Results of the WMT15 Metrics Shared Task

Results of the WMT15 Metrics Shared Task

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

Philipp Koehn
Johns Hopkins University
DCS
phi@jhu.edu

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

Abstract

This paper presents the results of the WMT15 Metrics Shared Task. We asked participants of this task to score the outputs of the MT systems involved in the WMT15 Shared Translation Task. We collected scores of 46 metrics from 11 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 WMT15 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).

1 Introduction

Automatic machine translation metrics play a very important role in the development of MT systems and their evaluation. There are many different metrics of diverse nature and one would like to assess their quality. For this reason, the Metrics Shared Task is held annually at the Workshop of Statistical Machine Translation[1], starting with Koehn and Monz (2006) and following up to Macháček and Bojar (2014).

The systems' outputs, human judgements and evaluated metrics are described in Section 2. The quality of the metrics in terms of system level correlation is reported in Section 3. Section 4 is devoted to segment level correlation.

2 Data

We used the translations of MT systems involved in WMT15 Shared Translation Task (Bojar et al., 2015) together with reference translations as the test set for the Metrics Task. This dataset consists of 87 systems' outputs and 10 reference translations in 10 translation directions (English from and into Czech, Finnish, French, German and Russian). The number of sentences in system and reference translations varies among language pairs ranging from 1370 for Finnish-English to 2818 for Russian-English. For more details, please see the WMT15 overview paper (Bojar et al., 2015).

2.1 Manual MT Quality Judgements

During the WMT15 Translation Task, a large scale manual annotation was conducted to compare the translation quality of participating systems. We used these collected human judgements for the evaluation of the automatic metrics.

The participants in the manual annotation were asked to evaluate system outputs by ranking translated sentences relative to each other. For each source segment that was included in the procedure, the annotator was shown five different outputs to which he or she was supposed to assign ranks. Ties were allowed.

These collected rank labels for each five-tuple of outputs were then interpreted as pairwise comparisons of systems and used to assign each system a score that reflects how high that system was usually ranked by the annotators. Several methods have been tested in the past for the exact score calculation and WMT15 has adopted TrueSkill as the official one. Please see the WMT15 overview paper for details on how this score is computed.

For the metrics task in 2014, we were still using the "Pre-TrueSkill" method called "> Others", see Bojar et al. (2011). Since we are now moving to the golden truth calculated by TrueSkill, we report also the average "Pre-TrueSkill" score in the relevant tables for comparison.

[1] http://www.statmt.org/wmt15

Proceedings of the Tenth Workshop on Statistical Machine Translation, pages 256–273,
Lisboa, Portugal, 17-18 September 2015.

Metric	Participant
BEER, BEER_TREEPEL	ILLC – University of Amsterdam (Stanojević and Sima'an, 2015)
BS	University of Zurich (Mark Fishel; no corresponding paper)
CHRF, CHRF3	DFKI (Popović, 2015)
DPMF, DPMFCOMB	Chinese Academy of Sciences and Dublin City University (Yu et al., 2015)
DREEM	National Research Council Canada (Chen et al., 2015)
LEBLEU-DEFAULT, LEBLEU-OPTIMIZED	Lingsoft and Aalto University (Virpioja and Grönroos, 2015)
METEOR-WSD, RATATOUILLE	LIMSI-CNRS (Marie and Apidianaki, 2015)
UOW-LSTM	University of Wolverhampton (Gupta et al., 2015a)
UPF-COBALT	Universitat Pompeu Fabra (Fomicheva et al., 2015)
USAAR-ZWICKEL-*	Saarland University (Vela and Tan, 2015)
VERTA-W, VERTA-EQ, VERTA-70ADEQ30FLU	University of Barcelona (Comelles and Atserias, 2015)

Table 1: Participants of WMT15 Metrics Shared Task

2.2 Participants of the Metrics Shared Task

Table 1 lists the participants of the WMT15 Shared Metrics Task, along with their metrics. We have collected 46 metrics from a total of 11 research groups.

Here we give a short description of each metric that performed the best on at least one language pair.

2.2.1 BEER and BEER_TREEPEL

BEER is a trained metric, a linear model that combines features capturing character n-grams and permutation trees. BEER has participated last year in sentence-level evalution. The main additions this year are corpus-level aggregation of sentence-level scores and a syntactic version called BEER_TREEPEL. BEER_TREEPEL includes features checking the match of each type of arc in the dependency trees of the hypothesis and the reference.

BEER was the best for en-de and en-ru at the system level and en-fi and en-ru at the sentence level. BEER_TREEPEL was the best for system-level evaluation of ru-en.

2.2.2 BS

The metric BS has no corresponding paper, so we include a summary by Mark Fishel here: The BS metric was an attempt of moving in a different direction than most state-of-the-art metrics and reduce complexity and language resource dependence to the minimum. The score is obtained from the number and lengths of "bad segments": continuous subsequences of words that are present only in the hypothesis or the reference, but not both. To account for morphologically complex languages and smooth the score for sparse word forms poor man's lemmatization is added: the floor of one third of each word's characters are removed from the word's end. The final score is either the log-sum of the bad segment lengths (BS) or a simple sum (TOTAL-BS).

BS and DPMF were the best for system-level English-French evaluation.

2.2.3 CHRF3

CHRF3 calculates a simple F-score combination of the precision and recall of character n-grams of length 6. The F-score is calculated with $\beta = 3$, giving triple the weight to recall.

CHRF3 was the best for en-fi and en-cs at the system level and en-cs at the sentence level.

2.2.4 DPMF and DPMFCOMB

DPMF is a syntax-based metric but unlike many syntax-based metrics, it does not compute score on substructures of the tree returned by a syntactic parser. Instead, DPMF parses the reference translation with a standard parser and trains a new parser on the tree of the reference translation. This new parser is then used for scoring the hypothesis. Additionally, DPMF uses F-score of unigrams in combination with the syntactic score.

DPMFCOMB is a combination of DPMF with several other metrics available in the evaluation tool Asiya[2].

DPMF and BS were the best for system-level evaluation of English-French. DPMF also tied for the best place with UOW-LSTM for French-English. DPMFCOMB was the best for fi-en, de-en and cs-en at the sentence level.

2.2.5 DREEM

DREEM uses distributed word and sentence representations of three different kinds: one-hot representation, a distributed representation learned with a neural network and a distributed sentence

[2] http://asiya-faust.cs.upc.edu/

representation learned with a recursive autoencoder. The final score is the cosine similarity of the representation of the hypothesis and the reference, multiplied with a length penalty.

DREEM was the best for fi-en system-level evaluation.

2.2.6 LEBLEU-OPTIMIZED

LEBLEU is a relaxation of the strict word n-gram matching that is used in standard BLEU. Unlike other similar relaxations, LEBLEU uses fuzzy matching of longer chunks of text that allows, for example, to match two independent words with a compound. LEBLEU-OPTIMIZED applies fuzzy match threshold and n-gram length optimized for each language pair.

LEBLEU-OPTIMIZED was the best for en-de at the sentence level.

2.2.7 RATATOUILLE

RATATOUILLE is a metric combination of BLEU, BEER, Meteor and few more metrics out of which METEOR-WSD is a novel contribution. METEOR-WSD is an extension of Meteor that includes synonym mappings to languages other than English based on alignments and rewards semantically adequate translations in context.

RATATOUILLE was the best for sentence-level French-English evaluation in both directions.

2.2.8 UOW-LSTM

UOW-LSTM uses dependency-tree recursive neural network to represent both the hypothesis and the reference with a dense vector. The final score is obtained from a neural network trained on judgements from previous years converted to similarity scores, taking into account both the distance and angle of the two representations.

