Results of the WMT16 Metrics Shared Task

Ondřej Bojar
Charles Univ. in Prague
MFF ÚFAL
bojar@ufal.mff.cuni.cz

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

Amir Kamran and **Miloš Stanojević**
Univ. of Amsterdam
ILLC
{a.kamran,m.stanojevic}@uva.nl

Abstract

This paper presents the results of the WMT16 Metrics Shared Task. We asked participants of this task to score the outputs of the MT systems involved in the WMT16 Shared Translation Task. We collected scores of 16 metrics from 9 research groups. In addition to that, we computed scores of 9 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 WMT16 official manual ranking of systems) and in terms of segment level correlation (how often a metric agrees with humans in comparing two translations of a particular sentence).

This year there are several additions to the setup: large number of language pairs (18 in total), datasets from different domains (news, IT and medical), and different kinds of judgments: relative ranking (RR), direct assessment (DA) and HUME manual semantic judgments. Finally, generation of large number of *hybrid systems* was trialed for provision of more conclusive system-level metric rankings.

1 Introduction

Automatic evaluation of machine translation quality is essential in the development and selection of machine translation systems. Many different automatic MT quality metrics are available and the Metrics Shared Task[1] is held annually at WMT to assess their quality, starting with Koehn and Monz (2006) and following up to Stanojević et al. (2015).

[1] http://www.statmt.org/wmt16/metrics-task/

Metrics participating in the metrics task rely on the existence of reference translations with which MT outputs are compared, and the metrics task itself then needs manual judgments of translation quality in order to check the extent to which the automatic metrics can approximate the judgment. A related WMT task on quality estimation assesses the performance of methods where no reference translations are needed, requiring only the manual quality judgments (Bojar et al., 2016b).

This year, we keep the two main types of metric evaluation: *system-level*, where a metric is expected to provide a quality score for the whole translated document, and *segment-level*, where the score is needed for every individual sentence.

We experiment with several novelties. Specifically, test sets this year come from three domains: *news*, *IT* and *medical/health-related* texts.

The added domains bring in an extended set of languages. In sum, the metrics task this year includes 18 language pairs, English paired with Basque, Bulgarian, Czech, Dutch, Finnish, German, Polish, Portuguese, Romanian, Russian, Spanish, and Turkish, in one or both directions.

On the evaluation side, we rely on three golden truths of manual judgment:

- *Relative Ranking (RR)* of up to 5 different translation candidates at a time, as collected in WMT in the past,
- *Direct Assessment (DA)* evaluating the adequacy of a translation candidate on an absolute scale in isolation from other translations,
- *HUME*, a composite segment-level score aggregated over manual judgments of translation quality of semantic units of the source sentence.

Additional changes to the task evaluation include a change in the way we compute confidence

Proceedings of the First Conference on Machine Translation, Volume 2: Shared Task Papers, pages 199–231,
Berlin, Germany, August 11-12, 2016.

Track	Test set	News Task	Tuning Task	IT Task	HimL Year 1	Hybrid	cs	de	ro	fi	ru	tr	cs	de	ro	fi	ru	tr	bg	es	eu	nl	pl	pt
		Systems					into-English						English into											
RRsysNews	newstest2016	✓	✓			✓	•	•	•	•	•	•	•	•	•	•	•	•						
RRsysIT	it-test2016			✓		✓							•	•					•	•	•	•		•
DAsysNews	newstest2016	✓	✓			✓	•	•	•	•	•	•	·	·	·	·	•	·						
RRsegNews	newstest2016	✓	✓				•	•	•	•	•	•	•	•	•	•	•	•						
DAsegNews	newstest2016	✓					•	•	•	•	•	•					•							
HUMEseg	himl2015				✓								•	•	•								•	

Table 1: Overview of "tracks" of the WMT16 metrics task. "•" indicates language pairs covered in the evaluation, "·" are language pairs planned but abandoned due to difficulties in obtaining human judgments.

intervals for metric correlations with human assessment, resulting in more reliable conclusions as to which metrics outperform others.

The official method of evaluation remains unchanged, relying on RR in both the system-level (TrueSkill) and segment-level (Kendall's τ) metrics, see below for details and references.

Our datasets are described in Section 2. This includes the test sets, system outputs, human judgments of translation quality as well as participating metrics across the tasks. Results of system-level metric evaluation are provided in Section 3.1 and Section 3.2, the results of the segment-level evaluation are provided in Section 3.3.

2 Data

Table 1 provides the complete picture of the golden truths, test sets, translation systems and language pairs involved in the metrics task this year. For simplicity, we called each of these setups a "track", indicating the underlying type of golden truth (RR/DA/HUME), system- or segment-level evaluation (sys/seg) and the particular test set.

While the set of setups is much larger this year, the participants of the task were affected rather minimally. Participants were only required to run metrics on the additional test sets and with an additional large set of hybrid systems in the system-level evaluation. As in the previous years, participants were allowed take part in any subset of language pairs and setups.

2.1 Test Sets

We use the following test sets:

newstest2016 is the main test set. It is the test set used in WMT16 News Translation Task (Bojar et al., 2016b), with approximately 3,000 sentences for each translation direction (with the exception of Romanian which only has 1,999 sentences). The set includes a single reference translation for each direction, except English→Finnish with two reference translations.

it-test2016 is the set of 1,000 sentences translated from English into seven other European languages. The IT test sentences typically contain instructions for operating commonly used software like web browsers, mail clients or image editors, e.g.: "In message box click on More > Archived."

himl2015 is part of the official test set created by the EU project HimL.[2] These are health-related texts from Cochrane summaries and NHS 24 online content. The texts originated in English and the target languages consist of Czech, German, Polish and Romanian versions created by post-edition of phrase-based MT output. From the full set of about 3,000 sentences, 800 were given as input to the participants of the metrics task and in the end about 340 sentences per language pair were used for evaluation, as those sentences have manual score suitable to employ as the golden truth for metric evaluation.

The sentences of NHS 24 tend to be shorter and simpler translations, e.g. "Choose lower fat options such as semi-skimmed milk and low fat yogurt.", while Cochrane summaries are longer and often contain specific terminology, e.g. "The purpose of this research was to determine how good the TEG and ROTEM assessments are at diagnosing TIC in adult trauma patients who are bleeding."

2.2 Translation Systems

Characteristics of the particular underlying translation task MT systems is likely an important fac-

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

tor affecting the difficulty of the metrics task. 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 could make the task easier for metrics, with metrics that are more sensitive to certain aspects of MT output performing better.

The MT systems included in evaluation of metrics are as follows:

News Task Systems are all MT systems participating in the WMT16 News Translation Task (Bojar et al., 2016b). These systems differ widely in nature (standard phrase-based, syntax-based, transfer-based or even rule-based systems, also with a large number of neural MT systems), with the precise set of systems and system types also depending on specific language pair.

Tuning Task Systems are all Moses phrase-based systems run by the organizers of the WMT16 Tuning Task (Jawaid et al., 2016). All of these systems share the same phrase tables and language models, they are trained on relatively large volumes of data, and differ only in the model weights as provided by the participants of the tuning task. Tuning task was limited to Czech↔English language pairs.

IT Task Systems are participants of the WMT16 IT-domain Translation Task (Bojar et al., 2016b), translating only from English to seven other European languages. This is generally a smaller set of systems and the number of covered system architectures here is also smaller. As far as we know, no neural system was involved in the task.

HimL Year 1 Systems are MT systems released in the first year of the EU project HimL[3]. They are all Moses-based and trained on available data in the medical or health-related domain.

Hybrid Systems were created by combining the output of two newstest2016 translation task systems, with the aim of providing a larger set of systems against which to evaluate metrics, as described further in Section 3.1. In short, we create 10K hybrid MT systems for each language pair.

Excluding the hybrid systems, we ended up with 171 system outputs across 18 language pairs and 3 test sets.

2.3 Manual MT Quality Judgments

There are three distinct "golden truths" employed to evaluate metrics this year: Relative Ranking (RR, as in previous year), Direct Assessment (DA) and HUME, a semantic-based manual metric.

The details of the methods are provided in this section, separately for system-level evaluation (Section 2.3.1, using RR and DA) and segment-level evaluation (Section 2.3.2, using RR, DA and HUME).

The RR manual judgments were provided by MT researchers taking part in WMT tasks, as in recent years of the campaign, after it was empirically established that judgments of RR collected through crowd-sourcing platforms were not reliable (Bojar et al., 2013). DA judgments are more robust in this respect and while the original plan was to collect DA from both researchers and crowd-sourced non-experts, only the latter ultimately took place due to time constraints.

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 document. Both our manual scoring methods RR and DA nevertheless proceed sentence by sentence, aggregating the final score in some way.

Relative Ranking (RR) As in previous WMT shared tasks, human assessors of MT output (only researchers this year) were presented with the source language input, target language reference translation and the output of five distinct MT output translations. Human assessors were required to rank the five translations from best to worse, with ties allowed. As introduced in WMT15, identical translations from distinct systems were collapsed into a single translation before running the human evaluation to increase the overall efficiency of RR human assessment.

Each five-tuple relative ranking was employed to produce 10 pairwise assessments, later combined into a score for each MT system that re-

[3] http://www.himl.eu/

flects the frequency by which the output of that system was preferred to the output of other systems. Several methods have been tested in the past for the exact score calculation and WMT16 has again adopted TrueSkill as the official ranking approach. Please see the WMT16 overview paper for details on how this score is computed.

To increase annotator efficiency, a maximum sentence length of 30 words was applied to RR human assessment.

Direct Assessment (DA) In addition to the standard relative ranking (RR) manual evaluation employed to yield official system rankings in WMT16 translation task, this year the translation task also trialed a new method of human evaluation, monolingual direct assessment (DA) of translation fluency (Graham et al., 2013) and adequacy (Graham et al., 2014; Graham et al., 2016). For investigatory purposes, therefore, we also include evaluation of metrics with reference to the newly trialed human assessment method.

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 for both fluency and adequacy. Furthermore, DA avoids bias that has been problematic in previous evaluations introduced by simultaneous assessment of several alternate translations of a given single source language input, where scores of systems for which translations were often compared to high or low quality translations resulted in an unfair advantage or disadvantage (Bojar et al., 2011). DA achieves this by assessment of individual translations in isolation from other outputs of the same source input.

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. Human 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.[4] Fluency assessment is similar to adequacy except that no reference is displayed and assessors are asked to rate how much they agree that a given translation is fluent target language text.

Large numbers of DA human assessments of translations for seven language pairs (targeting English and Russian) were collected on Amazon's Mechanical Turk,[5] 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 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 or fluency score.

Although the WMT16 Translation Task included both fluency and adequacy DA human assessment, the metrics task this year employed only DA adequacy scores. We hope to incorporate DA fluency into future metric evaluations, however.

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 WMT16.

2.3.2 Segment-level Manual Quality Judgments

Segment-level metrics have been evaluated against the pairwise judgments implied by the 5-way relative ranking annotation. This year, we add two new variants of human assessment: segment-level DA and HUME.

Segment-level DA Adequacy assessments were collected for translations sampled from the output of systems participating in WMT16 translation task for seven language pairs (Graham et al., 2015).[6] Since the actual MT system is not important for segment-level assessment, we sampled 500 translations per language pair at random.

[4] 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 %.

[5] http://www.mturk.com/

[6] Translations produced by ONLINEA were unfortunately omitted from segment-level DA due to submission and data collection timing constraints.

Metric	Participant
BEER	ILLC – University of Amsterdam (Stanojević and Sima'an, 2015)
CHARACTER	RWTH Aachen University (Wang et al., 2016)
CHRF1,2,3, WORDF1,2,3	Humboldt University of Berlin (Popović, 2016)
DEPCHECK	Charles University, no corresponding paper
DPMFCOMB-WITHOUT-RED	Chinese Academy of Sciences and Dublin City University (Yu et al., 2015)
MPEDA	Jiangxi Normal University (Zhang et al., 2016)
UOW.REVAL	University of Wolverhampton (Gupta et al., 2015b)
UPF-COBALT, COBALTF, METRICSF	Universitat Pompeu Fabra (Fomicheva et al., 2016)
DTED	University of St Andrews, (McCaffery and Nederhof, 2016)

Table 2: Participants of WMT16 Metrics Shared Task

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, a human assessment of each translation was collected from 15 distinct human assessors before combination into a 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 perfectly replicable. In repeat experiments, for all tested language pairs, a correlation of above 0.9 between (a) segment-level DA scores for translations collected in an initial experiment run and (b) the same collected in a repeat evaluation of the same translations, by combining assessments of a minimum of 15 human assessors (Graham et al., 2015).

A distinction between DA and RR is that while RR works off a single set of human assessments for evaluation of both system-level and segment-level metrics, DA additionally includes a variant of its methodology designed specifically for evaluation of segment-level metrics.

HUME The HUME metric (Birch et al., 2016) is a novel human evaluation measure that decomposes over the UCCA semantic units. 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 interrelations. HUME operates by aggregating human assessments of the translation quality of individual semantic units in the source sentence. We thus avoid the semantic annotation of machine-generated text, which is often 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 (Graham et al., 2015).

2.4 Participants of the Metrics Shared Task

Table 2 lists the participants of the WMT16 Shared Metrics Task, along with their metrics. We have collected 16 metrics from a total of 9 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.

2.4.1 BEER

BEER (Stanojević and Sima'an, 2015) is a trained evaluation metric with a linear model that combines features capturing character n-grams and permutation trees. BEER has participated in previous years of the evaluation task. This year the learning algorithm is improved (linear SVM instead of logistic regression) and some features that are relatively slow to compute are removed (paraphrasing, syntax and permutation trees) which resulted in a very large speed-up. BEER is usually trained for ranking but in this case there was a compromise: the initial model is trained for ranking (RR) with ranking SVM and then the output from SVM is scaled using trained regression model to approximate absolute judgment (DA).

2.4.2 CHARACTER

CHARACTER (Wang et al., 2016) is a novel 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.

2.4.3 CHRF and WORDF

WORDF1,2,3 (Popović, 2016) calculate a simple F-score combination of the precision and recall of word n-grams of maximal length 4 with different setting for the β parameter ($\beta = 1$, 2, or 3). Precision and recall that are used in computation of the F-score are arithmetic averages of precisions and recalls, respectively, for the different n-gram orders. CHRF1,2,3 calculate the F-score of character n-grams of maximal length 6. β parameter gives β times weight to recall: $\beta = 1$ implies equal weights for precision and recall.

2.4.4 DEPCHECK

DEPCHECK is based on the automatic post-editing tool Depfix (Rosa, 2014). For each sentence, DEPCHECK computes the percentage of nodes post-edited by Depfix, obtaining a "relative depcheck error rate" (RDER). The value of the DEPCHECK metric is then defined as $1 - \text{RDER}$. DEPCHECK does not distinguish the error types or whether there was more than one Depfix rule applied to a node. It is suggested for a future version of DEPCHECK to assign a weight (either by hand, or training from some golden data) to each rule that was applied to the MT output.

