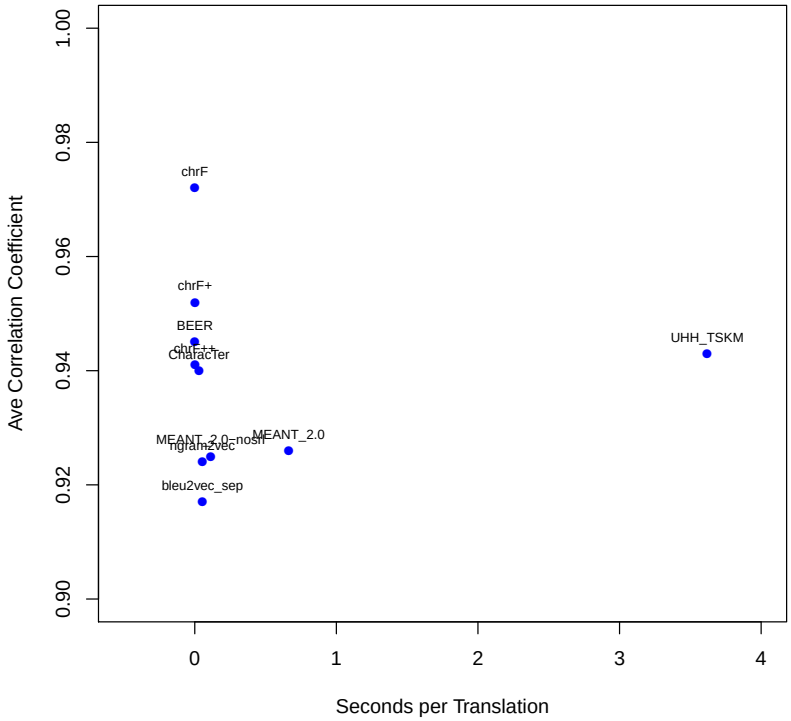
5 Conclusion

This paper summarizes the results of WMT17 shared task in machine translation evaluation, the Metrics Shared Task. Participating metrics were evaluated in terms of their correlation with human judgements 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 on average across several language pairs. For the latter, correlations between 0.4 and 0.6 Pearson's ρ or Kendall's τ are to be expected.

We confirm the main results from the previous year that character-level metrics, or metrics incorporating such a feature, generally perform better. Last year's conclusion that trained metrics generally perform better than non-trained ones is not that clear this year, good performance is observed for both trained metrics like BLEND, BEER (not retrained for this year) as well as non-trained metrics like CHRF, CHARACTER and also a new addition this year, MEANT_2.0.

[11] Some metric participants only submitted times for a subset of language pairs. In such cases, average correlations included in plots are only based on the correlations for which times were submitted.