UOW-LSTM tied for the best place in fr-en system-level evaluation with DPMF.

2.2.9 UPF-COBALT

UPF-COBALT pays an increased attention to syntactic context (for example arguments, complements, modifiers etc.) both in aligning the words of the hypothesis and reference as well as in scoring of the matched words. It relies on additional resources including stemmers, WordNet synsets, paraphrase databases and distributed word representations. UPF-COBALT system-level score was calculated by taking the ratio of sentences in which each system from a set of competitors was assigned the highest sentence-level score.

UPF-COBALT was the best on system-level evaluation for de-en and, together with VERTA-70ADEQ30FLU, for cs-en.

2.2.10 VERTA-70ADEQ30FLU

VERTA-70ADEQ30FLU aims at the combination of adequacy and fluency features that use many sources of different linguistic information: synonyms, lemmas, PoS tags, dependency parses and language models. On previous works VERTA's linguistic features combination were set depending on whether adequacy or fluency was evaluated. VERTA-70ADEQ30FLU is a weighted combination of VERTA setups for adequacy (0.70) and fluency (0.30).

VERTA-70ADEQ30FLU was, together with UPF-COBALT, the best on cs-en on system level.

2.2.11 Baseline Metrics

In addition to the submitted metrics, we have computed the following two groups of standard metrics as baselines for the system level:

- **Mteval.** The metrics BLEU (Papineni et al., 2002) and NIST (Doddington, 2002) were computed using the script `mteval-v13a.pl`[3] 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 TER (Snover et al., 2006), WER, PER and CDER (Leusch et al., 2006) were computed using 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.

For segment level baseline, we have used 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 judgements for segment level.

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

We have normalized all metrics' scores such that better translations get higher scores.

For computing the scores we used the same script from the last year metric task.

3 System-Level Results

Same as last year, we used Pearson correlation coefficient as the main measure for system level metrics correlation. We use the following formula to compute the Pearson's r for each metric and translation direction:

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

where H is the vector of human scores of all systems translating in the given direction, M is the vector of the corresponding scores as predicted by the given metric. $\bar{H}$ and $\bar{M}$ are their means respectively.

Since we have normalized all metrics such that better translations get higher score, we consider metrics with values of Pearson's r closer to 1 as better.

You can find the system-level correlations for translations into English in Table 2 and for translations out of English in Table 3. Each row in the tables contains correlations of a metric in each of the examined translation directions. The upper part of each table lists metrics that participated in all language pairs and it is sorted by average Pearson correlation coefficient across translation directions. The lower part contains metrics limited to a subset of the language pairs, so the average correlation cannot be directly compared with other metrics any more. The best results in each direction are in bold. The reported empirical confidence intervals of system level correlations were obtained through bootstrap resampling of 1000 samples (confidence level of 95%).

The move to TrueSkill golden truth slightly increased the correlations and changed the ranking of the metrics a little, but the general patterns hold. (The correlation between "Average" and "Pre-TrueSkill Average" is .999 for both directions.)

Both tables also include the average Spearman's rank correlation, which used to be the evaluation measure in the past. Spearman's rank correlation considers only the ranking of the systems and not the distances between them. It is thus more susceptible to instability if several systems have similar scores.

3.1 System-Level Discussion

As in the previous years, many metrics outperform BLEU both into as well as out of English. Note that the original BLEU was designed to work with 4 references and WMT provides just one; see Bojar et al. (2013) for details on BLEU correlation with varying number of references, up to several thousands. This year, BLEU with one reference reaches the average correlation of .92 into English or .78 out of English. The best performing metrics get up to .98 into English and .92 out of English. CDER is the best of the baselines, reaching .94 into English and .81 out of English.

The winning metric for each language pair is different, with interesting outliers: DREEM performed best when evaluating English translations from Finnish but on average, 12 other metrics into English performed better and DREEM appears to be among the worst metrics out of English. RATATOUILLE is fifth to tenth when evaluated by average Pearson but wins in both directions in average Spearman's rank correlation.

Two metrics confirm the effectiveness of character-level measures, esp. the winners for out of English evaluation: CHRF3 and BEER. The metric CHRF3 is particularly interesting because it does not require any resources whatsoever. It is defined as a simple F-measure of character-level 6-grams (spaces are ignored), with recall weighted 3 times more than precision. The balance between the precision and recall seems important depending on morphological richness of the target language: for evaluations into English, CHRF (equal weights) performs better than CHRF3.

As we already observed in the past, the winning metrics are trained on previous years of WMT. This holds for DPMFCOMB, UOW-LSTM and BEER including BEER_TREEPEL. DPMF and UPF-COBALT are not combination or trained metrics of any kind, DPMF is based on dependency analysis of the candidate and reference sentences and UPF-COBALT uses contextual information of compared words in the candidate and the reference.

We see an interesting difference in the performance of UOW-LSTM. It is the second metric in system-level correlation but falls among the worst

ones in segment-level correlations, see Table 4 below. Gupta et al. (2015b) suggest that the discrepancy in performance could be based by low inter-annotator agreement and Kendall's τ not reflecting the distances in translation quality between candidates, an issue similar to what we see with Pearson vs. Spearman's rank correlations.

Another dense-representation metric, DREEM, seems to suffer a similar discrepancy when evaluating into English. Out of English, DREEM did not perform very well.

An untested speculation is that the dense sentence-level representation present in some form in both UOW-LSTM as well as in DREEM confuses the metrics in their judgements of individual sentences.

3.2 Comparison with BLEU

In Appendix A, we provide two correlation plots for each language pair. The first plot visualizes the correlation of BLEU and manual judgements, the second plot shows the correlation for the best performing metric for that pair.

The BLEU plots include grey ellipses to indicate the confidence intervals of both BLEU as well as manual judgements. The ellipses are tilted only to indicate that BLEU and the manual score are dependent variables. Only the width and height of each ellipse represent a value, that is the confidence interval in each direction. The same vertical confidence intervals hold for plots in the right-hand column, but since we don't have any confidence estimates for the individual metrics, we omit them.

Czech-English plots indicate that UPF-COBALT was able to account for the very different behaviour of the transfer-based deep-syntactic system CU-TECTO. It was also able to appreciate the higher translation quality of montreal, UEDIN-* and online-b. The big cluster of systems labelled TT-* are submissions to the WMT15 Tuning Task (Stanojević et al., 2015).

For English-Czech, we see that UEDIN-JHU and MONTREAL are overfit for BLEU. In terms of BLEU, they are very close to the winning system CU-CHIMERA (a combination of CU-TECTO and phrase-based Moses, followed by automatic post-editing). CHRF3 is able to recognize the overfitting for MONTREAL, a neural-network based system, but not for UEDIN-JHU. CHRF3 also better recognizes the distance in quality between larger systems (from COMMERCIAL1 above) and the small-data tuning task systems.

For German-English, we see the same overfit of UEDIN-JHU towards BLEU. While neither UPF-COBALT nor CHRF3 could recognize this for translations involving Czech, the issue is spotted by UPF-COBALT for systems involving German. Syntax-based systems like UEDIN-SYNTAX for English-German and (presumably) ONLINE-B for German-English are among those where the correlation got most improved over BLEU.

The French dataset was in a different domain, which may explain why the best performing metric DPMF does actually not improve much above BLEU. DPMF uses a syntactic parser on the reference, and the performance of parsers on discussions is likely to be lower than the generally used news domain.