2.4.5 DPMFCOMB-WITHOUT-RED

The authors of DPMFCOMB-WITHOUT-RED follow the work on last year's metric DPMFCOMB (Yu et al., 2015), but modify it with two main differences. Firstly, they use the 'case insensitive' instead of 'case sensitive' option when using Asiya. Secondly, REDP are not used. Thus, DPMFCOMB-WITHOUT-RED is a combined metric including 57 single metrics. Weights of the individual metrics are trained with SVMrank, using training data from the English-targeted language pairs from WMT12 to WMT14. In the results DPMFCOMB-WITHOUT-RED is represented as DPMFCOMB for brevity.

2.4.6 DTED

DTED (McCaffery and Nederhof, 2016) is based on Tree Edit Distance. The scoring is done over the dependency parse tree of the output where the number of edit operations (insert, delete or substitute) needed to convert it to the correct (reference) dependency tree is used as an indicator of the translation quality. Unlike the majority of metrics which evaluate many aspects of translation, DTED evaluates only the word order.

2.4.7 MPEDA

MPEDA (Zhang et al., 2016) is developed on the basis of the METEOR metric. In order to accurately match words or phrases with the same or similar meaning, it extracts a domain-specific paraphrase table from the monolingual corpus and applies that paraphrase table to the METEOR metric to replace the general one. Unlike traditional paraphrase extraction approaches, it first filters out a domain-specific sub-corpus from a large general monolingual corpus and then extracts domain-specific paraphrase table from the sub-corpus by Markov Network model. Since the proposed paraphrase extraction approach can be used in all languages, MPEDA is language-independent.

2.4.8 UOW.REVAL

UOW.REVAL (Gupta et al., 2015b) uses dependency-tree Long Short Term Memory (LSTM) network to represent both the hypothesis and the reference with a dense vector. Training is performed using the judgements from WMT13 (Bojar et al., 2013) converted to similarity scores. The final score at the system level is obtained by averaging the segment level scores obtained from a neural network which takes into account both distance and Hadamard product of the two representations.

UOW.REVAL is the same as UOW_LSTM (Gupta et al., 2015a) that participated in the WMT15 task except that LSTM vector dimension is 150 for UoW.ReVal instead of 300.

Track	cs	de	ro	fi	ru	tr	cs	de	ro	fi	ru	tr	bg	es	eu	nl	pl	pt
	into-English						English into											
RRsysNews	T4,F3,T6	T4,F1	T4,F1	T4,F1	T4,F2	T4,F2	T5,F4,T6	T5,F5	T5,F6	T5,F6	T5,F2	T5,F6						
RRsysIT							T8,F4	T8,F5					T8	T8,F7	T8	T8,F7		T8,F7
DAsysNews	T4,F3,T7	T4,F1,T7	T4,F1,T7	T4,F1,T7	T4,F2,T7	T4,F2,T7					T5,F2,T7							
RRsegNews	T9	T9	T9	T9	T9	T9	T10	T10	T10	T10	T10	T10						
DAsegNews	T9,F8	T9,F8	T9,F8	T9,F8	T9,F8	T9,F8					T10,F9							
HUMEseg							T11,F10	T11,F10	T11,F10							T11,F10		

Table 3: Overview of tables (T) and figures (F) reporting results of the individual "tracks" and language pairs.

2.4.9 UPF-COBALT, COBALTF and METRICSF

UPF-COBALT (Fomicheva et al., 2016) is an alignment-based metric that examines the syntactic contexts of lexically similar candidate and reference words in order to distinguish meaning-preserving variations from the differences indicative of MT errors. This year the metric was improved by explicitly addressing MT fluency. The new version of the metric, COBALTF, combines various components of UPF-COBALT with a number of fine-grained features intended to capture the number and scale of disfluent fragments contained in MT sentences. METRICSF is a combination of three evaluation systems, BLEU, METEOR and UPF-COBALT, with the fluency-oriented features.

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 MTEVALBLEU (Papineni et al., 2002) and MTEVALNIST (Doddington, 2002) were computed using the script `mteval-v13a.pl`[7] which 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 MOSESBLEU, MOSESTER (Snover et al., 2006), MOSESWER, MOSESPER and MOSECDER (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, part of the Moses toolkit. It is a smoothed version of BLEU that correlates better with human judgments for segment-level.

For computing system-level scores, the same script was employed as in last year's metric task. New scripts have been added for system-level hybrids and segment-level evaluation.

3 Results

Table 3 provides an overview of all the tables and figures in the rest of the paper. We discuss system-level results for news task systems (including tuning task systems) in Section 3.1. The system-level results for the IT domain are discussed in Section 3.2. The segment-level results are in Section 3.3. We end with discussion in Section 3.4.

3.1 System-Level Results for News Task

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

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

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

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

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.

Table 4 includes results for system-level into-English metrics for evaluation of systems participating in the main translation task (newstest2016), evaluated against RR and DA human assessment variants, while Table 5 includes the same for the newstest2016 out-of-English language pairs (only Russian has the DA judgments). Tuning systems were excluded from Tables 4 and 5 and they are covered by Table 6 that shows correlations achieved by metrics with RR when the set of systems additionally includes tuning task systems.

In previous years, we reported empirical confidence intervals of system-level correlations obtained by bootstrap resampling human assessments data and computing confidence intervals for individual correlations with human assessment. Such confidence intervals reflect the variance due to particular sentences and assessors involved in the evaluation but lead to over-estimation of significant differences if employed to conclude which metrics outperform others. This year, as recommended by Graham and Baldwin (2014), instead we employ Williams significance test (Williams, 1959). 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 are highlighted in bold in Tables 4 and 5. Since RR is the official method of evaluation for this year's metrics task, bolded correlations under RR comprise official winners of the news domain portion of the system-level metrics task. DA results are included for comparison and are investigatory only.

With regard to which individual metric may or may not outperform other metrics, such as the important comparison as to which metrics significantly outperform the most widely employed metric BLEU (in its mteval or Moses scorer implementation), Figures 1, 2, 3, 4, 5, and 6 include significance test results for every competing pair of metrics including our baseline metrics. In heatmaps in Figures 1, 2, 3, 4, 5, and 6, the column labelled "MTEVALBLEU" or "MOSESBLEU" can be used to quickly observe which metrics achieve a significant increase in correlation with human assessment over that of BLEU, where a green cell in the column denotes outperformance of BLEU by the metric in that row.

	cs-en	**en-cs**
Human	RR + TT	RR + TT
Systems	12	20
WORDF2	.988	**.990**
WORDF1	.989	**.990**
MOSESBLEU	.989	**.987**
WORDF3	.988	**.989**
MTEVALBLEU	.985	**.986**
MOSESCDER	**.991**	.976
BEER	**.995**	.972
MPEDA	.988	.977
CHRF1	.990	.965
MTEVALNIST	.976	**.979**
CHRF2	.990	.952
CHRF3	.989	.935
CHARACTER	**.997**	.779
MOSESPER	.970	.803
MOSESTER	.974	.758
MOSESWER	.964	.755
UOW.REVAL	.982	-
newstest2016		

Table 6: Absolute Pearson correlation of cs-en and en-cs system-level metric scores with human assessment variant RR + TT, i.e. standard WMT relative ranking including tuning task systems.

For investigatory purposes only, we also include hybrid-supersample (Graham and Liu, 2016) results for system-level metrics. 10K hybrid systems were created per language pair, with corresponding DA human assessment scores, by sampling pairs of systems from WMT16 translation task and creating a hybrid system by combining translations from each system to create new hybrid output test set documents, each with a corresponding DA human assessment score. Not all metrics participating in the system-level metrics shared task submitted metric scores for the large set of hybrid systems, possibly due to the increased time required to run metrics on the large set of 10K systems. In this respect, DA hybrid may provide some indication of which metrics are likely to be more feasible to employ for tuning purposes in MT systems out-of-the-box. Due to time constraints, this year it was only possible to include hybrid-supersampling results for language pairs evaluated by the DA human assessment variant.

Correlations of metric scores with human assessment of the large set of hybrid systems are

	cs-en		de-en		fi-en		ro-en		ru-en		tr-en	
Human	RR	DA	RR	DA	RR	DA	RR	DA	RR	DA	RR	DA
Systems	6	6	10	10	9	9	7	7	10	10	8	8
MPEDA	**.996**	**.993**	**.956**	.937	**.967**	**.976**	**.938**	**.932**	**.986**	.929	**.972**	**.982**
UOW.REVAL	**.993**	**.986**	**.949**	**.985**	**.958**	**.970**	**.919**	**.957**	**.990**	**.976**	**.977**	**.958**
BEER	**.996**	**.990**	.949	.879	**.964**	**.972**	**.908**	.852	**.986**	.901	**.981**	**.982**
CHRF1	.993	.986	.934	.868	**.974**	**.980**	.903	.865	**.984**	.898	**.973**	.961
CHRF2	**.992**	.989	.952	.893	**.957**	**.967**	.913	.886	**.985**	.918	.937	.933
CHRF3	.991	.989	**.958**	.902	.946	.958	.915	**.892**	**.981**	.923	.918	.917
CHARACTER	**.997**	**.995**	**.985**	.929	.921	.927	**.970**	**.883**	.955	**.930**	.799	.827
MTEVALNIST	.988	.978	.887	.801	.924	.929	.834	.807	.966	.854	**.952**	**.938**
MTEVALBLEU	**.992**	**.989**	.905	.808	.858	.864	.899	.840	.962	.837	.899	.895
MOSESCDER	**.995**	**.988**	.927	.827	.846	.860	.925	.800	.968	.855	.836	.826
MOSESTER	**.983**	**.969**	.926	.834	.852	.846	.900	.793	.962	.847	.805	.788
WORDF2	**.991**	**.985**	.897	.786	.790	.806	.905	.815	.955	.831	.807	.787
WORDF3	**.991**	**.985**	.898	.787	.786	.803	.909	.818	.955	.833	.803	.786
WORDF1	**.992**	**.984**	.894	.780	.796	.808	.890	.804	.954	.825	.806	.776
MOSESPER	.981	.970	.843	.730	.770	.767	.791	.748	**.974**	.887	**.947**	.940
MOSESBLEU	**.991**	**.983**	.880	.757	.752	.759	.878	.793	.950	.817	.765	.739
MOSESWER	**.982**	**.967**	.926	.822	.773	.768	.895	.762	.958	.837	.680	.651
						newstest2016						

Table 4: Absolute Pearson correlation of to-English system-level metric scores with human assessment variants: RR = standard WMT relative ranking; DA = direct assessment of translation adequacy.

	en-cs		en-de		en-fi		en-ro		en-ru		en-tr	
Human	RR	DA	RR	DA	RR	DA	RR	DA	RR	DA	RR	DA
Systems	10		15		13		12		12	12	8	
CHARACTER	**.947**	-	**.915**	-	**.933**	-	**.959**	-	**.954**	**.966**	**.930**	-
BEER	**.973**	-	.732	-	.940	-	**.947**	-	.906	.922	**.956**	-
CHRF2	.954	-	.725	-	**.974**	-	.828	-	.930	**.955**	**.940**	-
CHRF3	.954	-	.745	-	**.974**	-	.818	-	**.936**	**.960**	**.916**	-
MOSESCDER	**.968**	-	.779	-	.910	-	**.952**	-	.874	.874	.791	-
CHRF1	.955	-	.645	-	.931	-	.858	-	.901	.928	**.938**	-
WORDF3	**.964**	-	.768	-	.901	-	**.931**	-	.836	.840	.714	-
WORDF2	**.964**	-	.766	-	.899	-	**.933**	-	.836	.840	.715	-
WORDF1	**.964**	-	.756	-	.888	-	**.937**	-	.836	.839	.711	-
MPEDA	**.964**	-	.684	-	**.944**	-	.786	-	.856	.866	.860	-
MOSESBLEU	**.968**	-	.784	-	.857	-	**.944**	-	.820	.820	.693	-
MTEVALBLEU	**.968**	-	.752	-	.868	-	.897	-	.835	.838	.745	-
MTEVALNIST	**.975**	-	.625	-	.886	-	.882	-	.890	.897	.788	-
MOSESTER	.940	-	.742	-	.863	-	.906	-	.882	.879	.644	-
MOSESWER	.935	-	.771	-	.855	-	**.912**	-	.882	.876	.570	-
MOSESPER	**.974**	-	.681	-	.700	-	**.944**	-	.857	.854	.641	-
CHRF3.2REF	-	-	-	-	**.973**	-	-	-	-	-	-	-
CHRF2.2REF	-	-	-	-	**.970**	-	-	-	-	-	-	-
CHRF1.2REF	-	-	-	-	.923	-	-	-	-	-	-	-
WORDF3.2REF	-	-	-	-	.890	-	-	-	-	-	-	-
WORDF2.2REF	-	-	-	-	.887	-	-	-	-	-	-	-
WORDF1.2REF	-	-	-	-	.876	-	-	-	-	-	-	-
						newstest2016						

Table 5: Absolute Pearson correlation of out-of-English system-level metric scores with human assessment variants: RR = standard WMT relative ranking; DA = direct assessment of translation adequacy.