(a) System-level



(b) Segment-level

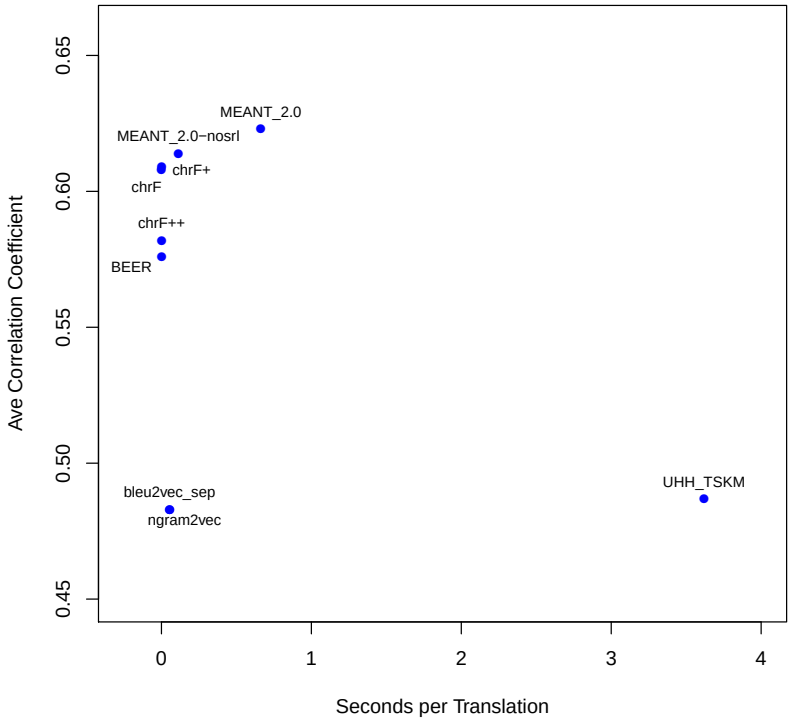


Figure 8: Scatter-plots of self-reported metric speed per translation (computed on the system-level news task datasets) versus average correlation with human assessment for (a) system-level performance and (b) segment-level performance.

Acknowledgments

We wouldn't be able to put this experiment together without tight collaboration with Christian Federmann who ran the core of WMT Shared Translation Task evaluation and also operated Appraise for us.

This study was supported in parts by the grants H2020-ICT-2014-1-645442 (QT21), H2020-ICT-2014-1-644402 (HimL), the Dutch organization for scientific research STW grant nr. 12271, 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.

References

Omri Abend and Ari Rappoport. 2013. Universal Conceptual Cognitive Annotation (UCCA). In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 228–238, Sofia, Bulgaria, August. Association for Computational Linguistics.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473.

Alexandra Birch, Omri Abend, Ondřej Bojar, and Barry Haddow. 2016. HUME: Human UCCA-Based Evaluation of Machine Translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1264–1274, Austin, Texas, November. Association for Computational Linguistics.

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. 2016a. 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 Miloš Stanojević. 2016b. Results of the WMT16 Metrics Shared Task . In *Proceedings of the First Conference on Machine Translation*, Berlin, Germany, August. Association for Computational Linguistics.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Raphael Rubino, Lucia Specia, and Marco Turchi. 2017a. Findings of the 2017 conference on machine translation (wmt17). In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Tasks Papers*, Copenhagen, Denmark, September. Association for Computational Linguistics.

Ondřej Bojar, Jindřich Helcl, Tom Kocmi, Jindřich Libovický, and Tomáš Musil. 2017b. Results of the wmt17 neural mt training task. In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Tasks Papers*, Copenhagen, Denmark, September. 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. 2014. 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. 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.

Jan Hajič. 2004. Complex Corpus Annotation: The Prague Dependency Treebank. In *Insight into Slovak and Czech Corpus Linguistics*, Bratislava, Slovakia. Jazykovedný ústav Ľ. Štúra, SAV.

Jindřich Helcl and Jindřich Libovický. 2017. Neural Monkey: An open-source tool for sequence learning. *The Prague Bulletin of Mathematical Linguistics*, 107:5–17.

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.

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, Philipp Dowling, and Dekai Wu. 2015. Improving evaluation and optimization of MT systems against MEANT. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, Lisboa, Portugal, September. 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.

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.

David Mareček, Ondřej Bojar, Ondřej Hübsch, Rudolf Rosa, and Dusan Varis. 2017. Cuni experiments for wmt17 metrics task. In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Tasks Papers*, Copenhagen, Denmark, September. Association for Computational Linguistics.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Haji, Christopher Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*, pages 1659–1666, Paris, France. European Language Resources Association.

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.

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.

Andre Tättar and Mark Fishel. 2017. bleu2vec: the painfully familiar metric on continuous vector space steroids. In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Tasks Papers*, Copenhagen, Denmark, September. Association for Computational Linguistics.

Yolanda Vazquez-Alvarez and Mark Huckvale. 2002. The reliability of the ITU-t p.85 standard for the evaluation of text-to-speech systems. In *Proc. of ICSLP - INTERSPEECH*.

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*, Berlin, Germany, August. Association for Computational Linguistics.

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

Hui Yu, Qingsong Ma, Xiaofeng Wu, and Qun Liu. 2015. CASICT-DCU Participation in WMT2015 Metrics Task. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, Lisboa, Portugal, September. Association for Computational Linguistics.