In Finnish results, we see again UEDIN-JHU and ABUMATRAN (Rubino et al., 2015) overvalued by BLEU. DREEM based on distributed representation of words and sentences is able to recognize this for translation into English but it falls among the worst metrics in the other direction. For translation into Finnish, character-based n-grams of CHRF3 are much more reliable. Variants of ABUMATRAN were again those most overvalued by BLEU. ABUMATRAN uses several types of morphological segmentation and reconstructs Finnish words from the segments by concatenation. ABUMATRAN is loaded with many other features, like web-crawled data and domain handling, and system combination of several approaches. The optimization towards BLEU (unreliable for Finnish, as we have learned in this task), could be among the main reasons behind the comparably lower manual scores.

For Russian, BEER is the best metric, in its syntax-aware variant BEER_TREEPEL for evaluating English. Compared to BLEU, the improvement in correlation is not that striking for Russian-English. (It would be interesting to know whether ONLINE-G is better than ONLINE-B because of English syntax or addressing source-side morphology better. BEER_TREEPEL captures both aspects.) In the other direction, targetting Russian, BLEU was effectively unable to rank the systems at all. It is probably the character-level features in BEER that allow it to reach a very good correlation, .97.

Correlation coefficient	Pearson Correlation Coefficient							Spearman's
Direction	fr-en	fi-en	de-en	cs-en	ru-en	Average	Pre-TrueSkill	Average
Considered Systems	7	14	13	16	13		Average	
DPMFCOMB	.995 ± .004	.958 ± .011	.973 ± .009	.991 ± .002	.974 ± .008	**.978** ± .007	.970 ± .012	.882 ± .041
UOW-LSTM	**.997** ± .003	.976 ± .008	.960 ± .010	.983 ± .003	.963 ± .009	.976 ± .007	≀**.976** ± .011	≀.916 ± .038
BEER_TREEPEL	.981 ± .008	.971 ± .010	.952 ± .012	.992 ± .002	**.981** ± .008	.975 ± .008	.962 ± .014	.861 ± .051
DPMF	**.997** ± .003	.951 ± .011	.960 ± .010	.984 ± .003	.973 ± .008	.973 ± .007	≀.965 ± .012	≀.893 ± .035
UPF-COBALT	.987 ± .006	.962 ± .010	**.981** ± .007	**.993** ± .002	.929 ± .014	.971 ± .008	≀.970 ± .012	.888 ± .040
METEOR-WSD	.982 ± .007	.950 ± .012	.953 ± .011	.983 ± .003	.976 ± .008	.969 ± .008	.960 ± .014	.832 ± .051
BEER	.979 ± .008	.965 ± .010	.946 ± .012	.983 ± .003	.971 ± .009	.969 ± .009	.958 ± .015	≀.838 ± .049
VERTA-70ADEQ30FLU	.982 ± .007	.949 ± .012	.934 ± .014	**.993** ± .002	.972 ± .010	.966 ± .009	.952 ± .015	≀.883 ± .038
VERTA-W	.977 ± .008	.955 ± .011	.928 ± .015	.988 ± .003	.964 ± .011	.963 ± .010	.949 ± .016	.873 ± .042
CHRF	.993 ± .005	.947 ± .012	.934 ± .014	.981 ± .004	.938 ± .013	.959 ± .009	.944 ± .016	.871 ± .037
RATATOUILLE	.986 ± .006	.902 ± .016	.958 ± .011	.961 ± .005	.955 ± .011	.952 ± .010	≀.956 ± .014	≀**.919** ± .039
VERTA-EQ	.983 ± .007	.921 ± .015	.906 ± .017	.990 ± .003	.953 ± .012	.950 ± .011	.934 ± .017	.857 ± .041
DREEM	.950 ± .012	**.977** ± .008	.889 ± .018	.986 ± .003	.929 ± .015	.946 ± .011	.927 ± .018	.825 ± .053
CDER	.983 ± .007	.966 ± .009	.890 ± .018	.960 ± .005	.920 ± .016	.944 ± .011	.923 ± .018	.814 ± .046
CHRF3	.979 ± .008	.903 ± .016	.956 ± .011	.968 ± .004	.898 ± .016	.941 ± .011	≀.944 ± .016	≀.818 ± .047
NIST	.980 ± .008	.894 ± .016	.901 ± .017	.973 ± .004	.910 ± .017	.932 ± .013	.906 ± .020	≀.828 ± .055
LEBLEU-DEFAULT	.955 ± .012	.900 ± .016	.916 ± .016	.947 ± .006	.908 ± .015	.925 ± .013	≀.926 ± .019	.814 ± .049
LEBLEU-OPTIMIZED	.984 ± .007	.900 ± .016	.916 ± .016	.976 ± .004	.842 ± .020	.923 ± .013	≀.928 ± .018	≀.855 ± .042
BS	.986 ± .007	.925 ± .014	.872 ± .019	.976 ± .004	.847 ± .021	.921 ± .013	.891 ± .021	.793 ± .045
PER	.978 ± .008	.871 ± .019	.846 ± .021	.963 ± .005	.931 ± .015	.918 ± .014	≀.898 ± .021	≀.811 ± .050
BLEU	.975 ± .009	.929 ± .014	.865 ± .020	.957 ± .006	.851 ± .022	.915 ± .014	.889 ± .021	.796 ± .052
TER	.979 ± .008	.872 ± .019	.890 ± .018	.907 ± .008	.907 ± .017	.911 ± .014	.884 ± .022	.768 ± .054
WER	.977 ± .009	.853 ± .020	.884 ± .018	.888 ± .008	.895 ± .018	.899 ± .015	.871 ± .023	.747 ± .057
USAAR-ZWICKEL-METEOR-MEDIAN	n/a	.936 ± .013	.961 ± .010	.976 ± .004	.965 ± .010	.959 ± .009	.955 ± .014	.871 ± .034
USAAR-ZWICKEL-METEOR-HARMONIC	n/a	.509 ± .032	.565 ± .030	.690 ± .013	.309 ± .034	.518 ± .027	.545 ± .041	.768 ± .033
USAAR-ZWICKEL-COSINE2METEOR-MEDIAN	n/a	−.220 ± .037	−.098 ± .037	.500 ± .015	.042 ± .035	.056 ± .031	.086 ± .046	−.038 ± .071
USAAR-ZWICKEL-METEOR-MEAN	n/a	.952 ± .011	.957 ± .011	.985 ± .003	.976 ± .008	.968 ± .008	.957 ± .014	.854 ± .034
USAAR-ZWICKEL-METEOR-ARIGEO	n/a	.952 ± .011	.957 ± .011	.985 ± .003	.976 ± .008	.968 ± .008	.957 ± .014	.854 ± .034
USAAR-ZWICKEL-METEOR-RMS	n/a	.958 ± .011	.944 ± .013	.988 ± .003	.974 ± .009	.966 ± .009	.947 ± .015	.861 ± .032
USAAR-ZWICKEL-COMET-RMS	n/a	.873 ± .019	.898 ± .016	.877 ± .009	.846 ± .019	.874 ± .016	.842 ± .025	.705 ± .050
USAAR-ZWICKEL-COMET-ARIGEO	n/a	.836 ± .021	.844 ± .020	.844 ± .010	.825 ± .021	.837 ± .018	.819 ± .028	.718 ± .049
USAAR-ZWICKEL-COSINE2METEOR-RMS	n/a	−.088 ± .038	−.302 ± .035	.390 ± .016	.379 ± .035	.095 ± .031	.087 ± .045	.038 ± .076
USAAR-ZWICKEL-COSINE-MEDIAN	n/a	−.414 ± .035	−.514 ± .033	.816 ± .010	.440 ± .035	.082 ± .028	.047 ± .041	−.020 ± .070
USAAR-ZWICKEL-COMET-MEAN	n/a	.836 ± .021	.844 ± .020	.844 ± .010	.825 ± .021	.837 ± .018	.819 ± .028	.718 ± .049
USAAR-ZWICKEL-COMET-HARMONIC	n/a	.445 ± .034	.525 ± .031	.602 ± .015	.307 ± .034	.470 ± .028	.487 ± .043	.561 ± .053
USAAR-ZWICKEL-COMET-MEDIAN	n/a	−.108 ± .038	.135 ± .036	.638 ± .013	.167 ± .035	.208 ± .030	.235 ± .046	.146 ± .069
USAAR-ZWICKEL-COSINE2METEOR-MEAN	n/a	−.119 ± .037	−.389 ± .034	.441 ± .016	.371 ± .035	.076 ± .031	.087 ± .045	.038 ± .076
USAAR-ZWICKEL-COSINE2METEOR-ARIGEO	n/a	−.119 ± .037	−.389 ± .034	.441 ± .016	.371 ± .035	.076 ± .031	.087 ± .045	.038 ± .076
USAAR-ZWICKEL-COSINE2METEOR-HARMONIC	n/a	−.341 ± .035	−.178 ± .038	−.050 ± .017	.253 ± .034	−.079 ± .031	−.083 ± .046	.025 ± .073
USAAR-ZWICKEL-COSINE-MEAN	n/a	nan	.002 ± .038	.906 ± .007	nan	nan	nan	.133 ± .052
USAAR-ZWICKEL-COSINE-HARMONIC	n/a	nan	−.124 ± .038	.897 ± .007	nan	nan	nan	.038 ± .048
USAAR-ZWICKEL-COSINE-RMS	n/a	nan	.064 ± .038	.910 ± .007	nan	nan	nan	.146 ± .052