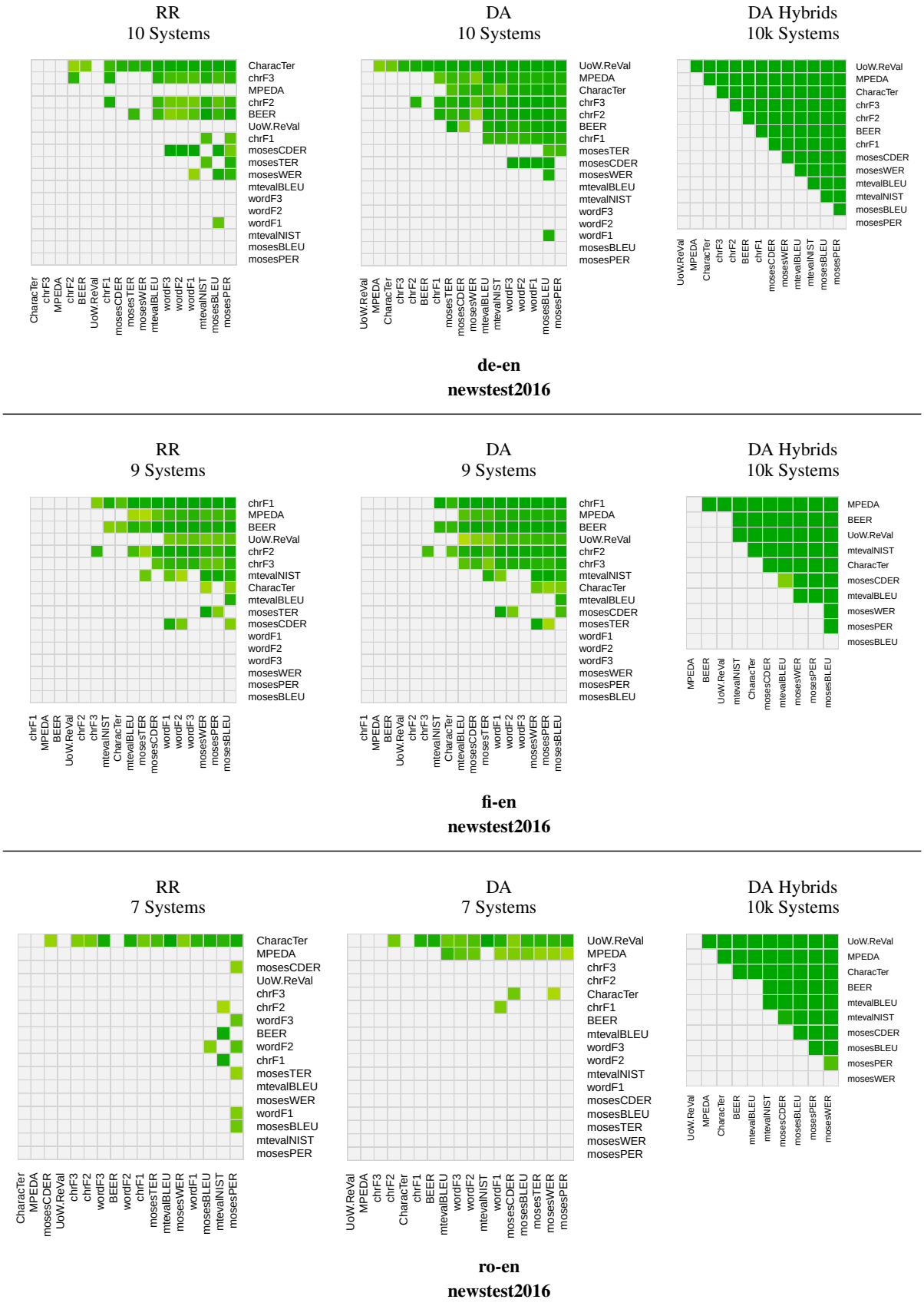


Figure 1: German-to-English (de-en), Finnish-to-English (fi-en) and Romanian-to-English (ro-en) system-level metric significance test results for human assessment variants; green cells denote a significant increase in correlation with human assessment for the metric in a given row over the metric in a given column according to Williams test; RR = standard WMT relative ranking for translation task systems only; DA = direct assessment of translation adequacy; DA Hybrids = direct assessment with hybrid super-sampling.

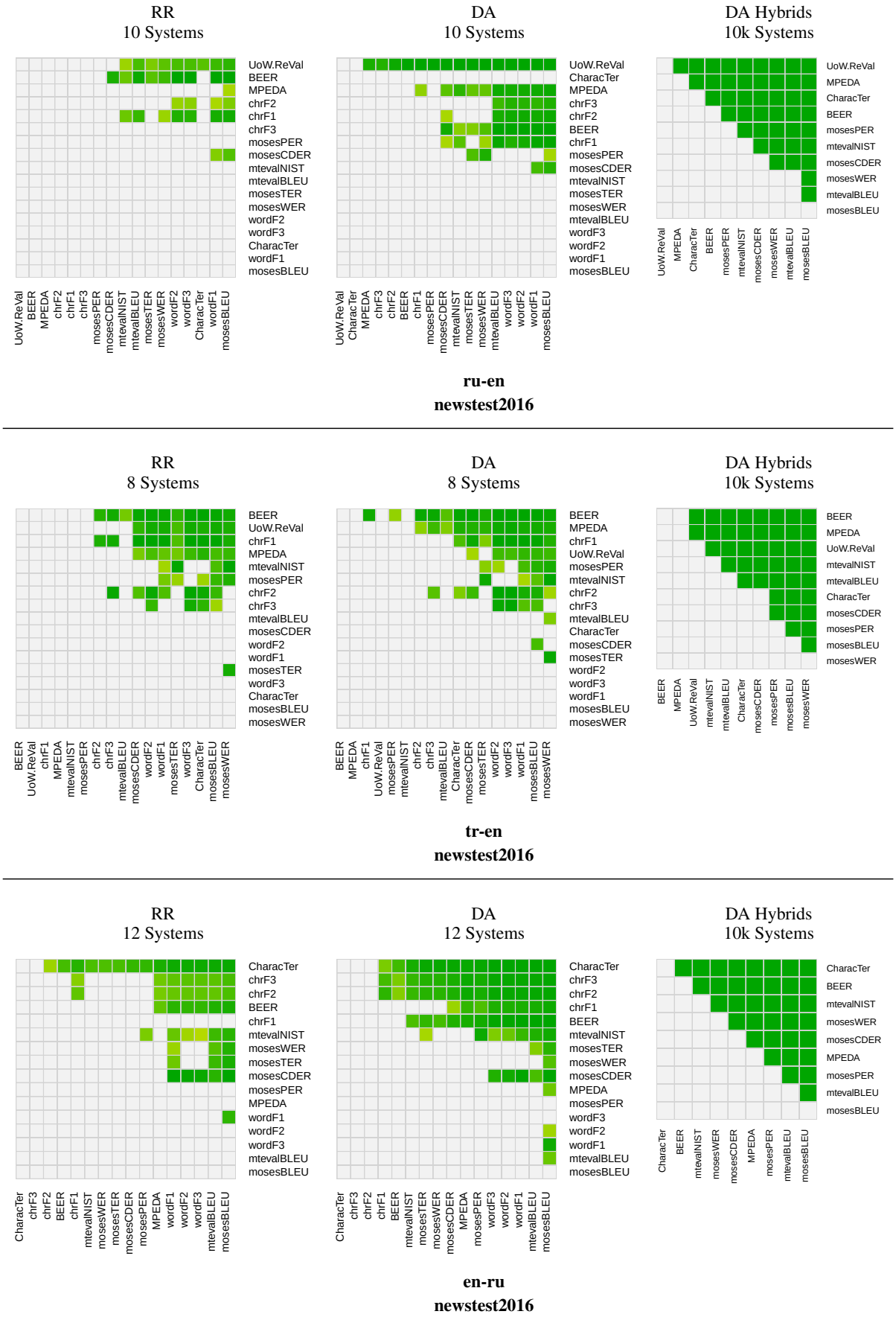


Figure 2: Russian-to-English (ru-en), Turkish-to-English (tr-en) and English-to-Russian (en-ru) system-level metric significance test results for human assessment variants; green cells denote a significant increase in correlation for the metric in a given row over the metric in a given column according to Williams test; RR = standard WMT relative ranking for translation task systems only; DA = direct assessment of translation adequacy; DA Hybrids = direct assessment with hybrid super-sampling.

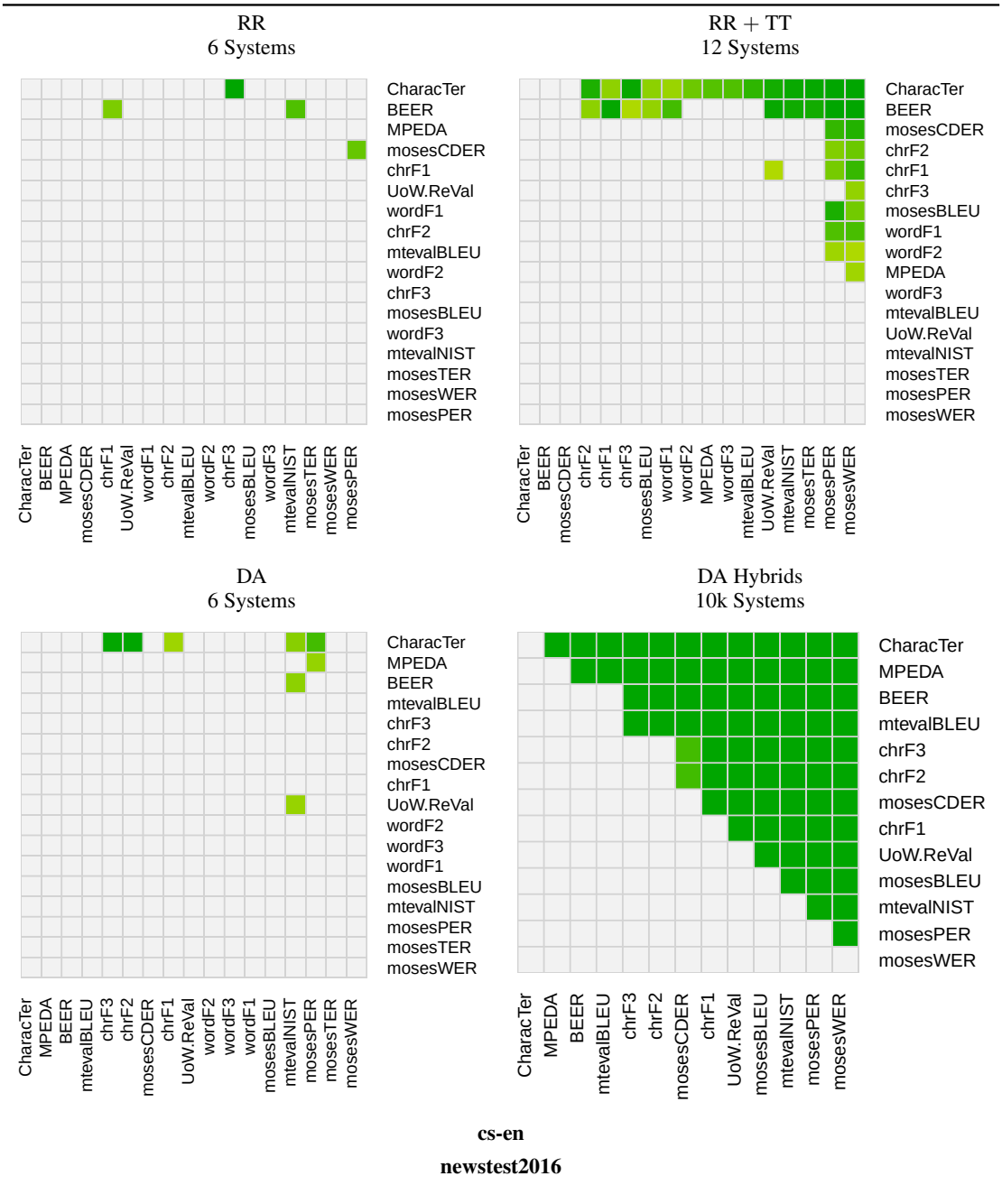


Figure 3: Czech-to-English (cs-en) system-level metric significance test results for human assessment variants; a green cell corresponds to a significant increase in correlation for the metric in a given row over the metric in a given column according to Williams test; RR = standard WMT relative ranking for translation task systems only; RR + TT = standard WMT relative ranking for all cs-en newstest2016 systems; DA = direct assessment of translation adequacy; DA Hybrids = direct assessment with hybrid super-sampling.

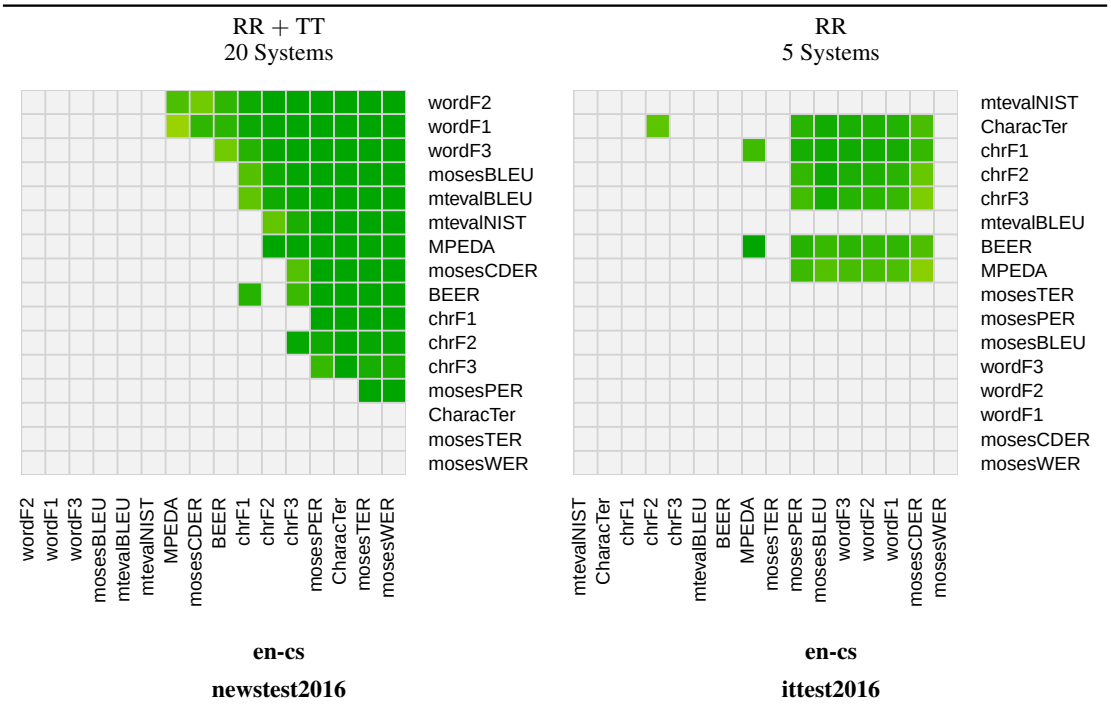


Figure 4: English-to-Czech (en-cs) system-level metric significance test results; a green cell corresponds to a significant increase in correlation for the metric in a given row over the metric in a given column according to Williams test; RR = standard WMT relative ranking; RR + TT = standard WMT relative ranking for translation and tuning task systems.