Table 2: System-level correlations of automatic evaluation metrics and the official WMT human scores when translating into English. The symbol "≀" indicates where the average is out of sequence compared to the main Pearson average.

Correlation coefficient	Pearson Correlation Coefficient							Spearman's
Direction	en-fr	en-fi	en-de	en-cs	en-ru	Average	Pre-TrueSkill	Average
Considered Systems	7	10	16	15	10		Average	
CHRF3	.932 ± .018	**.878** ± .017	.848 ± .020	**.977** ± .003	.946 ± .008	**.916** ± .013	.899 ± .021	.835 ± .032
BEER	.961 ± .014	.808 ± .021	**.879** ± .018	.962 ± .003	**.970** ± .006	**.916** ± .012	≀**.907** ± .018	≀.891 ± .036
LEBLEU-DEFAULT	.933 ± .018	.835 ± .020	.850 ± .019	.953 ± .004	.896 ± .011	.893 ± .014	.875 ± .021	.846 ± .042
LEBLEU-OPTIMIZED	.933 ± .018	.803 ± .022	.868 ± .019	.952 ± .004	.908 ± .010	.893 ± .014	≀.882 ± .021	.845 ± .043
RATATOUILLE	.957 ± .015	.763 ± .025	.862 ± .019	.965 ± .003	.913 ± .010	.892 ± .014	.868 ± .021	≀**.915** ± .029
CHRF	.930 ± .018	.841 ± .021	.690 ± .027	.971 ± .003	.915 ± .010	.869 ± .016	.846 ± .023	.837 ± .027
METEOR-WSD	.959 ± .014	.760 ± .024	.650 ± .029	.953 ± .004	.892 ± .011	.843 ± .017	.816 ± .024	.837 ± .036
CDER	.953 ± .015	.640 ± .029	.660 ± .028	.929 ± .004	.863 ± .012	.809 ± .018	.777 ± .025	.704 ± .051
NIST	.949 ± .015	.692 ± .028	.502 ± .032	.958 ± .003	.893 ± .003	.799 ± .018	.771 ± .026	≀.769 ± .047
TER	.948 ± .015	.614 ± .032	.564 ± .031	.917 ± .005	.883 ± .011	.785 ± .019	.755 ± .026	.724 ± .050
WER	.941 ± .016	.608 ± .032	.568 ± .030	.910 ± .005	.884 ± .011	.782 ± .019	.752 ± .027	.702 ± .051
BLEU	.948 ± .016	.602 ± .030	.573 ± .030	.936 ± .004	.841 ± .013	.780 ± .019	.751 ± .027	.691 ± .052
PER	.949 ± .016	.603 ± .031	.316 ± .035	.908 ± .004	.858 ± .013	.727 ± .020	.696 ± .028	.609 ± .030
BS	**.964** ± .013	−.336 ± .035	.714 ± .026	.953 ± .004	.852 ± .013	.629 ± .018	.625 ± .025	≀.686 ± .049
DREEM	.871 ± .023	.385 ± .032	−.074 ± .039	.883 ± .006	.968 ± .006	.607 ± .021	.608 ± .031	.682 ± .039
DPMF	**.964** ± .014	n/a	.724 ± .026	n/a	n/a	.844 ± .020	.827 ± .027	.823 ± .048
USAAR-ZWICKEL-METEOR-MEDIAN	n/a	n/a	.741 ± .025	n/a	n/a	.741 ± .025	.685 ± .038	.750 ± .046
USAAR-ZWICKEL-METEOR-MEAN	n/a	n/a	.635 ± .029	n/a	n/a	.635 ± .029	.581 ± .041	.615 ± .041
USAAR-ZWICKEL-METEOR-RMS	n/a	n/a	.542 ± .033	n/a	n/a	.542 ± .033	.494 ± .044	.541 ± .041
USAAR-ZWICKEL-COMET-HARMONIC	n/a	n/a	.396 ± .033	n/a	n/a	.396 ± .033	.386 ± .045	.309 ± .057
USAAR-ZWICKEL-METEOR-HARMONIC	n/a	n/a	.357 ± .032	n/a	n/a	.357 ± .032	.330 ± .048	≀.550 ± .053
USAAR-ZWICKEL-COSINE-MEDIAN	n/a	n/a	.310 ± .036	n/a	n/a	.310 ± .036	.330 ± .048	.291 ± .071
USAAR-ZWICKEL-COMET-ARIGEO	n/a	n/a	.310 ± .037	n/a	n/a	.310 ± .037	.304 ± .048	≀.671 ± .050
USAAR-ZWICKEL-COSINE2METEOR-MEDIAN	n/a	n/a	.044 ± .037	n/a	n/a	.044 ± .037	.031 ± .051	−.047 ± .066
USAAR-ZWICKEL-COSINE2METEOR-HARMONIC	n/a	n/a	−.004 ± .038	n/a	n/a	−.004 ± .038	.059 ± .050	≀.009 ± .044
USAAR-ZWICKEL-COMET-MEDIAN	n/a	n/a	−.048 ± .038	n/a	n/a	−.048 ± .038	−.061 ± .050	≀.032 ± .057
USAAR-ZWICKEL-COMET-RMS	n/a	n/a	−.117 ± .039	n/a	n/a	−.117 ± .039	−.127 ± .050	≀.415 ± .054
USAAR-ZWICKEL-COMET-MEAN	n/a	n/a	−.126 ± .039	n/a	n/a	−.126 ± .039	−.135 ± .051	.412 ± .050
USAAR-ZWICKEL-COSINE2METEOR-ARIGEO	n/a	n/a	−.155 ± .036	n/a	n/a	−.155 ± .036	−.156 ± .050	−.168 ± .065
USAAR-ZWICKEL-COSINE2METEOR-MEAN	n/a	n/a	−.155 ± .036	n/a	n/a	−.155 ± .036	−.156 ± .050	−.168 ± .065
USAAR-ZWICKEL-COSINE2METEOR-RMS	n/a	n/a	−.197 ± .035	n/a	n/a	−.197 ± .035	−.188 ± .050	≀−.188 ± .063
USAAR-ZWICKEL-METEOR-ARIGEO	n/a	n/a	−.419 ± .034	n/a	n/a	−.419 ± .034	−.336 ± .050	−.162 ± .071

Table 3: System-level correlations of automatic evaluation metrics and the official WMT human scores when translating out of English. The symbol "≀" indicates where the average is out of sequence compared to the main Pearson average.