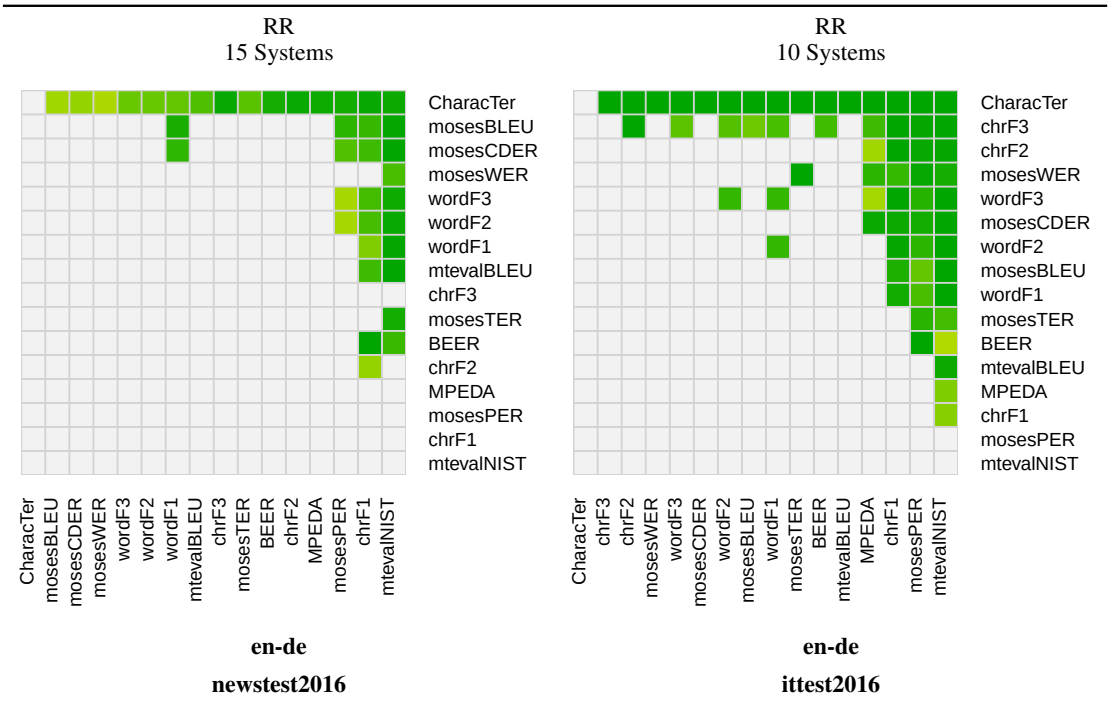


Figure 5: English-to-German (en-de) system-level metric significance test results; a green cell corresponds to a significant increase in correlation for the metric in a given row over the metric in a given column according to Williams test; RR = standard WMT relative ranking.

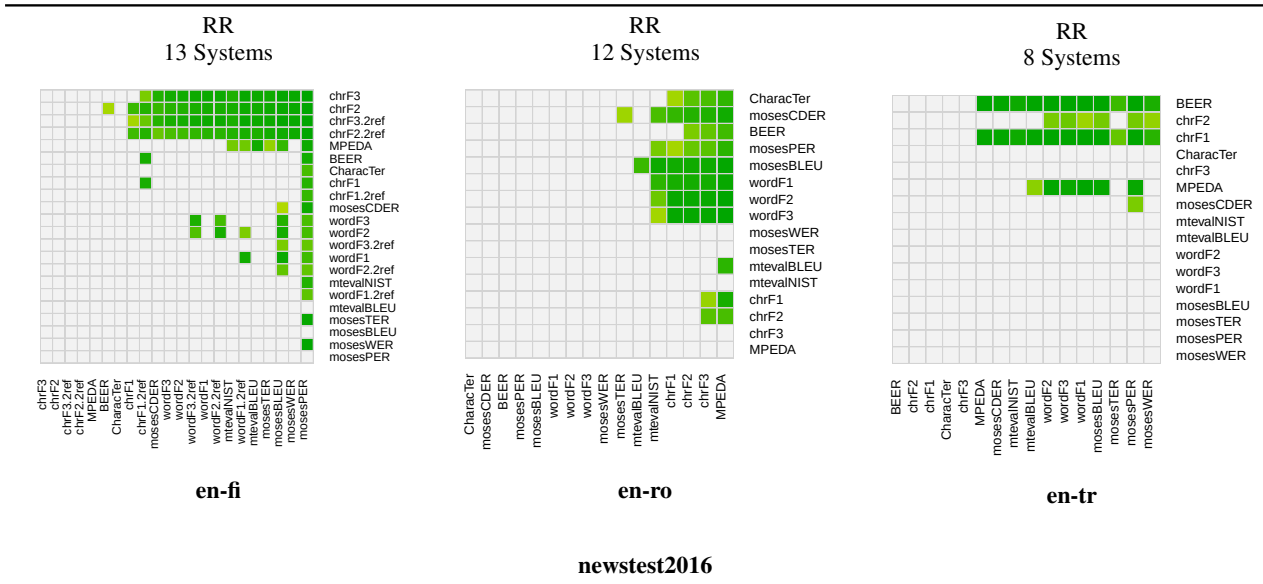


Figure 6: English-to-Finnish (en-fi), English-to-Romanian (en-ro) and English-to-Turkish (en-tr) system-level metric significance test results; a green cell corresponds to a significant increase in correlation for the metric in a given row over the metric in a given column according to Williams test; RR = standard WMT relative ranking.

shown in Table 7, where again metrics not significantly outperformed by any other are highlighted in bold. Results are for investigatory purposes only and do not indicate official winners, however. Figures 1, 2 and 3 also include significance test results for hybrid super-sampled correlations for all pairs of competing metrics for a given language pair.

In Appendix A, correlation plots for each language pair are also provided. The left-hand plot visualizes the correlation of MTEVALBLEU and manual judgements, while the right-hand plot shows the correlation for the best performing metrics for that pair according to both standard RR and DA, as per Tables 4, 5 and 7.

3.2 System-Level Results for IT Task

Since systems participating in the IT domain translation task were manually evaluated with RR, we include evaluation of metrics for translation of this specific domain. Results of all metrics evaluated on the IT domain MT systems are shown in Table 8, where official winning metrics for this domain are identified as those not significantly outperformed by any other metric according to Williams test, correlations for which are highlighted in bold.[8]

Full pairwise significance test results for every pair of competing metrics evaluated on IT domain systems for Spanish, Dutch and Portuguese are shown in Figure 7, German in Figure 5 and Czech in Figure 4. No significance tests are provided for IT domain Bulgarian and Basque, as all metrics achieved equal correlations.

We see from Table 8 and also Figure 7 that MOSESBLEU does not belong to the winners for several target languages (Czech, German, Dutch), but across the board, metrics are hard to distinguish on this specific test set.

3.3 Segment-Level Results

In WMT16, the official method for segment-level metric evaluation remains unchanged: a Kendall's Tau-like formulation of a given metric's agreement with pairwise human assessment of translations, collected through 5-way relative ranking (RR). However, we also trial evaluation of segment-level metrics with reference to segment-level DA human assessment (for the main translation task data set) and a semantic-based manual judgments HUME (for himl2015 data set).

[8] Bulgarian and Basque IT translation tasks included only two participating systems and all metrics were able to order them correctly, all resulting in a correlation of 1.0.

	cs-en	de-en	fi-en	ro-en	ru-en	tr-en	en-ru
Human	DA	DA	DA	DA	DA	DA	DA
Systems	10K	10K	10K	10K	10K	10K	10K
MPEDA	.988	.923	**.971**	.905	.923	**.975**	.860
BEER	.985	.871	.964	.828	.894	**.975**	.914
CHARACTER	**.989**	.918	.915	.850	.919	.822	**.954**
MTEVALNIST	.971	.790	.919	.784	.853	.919	.890
MTEVALBLEU	.985	.802	.849	.828	.833	.868	.831
MOSESCDER	.984	.819	.851	.777	.850	.822	.868
UOW.REVAL	.981	**.976**	.964	**.930**	**.967**	.951	-
MOSESPER	.970	.728	.758	.745	.877	.798	.846
MOSESWER	.962	.814	.758	.741	.834	.642	.870
MOSESBLEU	.979	.753	.747	.772	.819	.708	.813
CHRF3	.984	.892	-	-	-	-	-
CHRF2	.984	.882	-	-	-	-	-
CHRF1	.982	.856	-	-	-	-	-
			newstest2016				

Table 7: Absolute Pearson correlation of system-level metric scores with 10K hybrid systems: DA Hybrid = direct assessment of translation adequacy of 10K hybrid MT systems.

	en-bg	en-cs	en-de	en-es	en-eu	en-nl	en-pt
Human	RR	RR	RR	RR	RR	RR	RR
Systems	2	5	10	4	2	4	4
CHARACTER	**1.000**	**0.901**	**0.930**	**0.963**	**1.000**	**0.927**	0.976
CHRF3	**1.000**	**0.831**	0.700	**0.938**	**1.000**	**0.961**	0.990
CHRF2	**1.000**	0.837	0.672	0.933	**1.000**	**0.959**	0.986
BEER	**1.000**	**0.744**	0.621	0.931	**1.000**	**0.983**	**0.989**
CHRF1	**1.000**	**0.845**	0.588	0.915	**1.000**	**0.951**	0.967
MTEVALNIST	**1.000**	**0.905**	0.524	0.926	**1.000**	0.722	**0.993**
MPEDA	**1.000**	0.620	0.599	**0.951**	**1.000**	0.856	**0.989**
MOSESTER	**1.000**	**0.616**	0.628	**0.908**	**1.000**	0.835	**0.994**
MTEVALBLEU	**1.000**	**0.750**	0.621	**0.976**	**1.000**	0.596	**0.997**
MOSESWER	**1.000**	**0.009**	0.656	0.916	**1.000**	**0.903**	0.991
MOSESCDER	**1.000**	0.181	0.652	**0.932**	**1.000**	**0.914**	**0.997**
WORDF1	**1.000**	0.240	0.644	**0.959**	**1.000**	**0.911**	**0.997**
WORDF2	**1.000**	0.266	0.652	**0.965**	**1.000**	**0.900**	**0.997**
WORDF3	**1.000**	0.274	0.655	**0.966**	**1.000**	**0.897**	**0.996**
MOSESBLEU	**1.000**	0.296	0.650	**0.974**	**1.000**	0.886	**0.992**
MOSESPER	**1.000**	0.307	0.548	0.911	**1.000**	**0.938**	**0.998**
			ittest2016				

Table 8: System-level metric results (ittest2016): Pearson correlation of system-level metric scores with human assessment computed over standard WMT relative ranking (RR) human assessments; absolute values of correlation coefficients reported for all metrics.

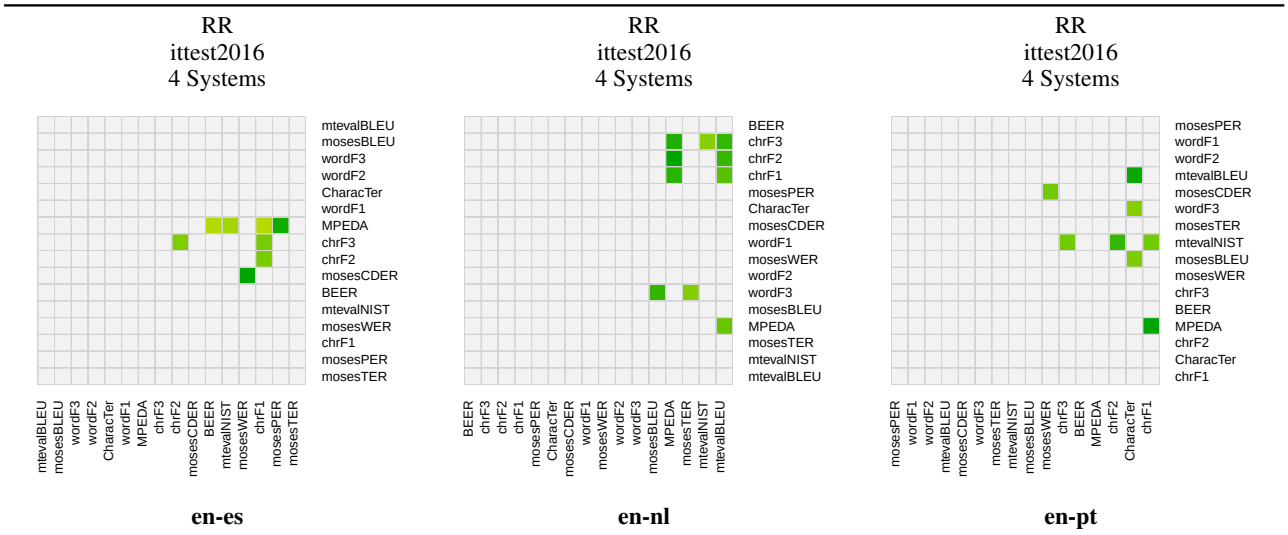


Figure 7: System-level metric ittest2016 significance test results for differences in metric correlation with human assessment for remaining out-of-English language pairs evaluated with relative ranking (RR) human assessment.

Segment-level DA Evaluation Segment-level DA adequacy scores, as described in Section 2.3.2, are employed as gold standard human scores for translations. Since DA segment-level scores are absolute judgments, in their raw (non-standardized) form corresponding simply to a percentage of the absolute adequacy of a given translation, evaluation of metrics simply takes the form of the computation of a Pearson correlation coefficient between metric and DA scores for translations. Significance of differences in metric performance, as in system-level DA metric evaluation, takes the form of Williams test for the significance of a difference in dependent correlations (Williams, 1959; Graham et al., 2015).

Segment-level HUME evaluation The evaluation of segment-level metrics with reference to HUME scores operates in a similar way to DA, by computing the Pearson correlation of HUME evaluation scores for individual translations with metric scores. Williams test is also applied to test for significant differences in metric performance.

Kendall's Tau-like Formulation We measure the quality of metrics' segment-level scores 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 RR 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, for segment-level evaluation is as follows:

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

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 judgment) 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:

Given such a matrix $C_{h,m}$ where $h, m \in \{<, =, >\}$[9] and a metric, we compute the Kendall's τ for the metric the following way:

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

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

$$\tau = \frac{\sum_{\substack{h,m \in \{<,=,>\} \\ C_{h,m} \neq \mathrm{X}}} C_{h,m} |S_{h,m}|}{\sum_{\substack{h,m \in \{<,=,>\} \\ C_{h,m} \neq \mathrm{X}}} |S_{h,m}|} \quad (3)$$

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 $C_{h,m} = \mathrm{X}$.

To summarize, the WMT16 matrix specifies to:

- exclude all human ties,
- count metric's ties only for the denominator (thus giving no credit for giving a tie),
- all cases of disagreement between human and metric judgments are counted as $Discordant$,
- all cases of agreement between human and metric judgments are counted as $Concordant$.

In previous years, we reported confidence intervals for the Kendall's Tau formulation, see Bojar et al. (2015) for details. However, since the formulation of Kendall's Tau is not computed in the standard way (we do not have a single overall ranking of translations, but rather rankings of sets of 5 translations), the accuracy of confidence intervals computed in this way is difficult to verify. To avoid the risk of drawing incorrect conclusions of significant differences in metric performance, we do not include confidence intervals with this year's Kendall's Tau formulation results.

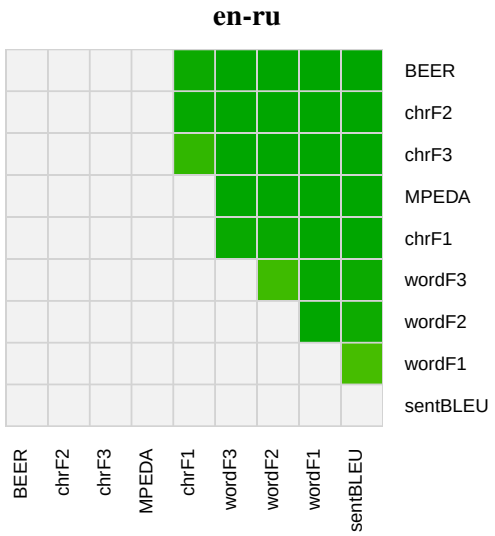


Figure 9: Direct Assessment (DA) segment-level metric significance test results for English to Russian (newstest2016): Green cells denote a significant win for the metric in a given row over the metric in a given column according to Williams test for difference in dependent correlation.