4 Segment-Level Results

We measure the quality of metrics' segment-level scores using Kendall's τ rank correlation coefficient. In this type of evaluation, a metric is expected to predict the result of the manual pairwise comparison of two systems. Note that the golden truth is obtained from a compact annotation of five systems at once, while an experiment with text-to-speech evaluation techniques by Vazquez-Alvarez and Huckvale (2002) suggest that a genuine pairwise comparison is likely to lead to more stable results.

The basic formula for Kendall's τ is:

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

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 changed each year of WMT metric tasks. Here we adopt the version from WMT14. 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 $h, m \in \{<, =, >\}$[4] and a metric, we compute the Kendall's τ for the metric the following way:

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 Kendall's τ, 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 of Kendall's τ, we simply sum the sizes of all the sets $S_{h,m}$ except those where $C_{h,m} = \text{X}$. To define it formally:

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

To summarize, the WMT14 matrix specifies to:

- exclude all human ties,
- count metric's ties only for the denominator of Kendall τ (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 $Concordant$.

You can find the system-level correlations for translations into English in Table 4 and for translations out of English in Table 5. Again, the upper part of each table contains metrics participating in all language pairs and it is sorted by average τ across translation directions. The lower part contains metrics limited to a subset of the language pairs, so the average cannot be directly compared with other metrics any more.

4.1 Segment-Level Discussion

As usual, segment-level correlations are significantly lower than system-level ones. The highest correlation is reached by DPMFCOMB on Czech-to-English: .495 of Kendall's τ. The correlations reach on average .447 into English and .400 out of English.

DPMFCOMB is the clear winner into English, followed by BEER_TREEPEL, both of which consider syntactic structure of the sentence, combined with several other independent features or metrics.

RATATOUILLE, also a combined metric, is the best option for evaluation to and from French.

Metrics considering character-level n-grams (BEER and CHRF3) are particularly good for

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

Direction	**fr-en**	**fi-en**	**de-en**	**cs-en**	**ru-en**	**Average**
Extracted-pairs	29770	31577	40535	85877	44539	
DPMFCOMB	$.395 \pm .012$	$\mathbf{.445} \pm .012$	$\mathbf{.482} \pm .009$	$\mathbf{.495} \pm .007$	$\mathbf{.418} \pm .013$	$\mathbf{.447} \pm .011$
BEER_TREEPEL	$.389 \pm .014$	$.438 \pm .010$	$.447 \pm .008$	$.471 \pm .007$	$.403 \pm .014$	$.429 \pm .011$
RATATOUILLE	$\mathbf{.398} \pm .010$	$.421 \pm .011$	$.441 \pm .010$	$.472 \pm .007$	$.393 \pm .013$	$.425 \pm .010$
UPF-COBALT	$.386 \pm .012$	$.437 \pm .013$	$.427 \pm .011$	$.457 \pm .007$	$.402 \pm .013$	$.422 \pm .011$
BEER	$.393 \pm .012$	$.422 \pm .012$	$.438 \pm .010$	$.457 \pm .008$	$.396 \pm .014$	$.421 \pm .011$
CHRF	$.383 \pm .011$	$.417 \pm .012$	$.424 \pm .010$	$.446 \pm .008$	$.384 \pm .014$	$.411 \pm .011$
CHRF3	$.383 \pm .013$	$.397 \pm .011$	$.421 \pm .010$	$.449 \pm .008$	$.386 \pm .013$	$.407 \pm .011$
METEOR-WSD	$.375 \pm .012$	$.406 \pm .010$	$.420 \pm .011$	$.438 \pm .008$	$.387 \pm .012$	$.405 \pm .010$
DPMF	$.368 \pm .012$	$.411 \pm .011$	$.418 \pm .011$	$.436 \pm .008$	$.378 \pm .011$	$.402 \pm .011$
LEBLEU-OPTIMIZED	$.376 \pm .013$	$.391 \pm .010$	$.399 \pm .010$	$.438 \pm .008$	$.374 \pm .012$	$.396 \pm .011$
LEBLEU-DEFAULT	$.373 \pm .013$	$.383 \pm .011$	$.402 \pm .009$	$.436 \pm .007$	$.376 \pm .011$	$.394 \pm .010$
VERTA-EQ	$.388 \pm .012$	$.369 \pm .013$	$.410 \pm .011$	$.447 \pm .007$	$.346 \pm .013$	$.392 \pm .011$
VERTA-70ADEQ30FLU	$.374 \pm .012$	$.365 \pm .014$	$.418 \pm .011$	$.438 \pm .007$	$.344 \pm .013$	$.388 \pm .011$
VERTA-W	$.383 \pm .010$	$.344 \pm .014$	$.416 \pm .010$	$.445 \pm .007$	$.345 \pm .013$	$.387 \pm .011$
DREEM	$.362 \pm .012$	$.340 \pm .010$	$.368 \pm .011$	$.423 \pm .007$	$.348 \pm .013$	$.368 \pm .011$
UOW-LSTM	$.332 \pm .011$	$.376 \pm .012$	$.375 \pm .011$	$.385 \pm .008$	$.356 \pm .010$	$.365 \pm .011$
SENTBLEU	$.358 \pm .013$	$.308 \pm .012$	$.360 \pm .011$	$.391 \pm .006$	$.329 \pm .011$	$.349 \pm .011$
TOTAL-BS	$.332 \pm .013$	$.319 \pm .013$	$.333 \pm .010$	$.381 \pm .007$	$.321 \pm .011$	$.337 \pm .011$
USAAR-ZWICKEL-METEOR	n/a	$.406 \pm .011$	$.422 \pm .011$	$.439 \pm .008$	$.386 \pm .012$	$.413 \pm .011$
USAAR-ZWICKEL-COMET	n/a	$.021 \pm .013$	$.050 \pm .010$	$.072 \pm .009$	$.084 \pm .010$	$.057 \pm .011$
USAAR-ZWICKEL-COSINE2METEOR	n/a	$.001 \pm .013$	$-.011 \pm .010$	$.020 \pm .009$	$.041 \pm .010$	$.013 \pm .011$
USAAR-ZWICKEL-COSINE	n/a	$-.035 \pm .013$	$-.019 \pm .010$	$.090 \pm .008$	$.014 \pm .013$	$.012 \pm .011$

Table 4: Segment-level Kendall's τ correlations of automatic evaluation metrics and the official WMT human judgements when translating into English.