Results of the segment-level human evaluation for translations sampled from the main translation task are shown in Tables 9 and 10, where metric correlations (for DA human assessment variant only) not significantly outperformed by any other metric are highlighted in bold. Since Kendall's Tau are traditionally employed to conclude task winners, while at the same time we currently lack a known reliable method of identifying significant differences between metrics, we postpone announcement of official winning segment-level metrics until further research has been carried out to establish a reliable method in this respect.

DA human assessment pairwise significance test results for differences in metric performance are included for investigatory purposes only in Figures 8 and 9.

Results of segment-level metrics task evaluated with HUME on the himl2015 data set are shown in Table 11, where metrics not significantly outperformed by any other in a given language pair are highlighted in bold, and these metrics are official winners of the himl2015 segment-level metric evaluation. Full pairwise significance test results for all metrics are shown in Figure 10.

Direction	**cs-en**		**de-en**		**fi-en**		**ro-en**		**ru-en**		**tr-en**	
Human Gold	RR	DA	RR	DA	RR	DA	RR	DA	RR	DA	RR	DA
# Assessments	70k	12k	15k	12k	19k	14k	11k	12k	18k	13k	7k	13k
# Translations	8.6k	560	2.4k	560	4.6k	560	2.2k	560	4.7k	560	2.2k	560
Correlation	τ	r	τ	r	τ	r	τ	r	τ	r	τ	r
DPMFCOMB	.388	**.713**	.420	**.584**	.481	**.598**	.383	.627	.420	**.615**	.401	**.663**
METRICS-F	.345	**.696**	.421	**.601**	.447	.557	.388	**.662**	.412	**.618**	.424	**.649**
COBALT-F.	.336	.671	.415	**.591**	.433	.554	.361	.639	.397	**.618**	.423	.627
UPF-COBA.	.359	.652	.387	.550	.436	.490	.356	.616	.394	.556	.379	.626
BEER	.342	.661	.371	.462	.416	.471	.331	.551	.376	.533	.372	.545
MPEDA	.331	.644	.375	.538	.425	.513	.339	.587	.387	.545	.335	.616
CHRF2	.341	.658	.358	.457	.418	.469	.344	.581	.383	.534	.346	.556
CHRF3	.343	.660	.351	.455	.421	.472	.341	.582	.382	.535	.345	.555
CHRF1	.323	.644	.372	.454	.410	.452	.339	.570	.379	.522	.345	.551
UOW-REVAL	.261	.577	.329	.528	.376	.471	.313	.547	.314	.528	.342	.531
WORDF3	.299	.599	.293	.447	.377	.473	.304	.525	.343	.504	.287	.536
WORDF2	.297	.596	.296	.445	.378	.471	.300	.522	.341	.503	.283	.537
WORDF1	.290	.585	.293	.435	.369	.464	.293	.508	.336	.497	.275	.535
SENTBLEU	.284	.557	.265	.448	.368	.484	.272	.499	.330	.502	.245	.532
DTED	.201	.394	.130	.254	.209	.361	.144	.329	.201	.375	.142	.267

newstest-2016

Table 9: Segment-level metric results for to-English language pairs (newstest2016): Correlation of segment-level metric scores with human assessment variants, where τ are official results computed similar to Kendall's τ and over standard WMT relative ranking (RR) human assessments; r are Pearson correlation coefficients of metric scores with direct assessment (DA) of absolute translation adequacy; absolute value of correlation coefficients reported for all metrics.

Direction	**en-cs**		**en-de**		**en-fi**		**en-ro**		**en-ru**		**en-tr**	
Human Gold	RR	DA	RR	DA	RR	DA	RR	DA	RR	DA	RR	DA
# Assessments	118k	-	35k	-	31k	-	7k	-	21k	20k	7k	-
# Translations	12.9k	-	6.2k	-	4.1k	-	1.9k	-	6.0k	-	3.0k	-
Correlation	τ	r	τ	r	τ	r	τ	r	τ	r	τ	r
BEER	.422	-	.333	-	.364	-	.307	-	.405	**.666**	.337	-
CHRF2	.420	-	.329	-	.374	-	.304	-	.406	**.661**	.330	-
CHRF3	.421	-	.327	-	.380	-	.304	-	.400	**.661**	.326	-
CHRF1	.402	-	.320	-	.350	-	.305	-	.389	.642	.320	-
MPEDA	.393	-	.274	-	.342	-	.238	-	.372	**.645**	.255	-
WORDF2	.373	-	.247	-	.313	-	.250	-	.358	.580	.218	-
WORDF3	.373	-	.247	-	.314	-	.245	-	.359	.582	.216	-
WORDF1	.369	-	.245	-	.311	-	.248	-	.351	.573	.209	-
SENTBLEU	.359	-	.236	-	.306	-	.233	-	.328	.550	.222	-
CHRF3-2R.	-	-	.334	-	-	-	-	-	-	-	-	-
CHRF2-2R.	-	-	.331	-	-	-	-	-	-	-	-	-
CHRF1-2R.	-	-	.324	-	-	-	-	-	-	-	-	-
WORDF3-2.	-	-	.251	-	-	-	-	-	-	-	-	-
WORDF2-2.	-	-	.251	-	-	-	-	-	-	-	-	-
WORDF1-2.	-	-	.250	-	-	-	-	-	-	-	-	-
DEPCHECK	.109	-	-	-	-	-	-	-	-	-	-	-

newstest-2016

Table 10: Segment-level metric results for out-of-English language pairs (newstest2016): Absolute correlation of segment-level metric scores with human assessment variants, where τ are official results computed similar to Kendall's τ and over standard WMT relative ranking (RR) human assessments; r are Pearson correlation coefficients of metric scores with direct assessment (DA) of absolute translation adequacy; absolute value of correlation coefficients reported for all metrics.

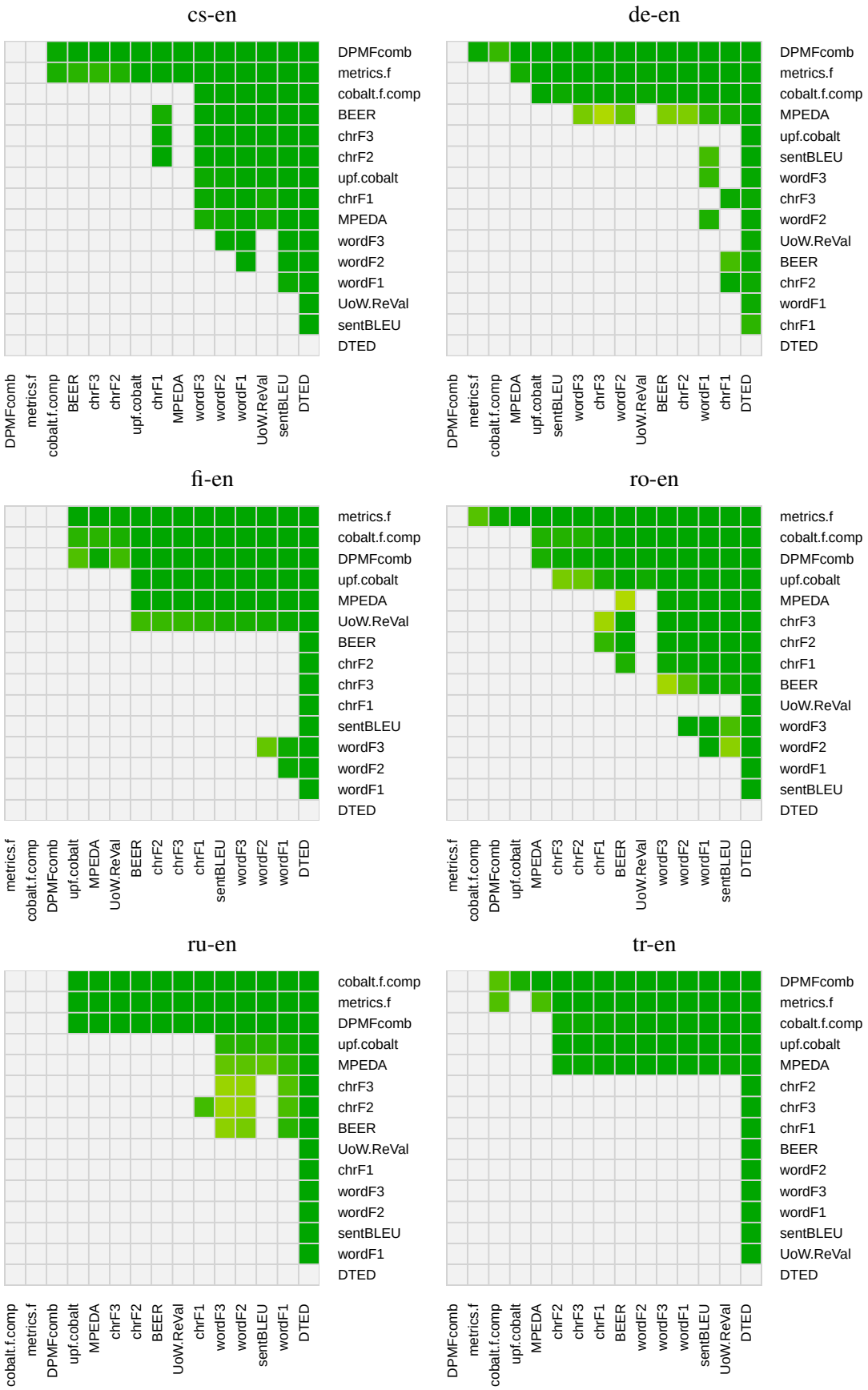


Figure 8: Direct Assessment (DA) segment-level metric significance test results for to-English language pairs (newstest2016): Green cells denote a significant win for the metric in a given row over the metric in a given column according to Williams test for difference in dependent correlation.

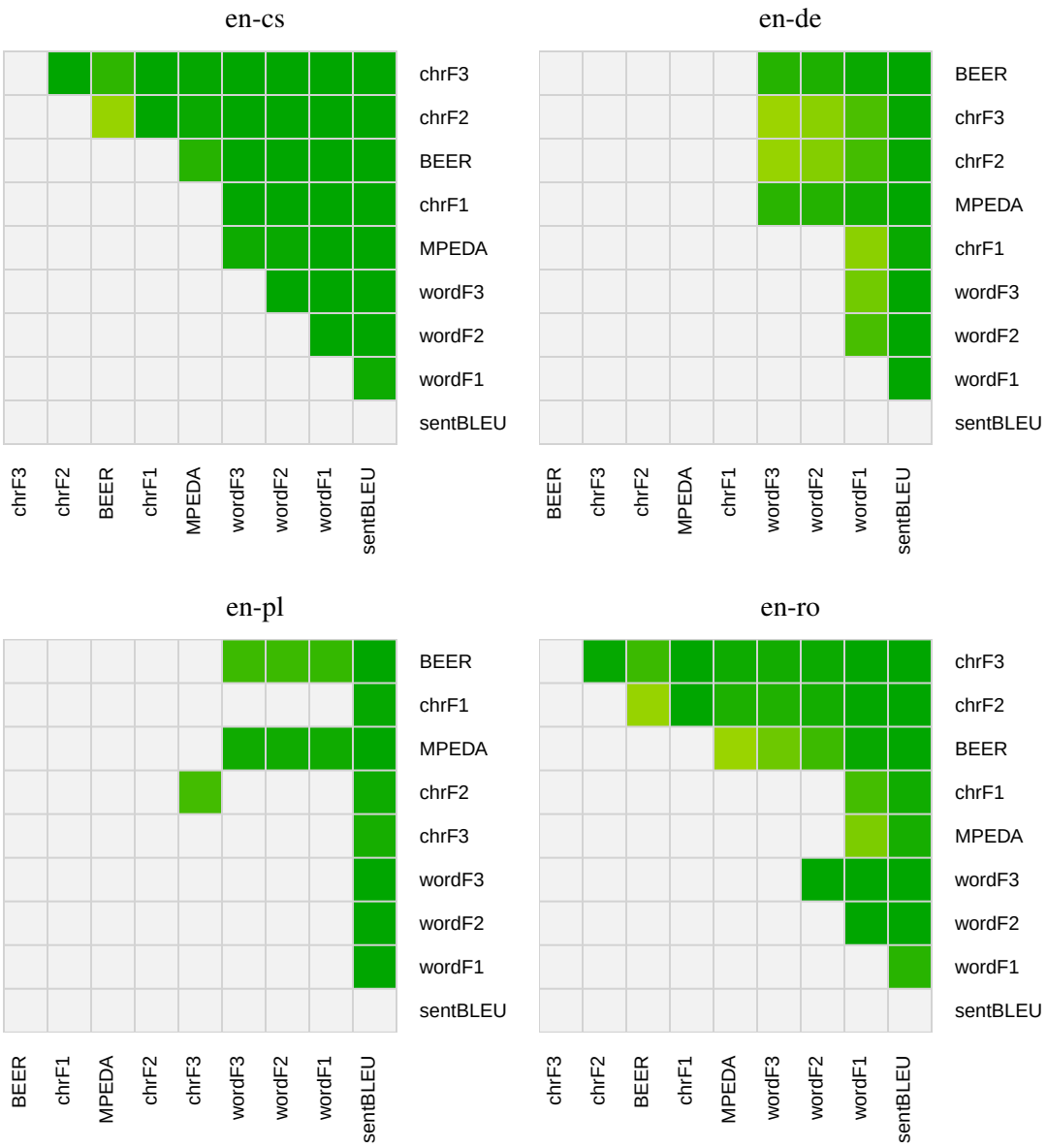


Figure 10: HUME segment-level metric significance test results (himl2015): Green cells denote a significant win for the metric in a given row over the metric in a given column according to Williams test for difference in dependent correlation; Winning metrics are those not significantly outperformed by any other (en-cs: CHRF3; en-de: BEER, CHRF3, CHRF2, MPEDA, CHRF1; en-pl: BEER, CHRF1, MPEDA, CHRF2; en-ro: CHRF3).

Direction	en-cs	en-de	en-ro	en-pl
Human Gold	HUME	HUME	HUME	HUME
n	339	330	349	345
Correlation	r	r	r	r
CHRF3	**.544**	**.480**	**.639**	.413
CHRF2	.537	**.479**	.634	**.417**
BEER	.516	**.480**	.620	**.435**
CHRF1	.506	**.467**	.611	**.427**
MPEDA	.468	**.478**	.595	**.425**
WORDF3	.413	.425	.587	.383
WORDF2	.408	.424	.583	.383
WORDF1	.392	.415	.569	.381
SENTBLEU	.349	.377	.550	.328
		himl-2015		

Table 11: Pearson correlation of segment-level metric scores with HUME human assessment variant.