Direction	**en-fr**	**en-fi**	**en-de**	**en-cs**	**en-ru**	**Average**
Extracted-pairs	34512	32694	54447	136890	49302	
BEER	$.352 \pm .010$	**.380** $\pm .010$	$.393 \pm .010$	$.435 \pm .006$	**.439** $\pm .010$	**.400** $\pm .009$
CHRF3	$.335 \pm .013$	$.373 \pm .012$	$.398 \pm .008$	**.446** $\pm .005$	$.420 \pm .010$	$.395 \pm .010$
RATATOUILLE	**.366** $\pm .013$	$.318 \pm .011$	$.381 \pm .008$	$.429 \pm .006$	$.436 \pm .010$	$.386 \pm .010$
LEBLEU-OPTIMIZED	$.347 \pm .009$	$.368 \pm .010$	**.399** $\pm .008$	$.410 \pm .006$	$.404 \pm .011$	$.386 \pm .009$
CHRF	$.342 \pm .012$	$.359 \pm .010$	$.372 \pm .010$	$.444 \pm .005$	$.410 \pm .011$	$.385 \pm .010$
LEBLEU-DEFAULT	$.345 \pm .010$	$.368 \pm .010$	$.398 \pm .009$	$.406 \pm .006$	$.404 \pm .012$	$.384 \pm .009$
METEOR-WSD	$.342 \pm .012$	$.286 \pm .010$	$.344 \pm .007$	$.390 \pm .006$	$.399 \pm .010$	$.352 \pm .009$
DREEM	$.338 \pm .012$	$.280 \pm .011$	$.317 \pm .010$	$.395 \pm .006$	$.366 \pm .010$	$.339 \pm .010$
SENTBLEU	$.318 \pm .011$	$.227 \pm .011$	$.294 \pm .009$	$.360 \pm .005$	$.347 \pm .010$	$.309 \pm .009$
TOTAL-BS	$.297 \pm .011$	$.223 \pm .009$	$.278 \pm .009$	$.345 \pm .005$	$.356 \pm .011$	$.300 \pm .009$
DPMF	$.335 \pm .012$	n/a	$.350 \pm .009$	n/a	n/a	$.343 \pm .010$
USAAR-ZWICKEL-METEOR	n/a	n/a	$.342 \pm .008$	n/a	n/a	$.342 \pm .008$
USAAR-ZWICKEL-COMET	n/a	n/a	$.056 \pm .019$	n/a	n/a	$.056 \pm .009$
USAAR-ZWICKEL-COSINE	n/a	n/a	$-.007 \pm .010$	n/a	n/a	$-.007 \pm .010$
USAAR-ZWICKEL-COSINE2METEOR	n/a	n/a	$-.027 \pm .019$	n/a	n/a	$-.027 \pm .009$

Table 5: Segment-level Kendall's τ correlations of automatic evaluation metrics and the official WMT human judgements when translating out of English.

		2014	2015	Delta
BEER	Average en→*	.319±.011	.401±.009	0.082
	en-cs	.344±.009	.435±.006	0.091
	en-de	.268±.009	.396±.008	0.128
	en-fr	.292±.012	.352±.010	0.060
	en-ru	.440±.013	.440±.012	0.000
	Average *→en	.362±.013	.423±.010	0.061
	cs-en	.284±.016	.457±.008	0.173
	de-en	.337±.014	.438±.010	0.101
	fr-en	.417±.013	.393±.012	-0.024
	ru-en	.333±.011	.406±.009	0.073
SENTBLEU	Average en→*	.269±.011	.310±.009	0.041
	en-cs	.290±.009	.360±.005	0.070
	en-de	.191±.009	.296±.010	0.105
	en-fr	.256±.012	.318±.011	0.062
	en-ru	.381±.013	.347±.010	-0.034
	Average *→en	.285±.013	.351±.011	0.066
	cs-en	.213±.016	.391±.006	0.178
	de-en	.271±.014	.360±.011	0.089
	fr-en	.378±.013	.358±.013	-0.020
	ru-en	.263±.011	.340±.012	0.077
Average				0.07±0.06

Table 6: Kendall's τ scores for two metrics across years.

evaluation out of English and their margin seems to the highest for English-to-Finnish, up to .06 points.

Only two segment-level metrics took part in 2014 and 2015, BEER in a slightly improved implementation (with some small effect on the scores) and SENTBLEU in exactly the same implementation. Table 6 documents that this year, the scores are on average slightly higher. The main reason lies probably in the test set, which may be somewhat easier this year. French is different, the correlations decreased somewhat this year, which can be easily explained by the domain change: news in 2014 and discussions in 2015. The increase should not be caused by the redundancy cleanup of WMT manual rankings, see Bojar et al. (2015), since the collapsed systems get a tie after expanding and our implementation ignores all tied manual comparisons.

5 Conclusion

In this paper, we summarized the results of the WMT15 Metrics Shared Task, which assesses the quality of various automatic machine translation metrics. As in previous years, human judgements collected in WMT15 serve as the golden truth and we check how well the metrics predict the judgements at the level of individual sentences as well as at the level of the whole test set (system-level).

Across the two types of evaluation and the 10 language pairs, we saw great performance of trained and combined metrics (DPMFCOMB, BEER, RATATOUILLE and others). Neural networks for continuous word and sentence representations have also shown their generalization power, with an interesting discrepancy in system- vs. segment-level performance of UOW-LSTM and to a smaller degree of DREEM.

We value high the metric CHRF or CHRF3 for its extreme simplicity and very good performance at both system and segment level and especially out of English. We are curious to see if CHRF3 has the potential of becoming "the BLEU for the next five years". It would be very interesting to test its usability in system tuning. It is known that in tuning, metrics putting too much attention to recall can be easily tricked, but perhaps a careful setting of CHRF's β will be sufficient.

The WMT Metrics Task again attracted a good number of participants and the majority of submitted metrics are actually new ones. This is good news, indicating that MT metrics are an active field of research. Most, if not all metrics come with the source code, so it should be relatively easy to use them in own experiments. Still, we would expect much wider adoption of the metrics, if they made it for example to the standard Moses scorer or at least to the Asyia toolkit.

Acknowledgments

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreements n° 645452 (QT21) and n° 644402 (HimL). The work on this project was also supported by the Dutch organisation for scientific research STW grant nr. 12271.

References

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

Ondřej Bojar, Matouš Macháček, Aleš Tamchyna, and Daniel Zeman. 2013. Scratching the Surface of Possible Translations. In *Proc. of TSD 2013*, Lecture Notes in Artificial Intelligence, Berlin / Heidelberg. Západočeská univerzita v Plzni, Springer Verlag.

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.

Boxing Chen, Hongyu Guo, and Roland Kuhn. 2015. Multi-level Evaluation for Machine Translation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, Lisboa, Portugal, September. Association for Computational Linguistics.

Elisabet Comelles and Jordi Atserias. 2015. VERTa: a Linguistically-motivated Metric at the WMT15 Metrics Task. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, Lisboa, Portugal, 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.

Marina Fomicheva, Núria Bel, Iria da Cunha, and Anton Malinovskiy. 2015. UPF-Cobalt Submission to WMT15 Metrics Task. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, Lisboa, Portugal, September. Association for Computational Linguistics.

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 Orasan, and Josef van Genabith. 2015b. ReVal: A Simple and Effective Machine Translation Evaluation Metric Based on Recurrent Neural Networks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '15, Lisbon, Portugal.

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. 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.

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

Benjamin Marie and Marianna Apidianaki. 2015. Alignment-based sense selection in METEOR and the RATATOUILLE recipe. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, Lisboa, Portugal, September. 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, Stroudsburg, PA, USA. Association for Computational Linguistics.

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

Raphael Rubino, Tommi Pirinen, Miquel Esplà-Gomis, Nikola Ljubešić, Sergio Ortiz Rojas, Vassilis Papavassiliou, Prokopis Prokopidis, and Antonio Toral. 2015. Abu-MaTran at WMT 2015 Translation Task: Morphological Segmentation and Web Crawling. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, Lisboa, Portugal, 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.

Miloš Stanojević, Amir Kamran, and Ondřej Bojar. 2015. Results of the WMT15 Tuning 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*.

Mihaela Vela and Liling Tan. 2015. Predicting Machine Translation Adequacy with Document Embeddings. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, Lisboa, Portugal, September. Association for Computational Linguistics.

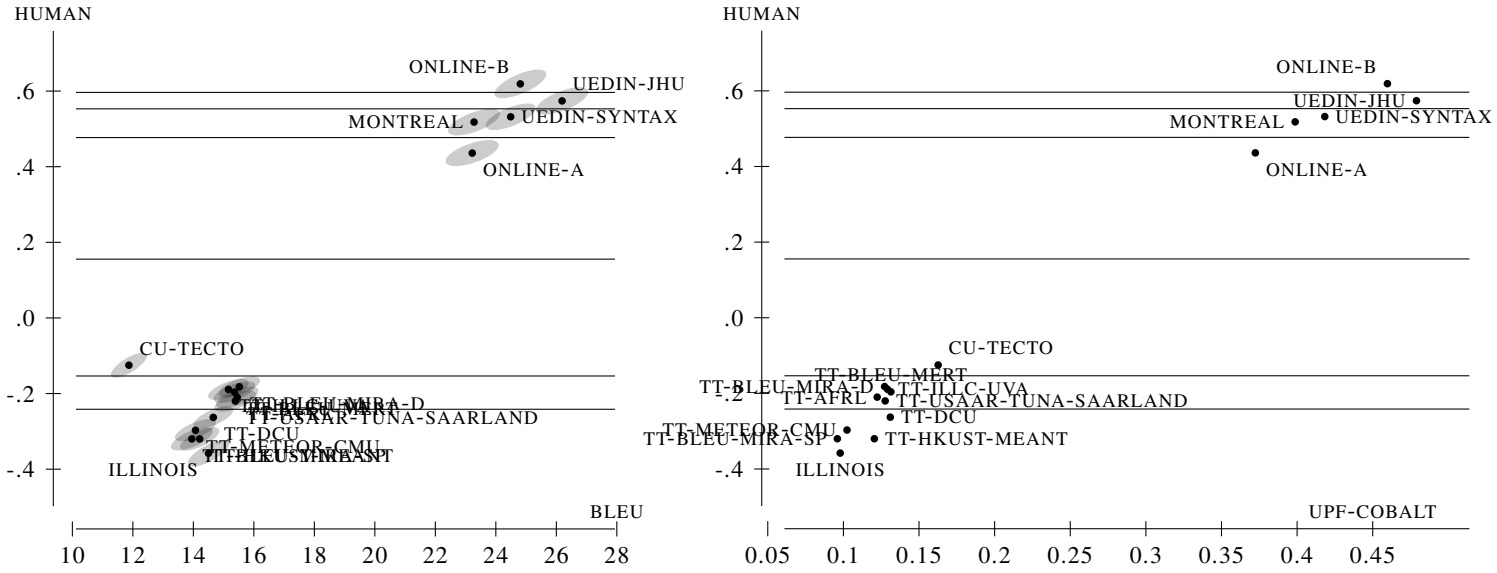
Sami Virpioja and Stig-Arne Grönroos. 2015. LeBLEU: N-gram-based Translation Evaluation Score for Morphologically Complex Languages. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, Lisboa, Portugal, September. Association for Computational Linguistics.

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.

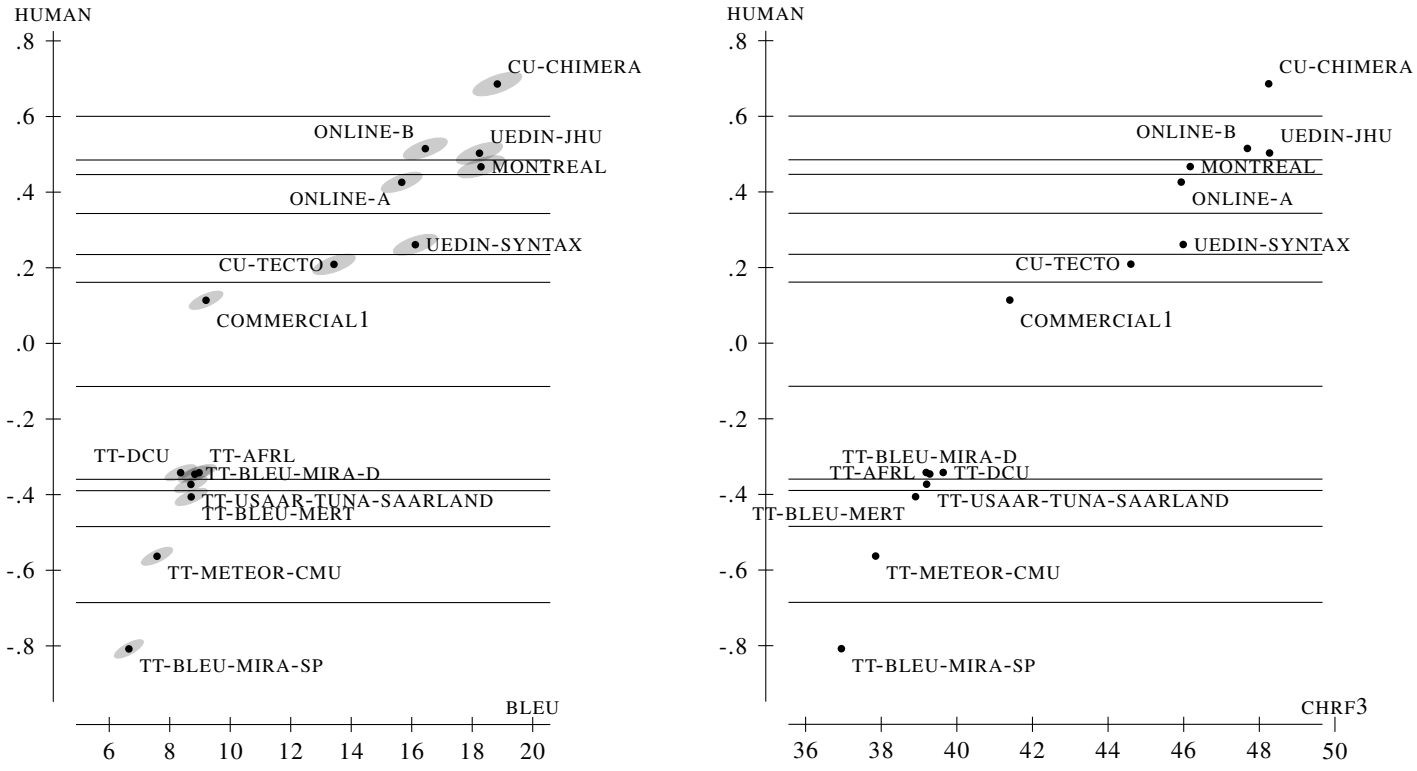
A System-Level Correlation Plots

The following figures plot the system-level results of BLEU (left-hand plots) and the best performing metric for the given language pair (right-hand plots) against manual score. See the discussion in Section 3.2.