3.4 Discussion

During the task, the DA evaluation, other than being more principled and discerning, has proved more reliable for crowd-sourcing human evaluation of MT.

It should be noted that DA requires distinct DA human evaluation variants for system and segment-level evaluation, but we may not see this as a negative but rather that DA provides a new method of human evaluation devised specifically for accurate evaluation of segment-level metrics.

Although this year DA was carried out through crowd-sourcing, while RR was completed by researchers, DA is not restricted to crowd-sourcing and could be carried out as-is by researchers or by slight modification by removal of the overhead of translation assessments included in DA for quality control. With any method of human evaluation, if we aim at crowd-sourcing, we must keep in mind that some languages are difficult to obtain workers for, observed in the fact that this year's WMT only collected crowd-sourced assessment for English and Russian as a target language. Although we employed a minimum of 15 human assessors for segment-level evaluation of metrics per segment, it might be worth noting that preliminary empirical evaluation has shown that the 15 human assessments we acquire do not need to be from distinct workers and when repeat assessments are allowed from the same worker, this also yields a correlation of above 0.9 with assessments of translations collected from strictly distinct workers. In other words, DA should be technically viable for all language pairs, if we employ researchers as opposed to crowd-sourced assessors (who may not be available for the language) and if we allow repeated assessments of the same segment by the same person.

Hybrid supersampling is a novel way of doing meta-evaluation of metric performance and it provided more conclusive results. Although we carried out hybrid supersampling for DA human evaluation only, the method is not DA specific, and it would be interesting to trial it with RR the future.

Character-level metrics again gave very good results on both system and segment level. The trend that started on WMT14 with BEER, then continued on WMT15 with BEER and CHRF, now happens with BEER, CHRF and CHARACTER. This growing number of character-level metrics suggests that community (at least the one that develops metrics) had started to adopt character-level matching as an important component of evaluation.

Just like in previous years, metrics that train their parameters get very high correlation with human judgment as exemplified with BEER and UOW.REVAL. This year's edition of the metrics task introduced different types of golden truths that opens the question towards which golden truth should metrics be trained. Should it be for RR by using some learning-to-rank algorithms, or for DA by using regression algorithms or some combination of the two.

The results this year again include surprises. For instance, evaluation of English-to-Czech this year suggests that WORDF, BLEU and NIST outperform CHRF under evaluation against RR both with and without tuning systems (Figure 4) on the news domain, whereas we have seen the exact op-

posite last year. The IT domain for English-to-Czech stays in line with last year's observations.

BLEU (and especially its Moses implementation) has been clearly outperformed by many metrics. That again highlights the question in MT as to why almost all systems remain to be optimized for BLEU. Optimization towards BLEU has driven system development and certainly achieved results in the past, but the relatively low correlation with human judgment is a sign that some alternative metrics should be considered. For this reason, we encourage metrics developers to add their metric to Moses scorer so that the MT community can more easily experiment with employing them as optimization objective functions. An additional motivation should also be so that valuable development work on metrics is not lost in the future. If added to Moses scorer, future metrics tasks could run easily these metrics as baselines, even if their authors are not participating in the task that year. That way, good performing metrics will live on and the results of the metrics task will be more comparable across years.

4 Conclusion

In this paper, we summarized the results of the WMT16 Metrics Shared Task, which assesses the quality of various automatic machine translation metrics. As in previous years, human judgments collected in WMT16 serve as the golden truth and we check how well the metrics predict the judgments at the level of individual sentences as well as at the level of the whole test set (system-level).

The more extensive meta-evaluation in this years task that involved large number of language pairs, different types of judgments and better measurements of the significance would hopefully shed some more light on the qualities of different metrics.

The patterns that can be observed in the results are that character-level metrics perform really well and that the number of them is growing over the years. Also, the trained metrics on average are performing better than non-trained metrics, especially for into-English language pairs.

Acknowledgments

We wouldn't be able to put this experiment together without tight collaboration with Matt Post and Christian Federmann who were running the core of WMT Shared Translation Task. This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreements nº 645452 (QT21) and nº 644402 (HimL). The work on this project was also supported by the Dutch organization for scientific research STW grant nr. 12271.

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.

Alexandra Birch, Barry Haddow, Ondřej Bojar, and Omri Abend. 2016. Hume: Human ucca-based evaluation of machine translation. *arXiv preprint arXiv:1607.00030*.

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 Buck, Chris Callison-Burch, Christian Federmann, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2013. Findings of the 2013 Workshop on Statistical Machine Translation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 1–44, Sofia, Bulgaria, August. Association for Computational Linguistics.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Barry Haddow, Matthias Huck, Chris Hokamp, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Carolina Scarton, Lucia Specia, and Marco Turchi. 2015. Findings of the 2015 Workshop on Statistical Machine Translation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, Lisboa, Portugal, September. 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, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurelie Neveol, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin Verspoor, and Marcos Zampieri. 2016b. Findings of the 2016 Conference

on Machine Translation. In *Proceedings of the First Conference on Machine Translation*, Berlin, Germany, August. 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.

Marina Fomicheva, Núria Bel, Lucia Specia, Iria da Cunha, and Anton Malinovskiy. 2016. CobaltF: A Fluent Metric for MT Evaluation. In *Proceedings of the First Conference on Machine Translation*, Berlin, Germany, August. 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.

Rohit Gupta, Constantin Orasan, and Josef van Genabith. 2015a. Machine Translation Evaluation using Recurrent Neural Networks. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, Lisboa, Portugal, September. Association for Computational Linguistics.

Rohit Gupta, Constantin Orăsan, and Josef van Genabith. 2015b. ReVal: A Simple and Effective Machine Translation Evaluation Metric Based on Recurrent Neural Networks. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Lisbon, Portugal.

Bushra Jawaid, Amir Kamran, Miloš Stanojević, and Ondřej ojar. 2016. Results of the WMT16 Tuning Shared Task. In *Proceedings of the First Conference on Machine Translation*, Berlin, Germany, August. 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.

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

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.

Martin McCaffery and Mark-Jan Nederhof. 2016. DTED: Evaluation of Machine Translation Structure Using Dependency Parsing and Tree Edit Distance. In *Proceedings of the First Conference on Machine Translation*, Berlin, Germany, August. 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. 2016. chrF deconstructed: beta parameters and n-gram weights. In *Proceedings of the First Conference on Machine Translation*, Berlin, Germany, August. Association for Computational Linguistics.

Rudolf Rosa. 2014. Depfix, a tool for automatic rule-based post-editing of SMT. *The Prague Bulletin of Mathematical Linguistics*, 102:47–56.

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.

Miloš Stanojević, Amir Kamran, Philipp Koehn, and Ondřej Bojar. 2015. Results of the WMT15 Metrics Shared Task. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, Lisboa, Portugal, 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.

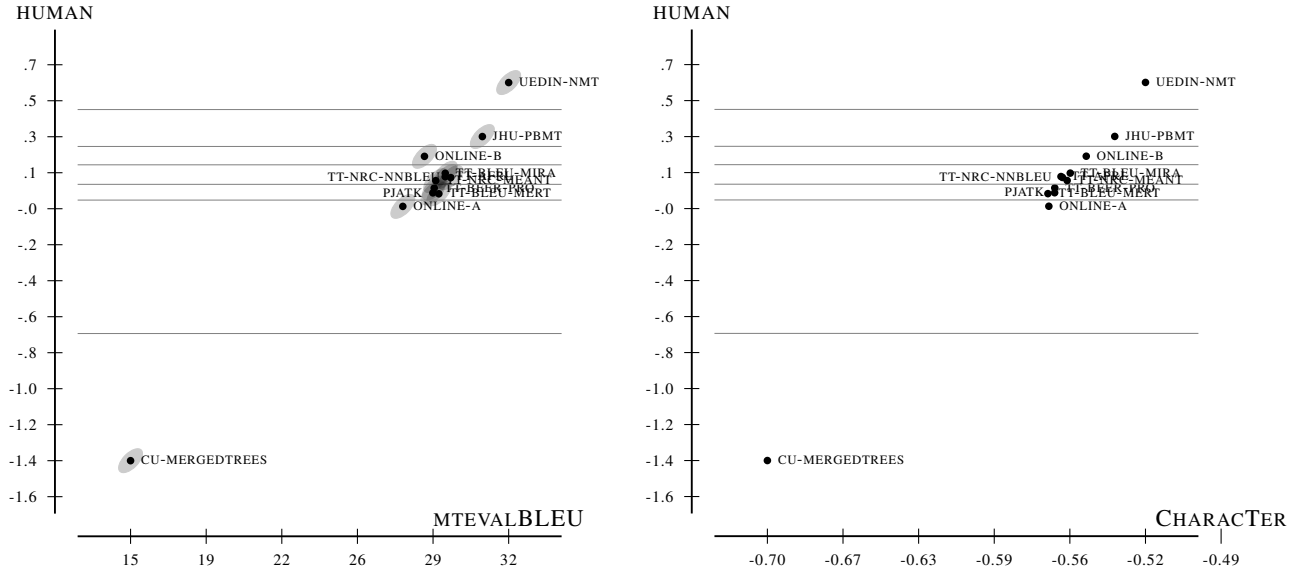
Hui Yu, Qingsong Ma, Xiaofeng Wu, and Qun Liu. 2015. CASICT-DCU Participation in WMT2015 Metrics Task. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, Lisboa, Portugal, September. Association for Computational Linguistics.

Lilin Zhang, Zhen Weng, Wenyan Xiao, Jianyi Wan, Zhiming Chen, Yiming Tan, Maoxi Li, and Mingwen Wang. 2016. Extract Domain-specific Paraphrase from Monolingual Corpus for Automatic Evaluation of Machine Translation. In *Proceedings of the First Conference on Machine Translation*, Berlin, Germany, August. Association for Computational Linguistics.

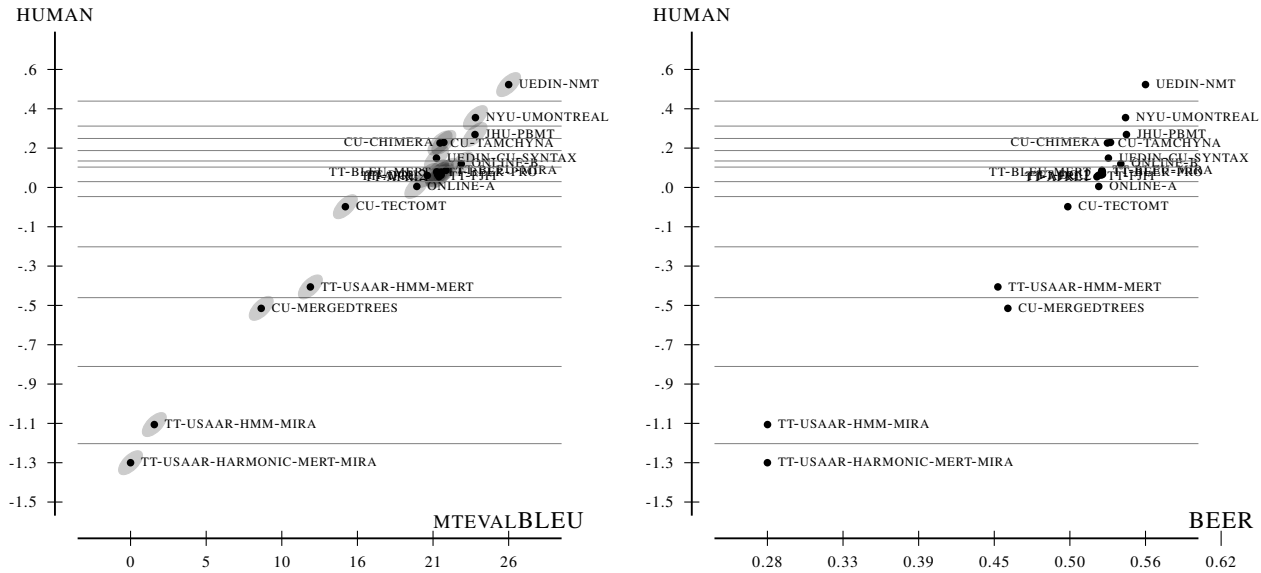
A System-Level Correlation Plots

The following figures plot the system-level results of MTEVALBLEU (left-hand plots) and the best performing (according to RR and DA, see Tables 4, 5 and 7) metrics for the given language pair (right-hand plots) against manual score.

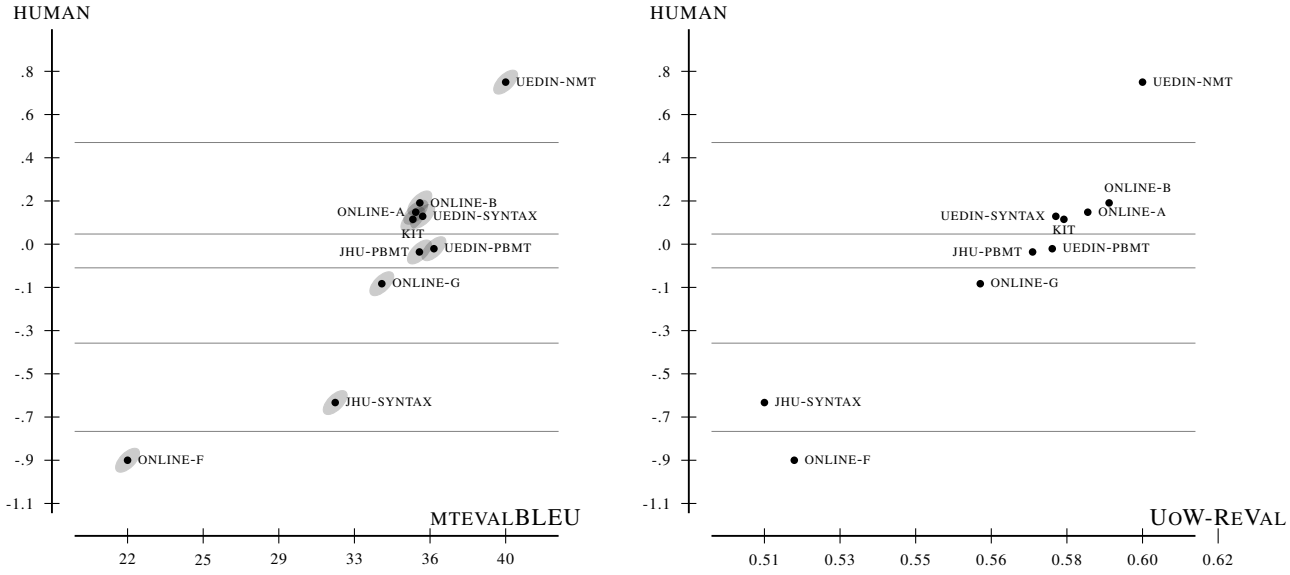
Czech-English



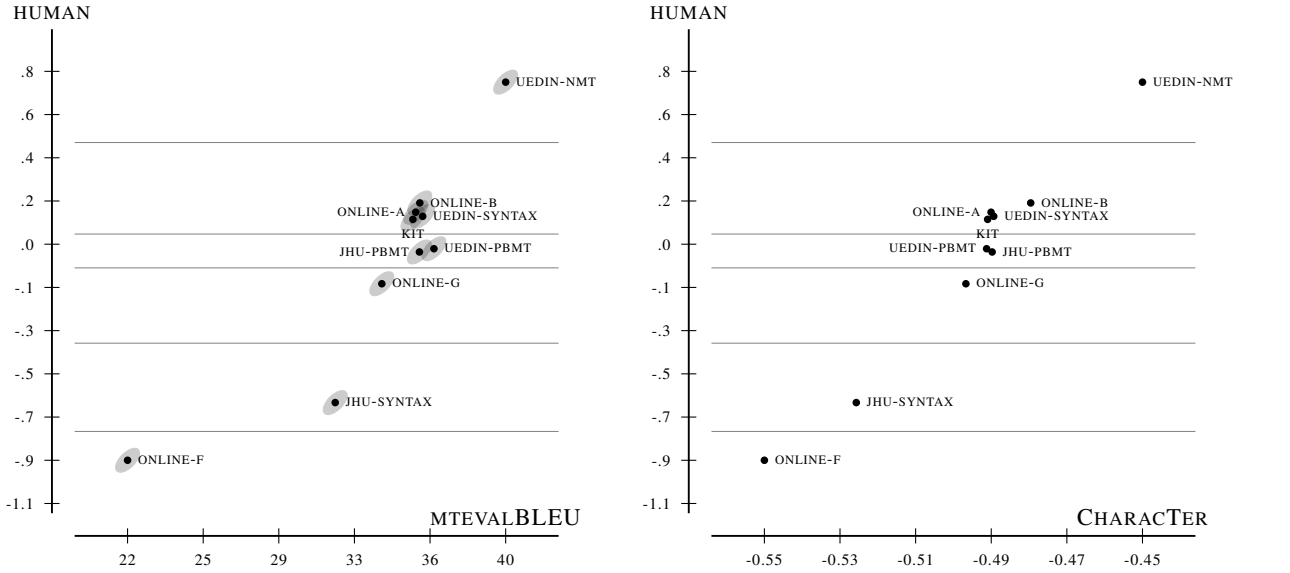
English-Czech



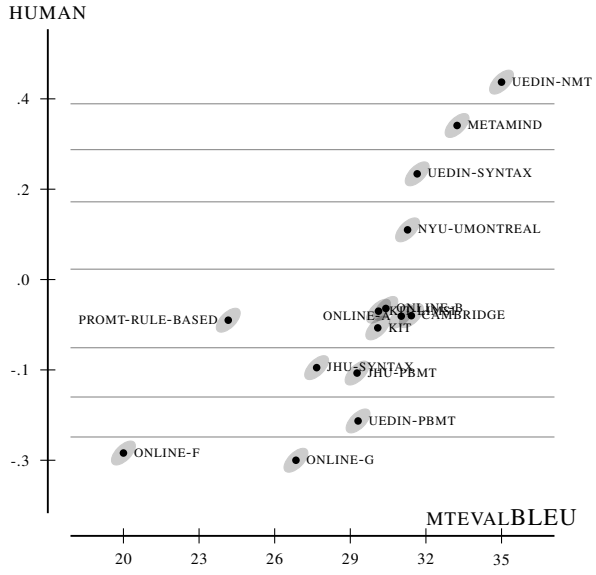
German-English

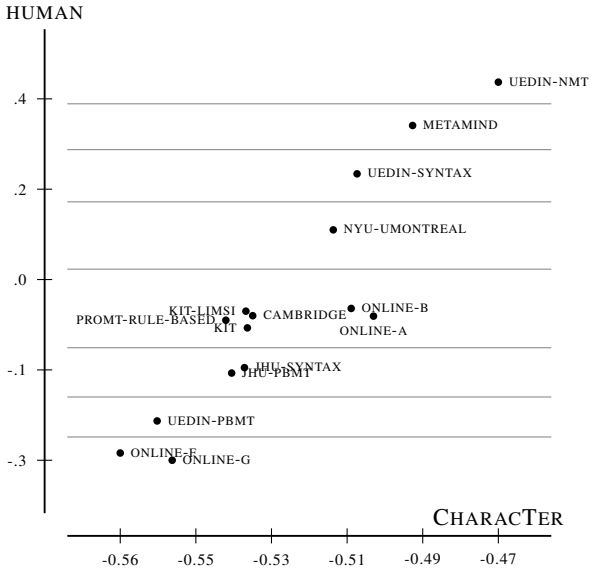


German-English

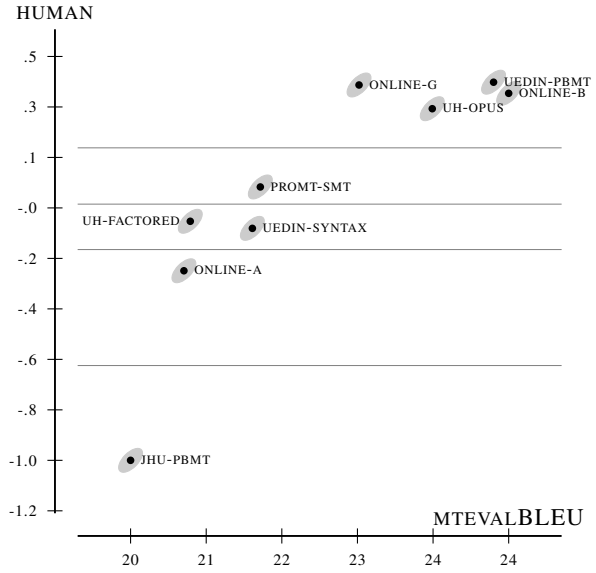


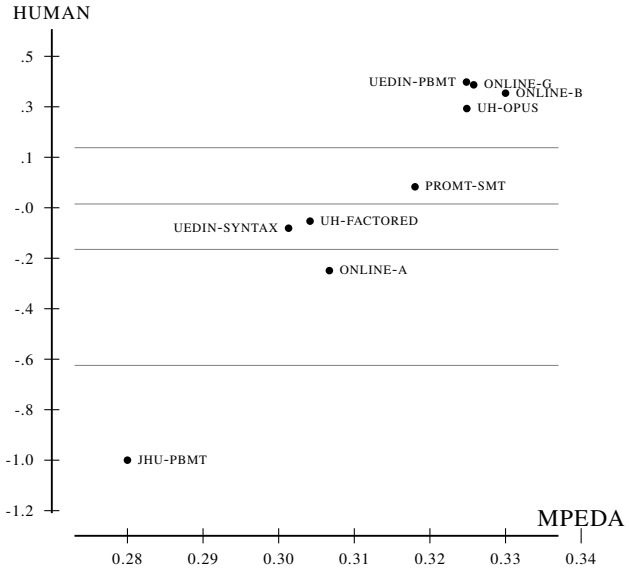
English-German



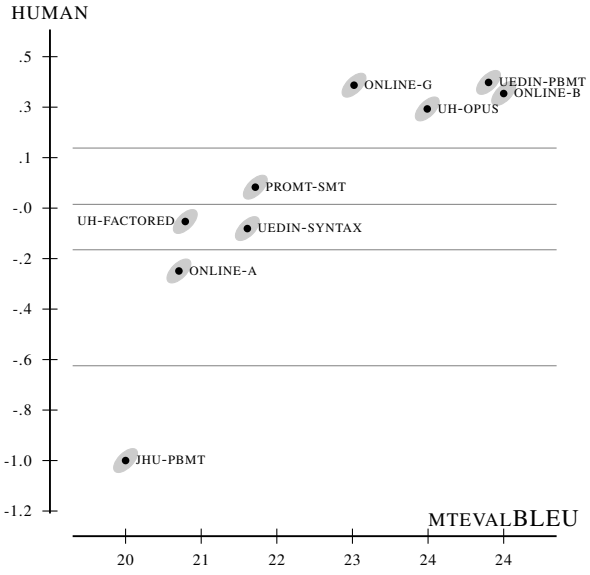


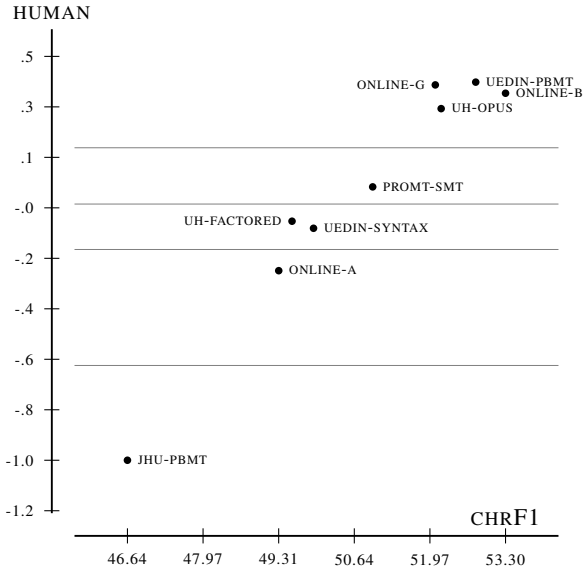
Finnish-English



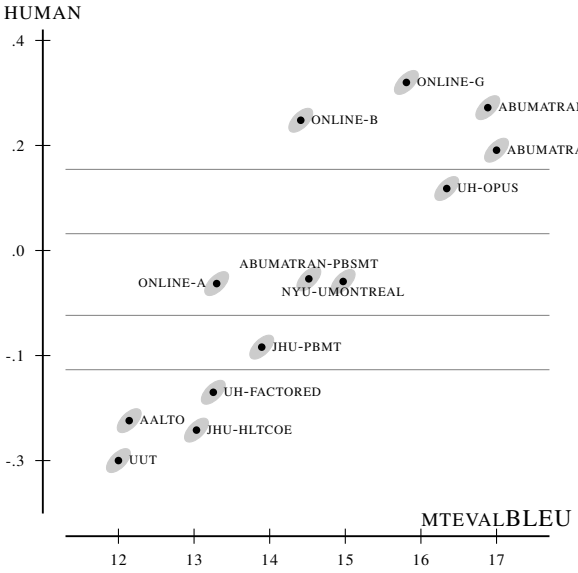


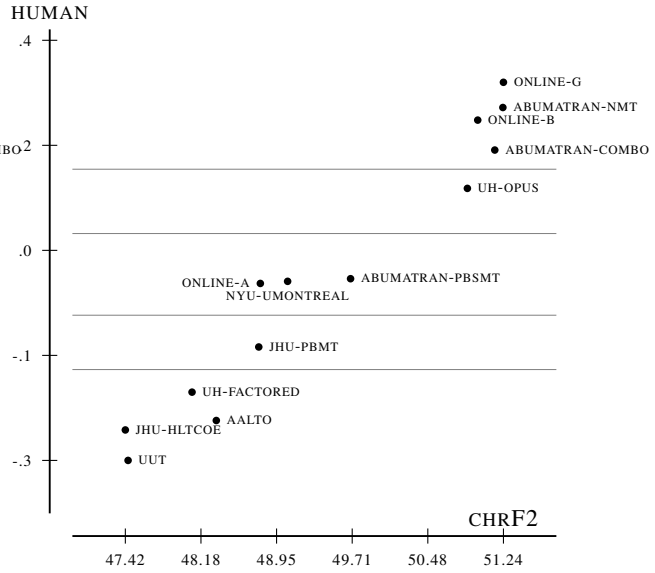
Finnish-English



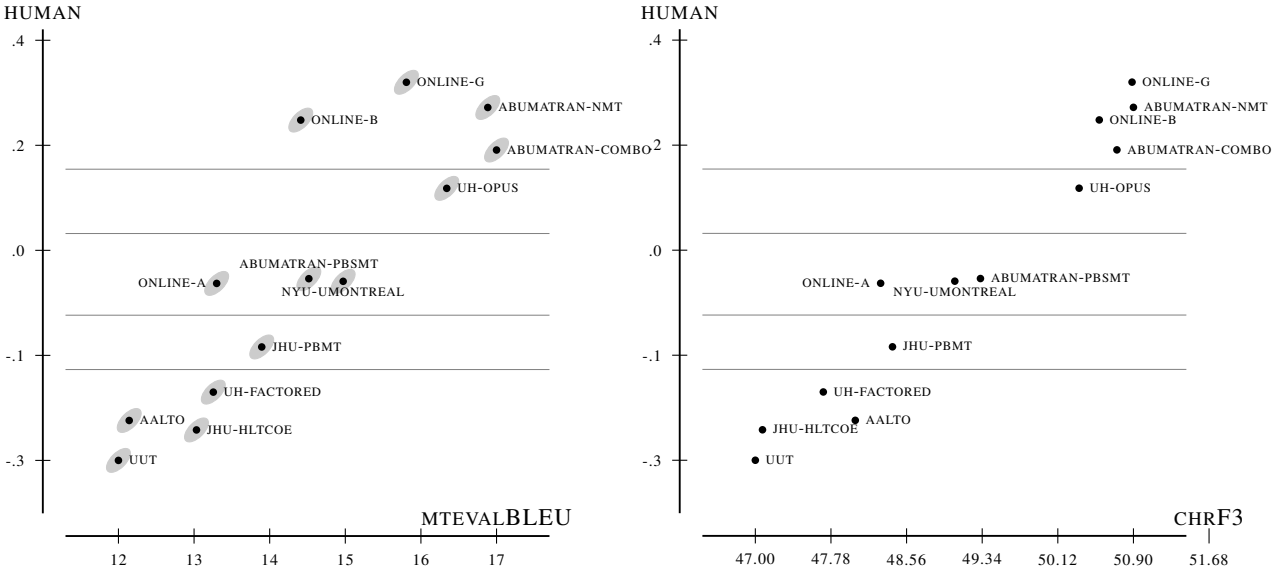


English-Finnish

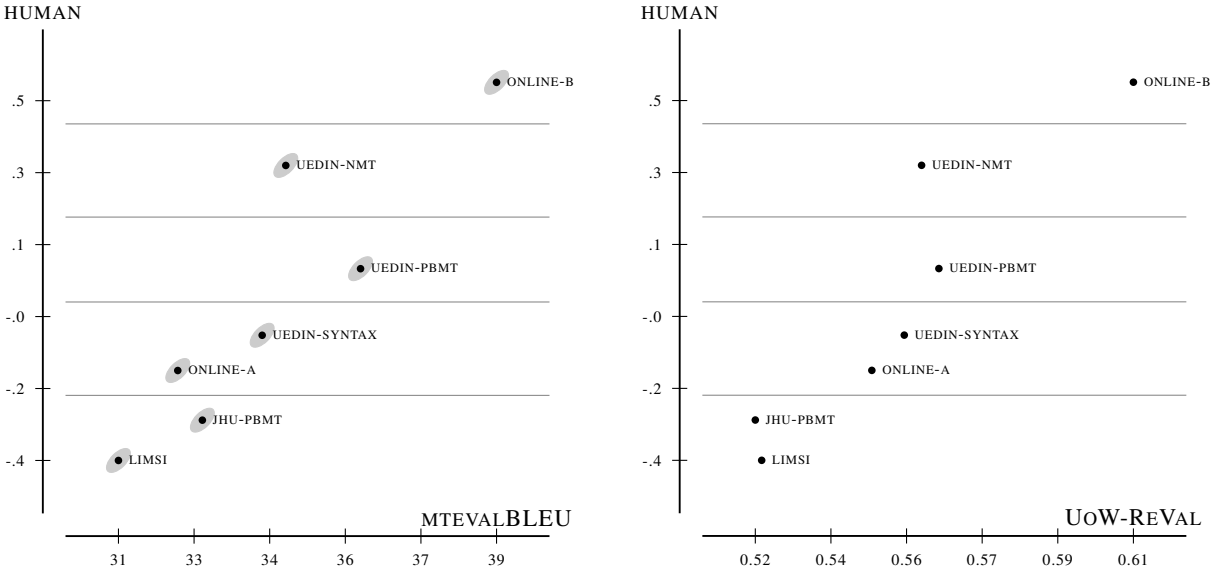




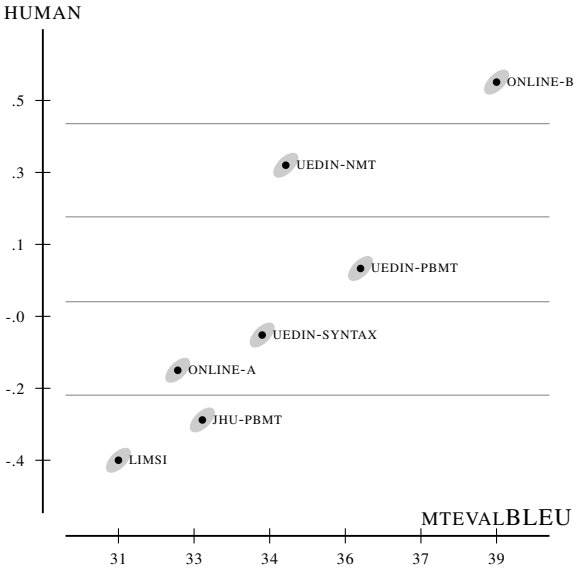
English-Finnish

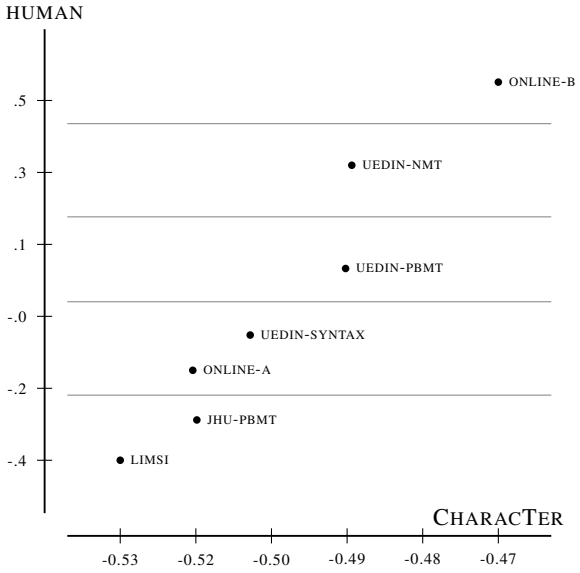


Romanian-English

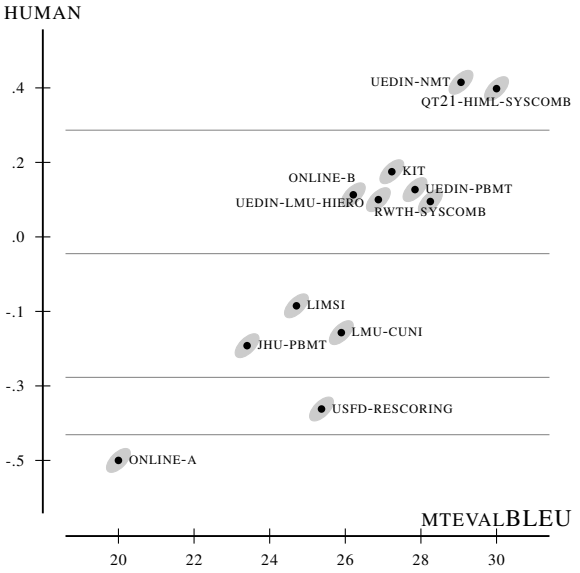


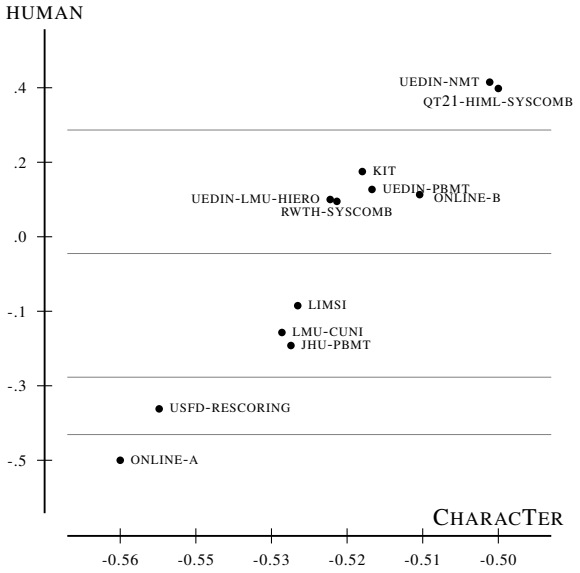
Romanian-English



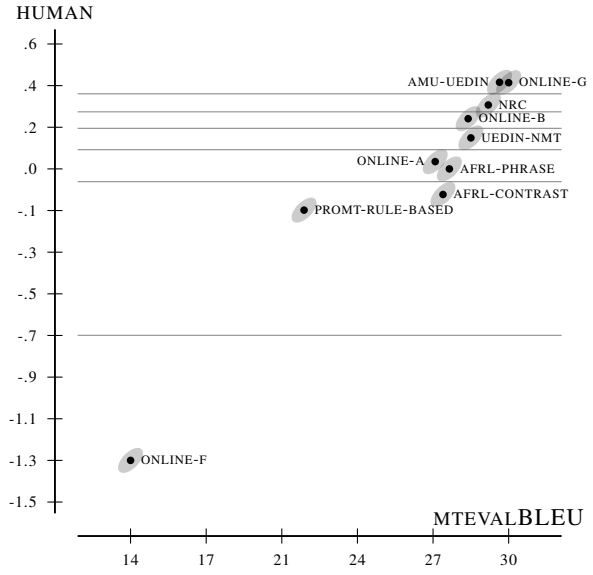


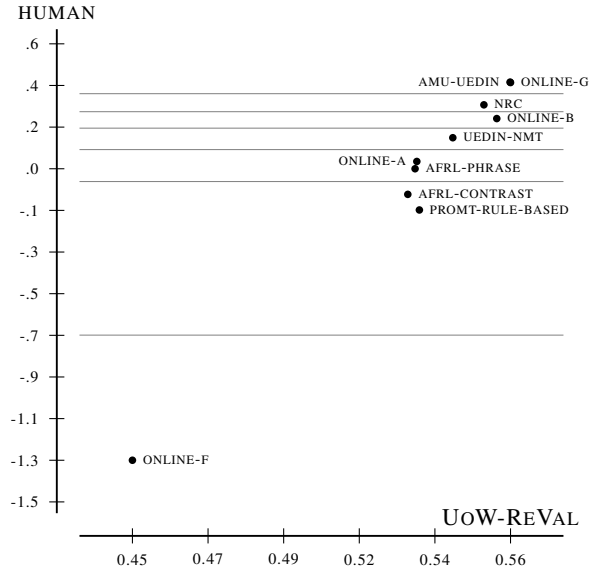
English-Romanian



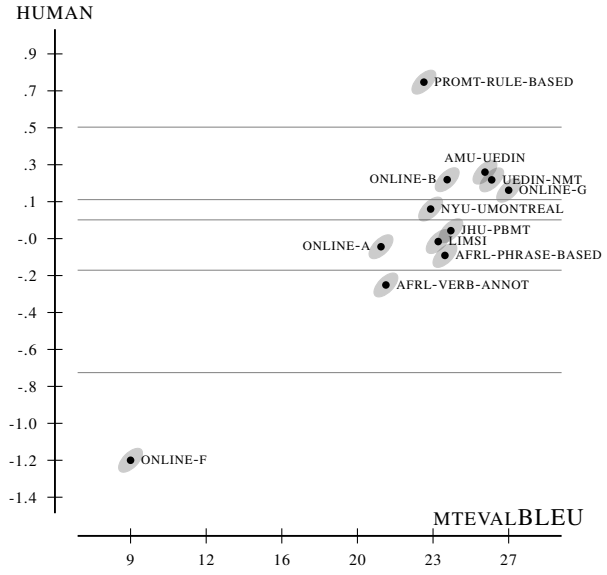


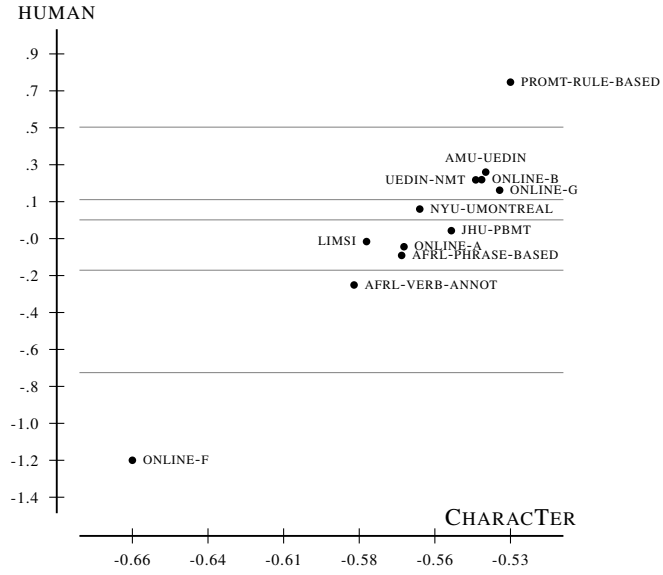
Russian-English



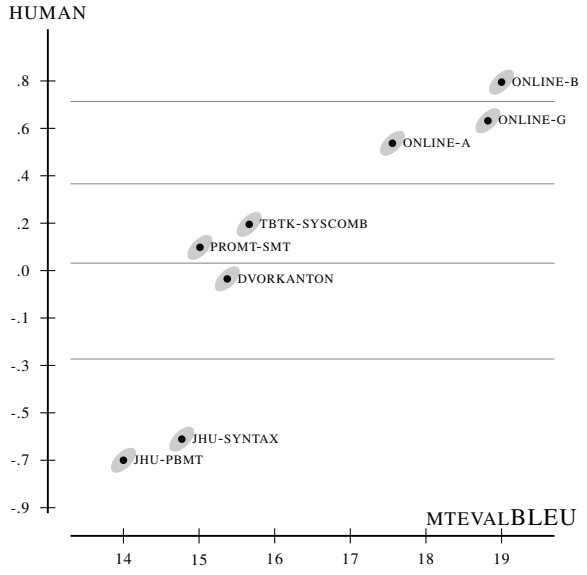


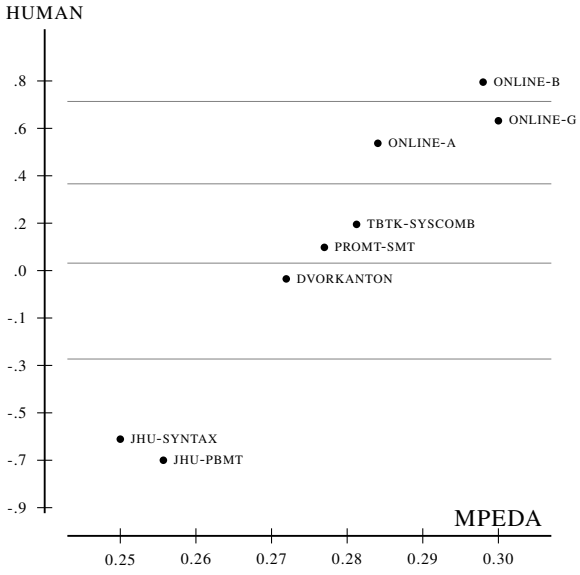
English-Russian



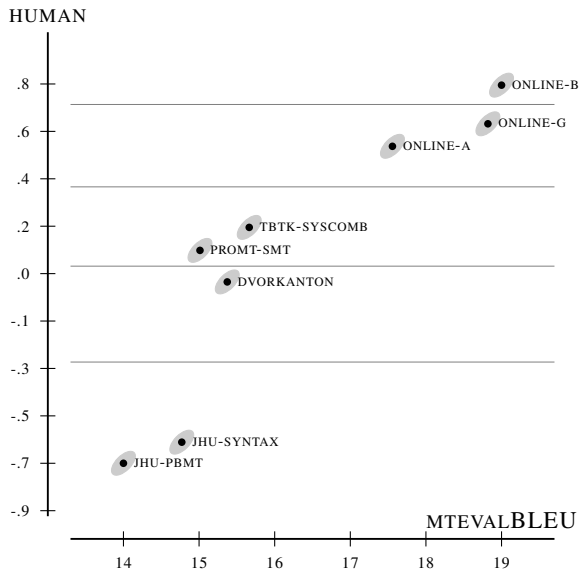


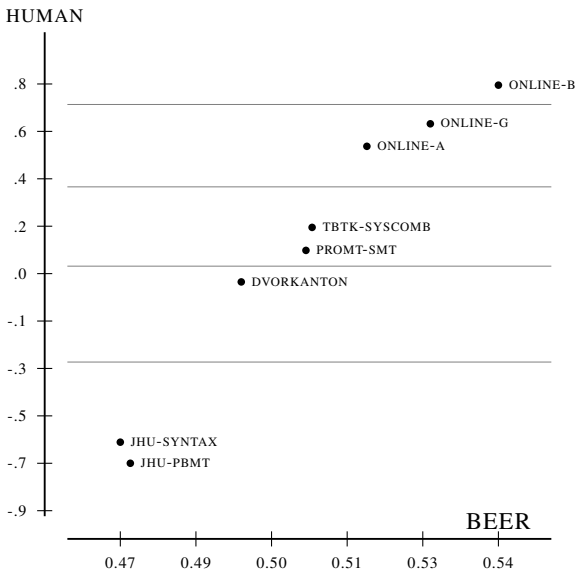
Turkish-English





Turkish-English





English-Turkish