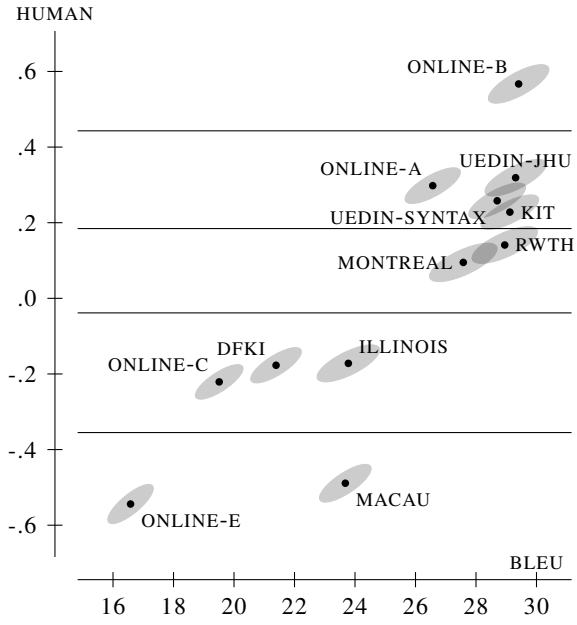
Czech-English

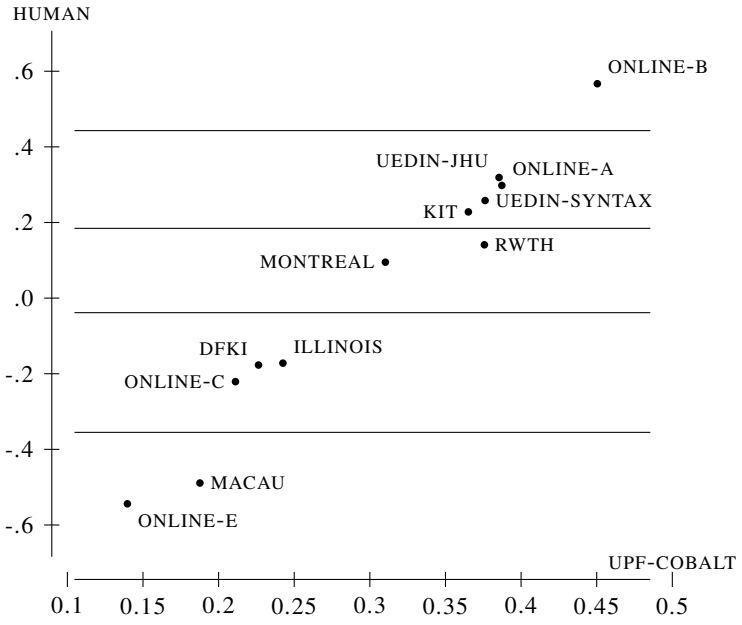


English-Czech

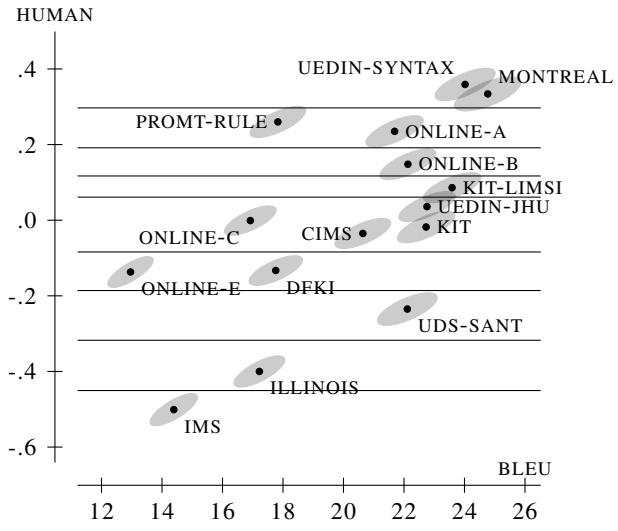


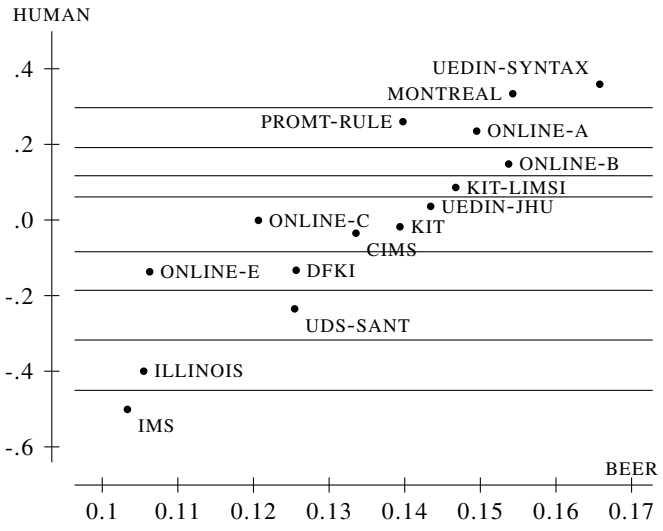
German-English



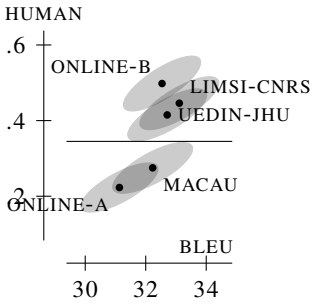


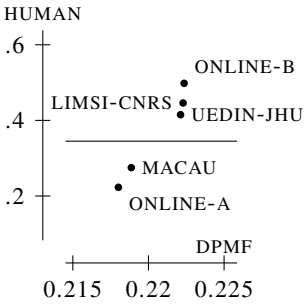
English-German



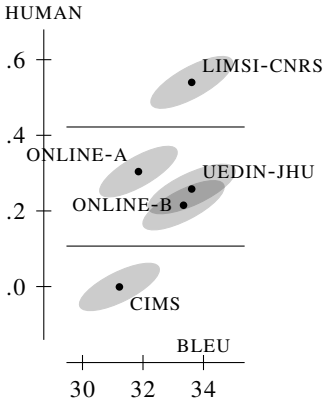


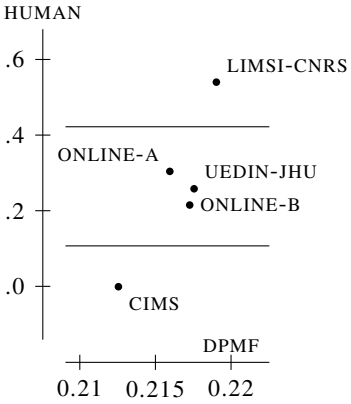
French-English



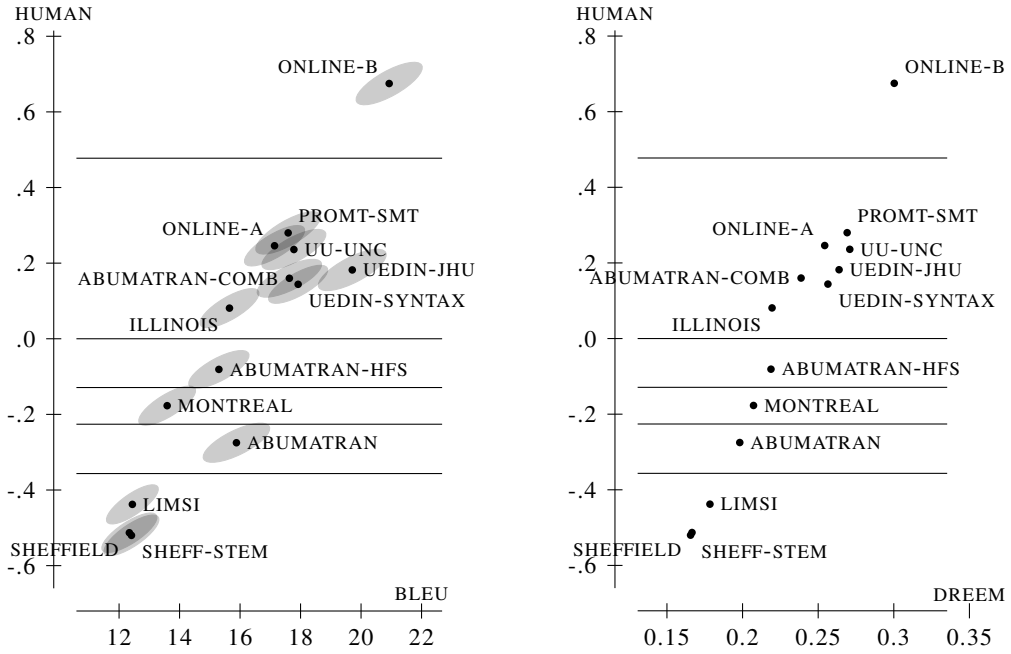


English-French

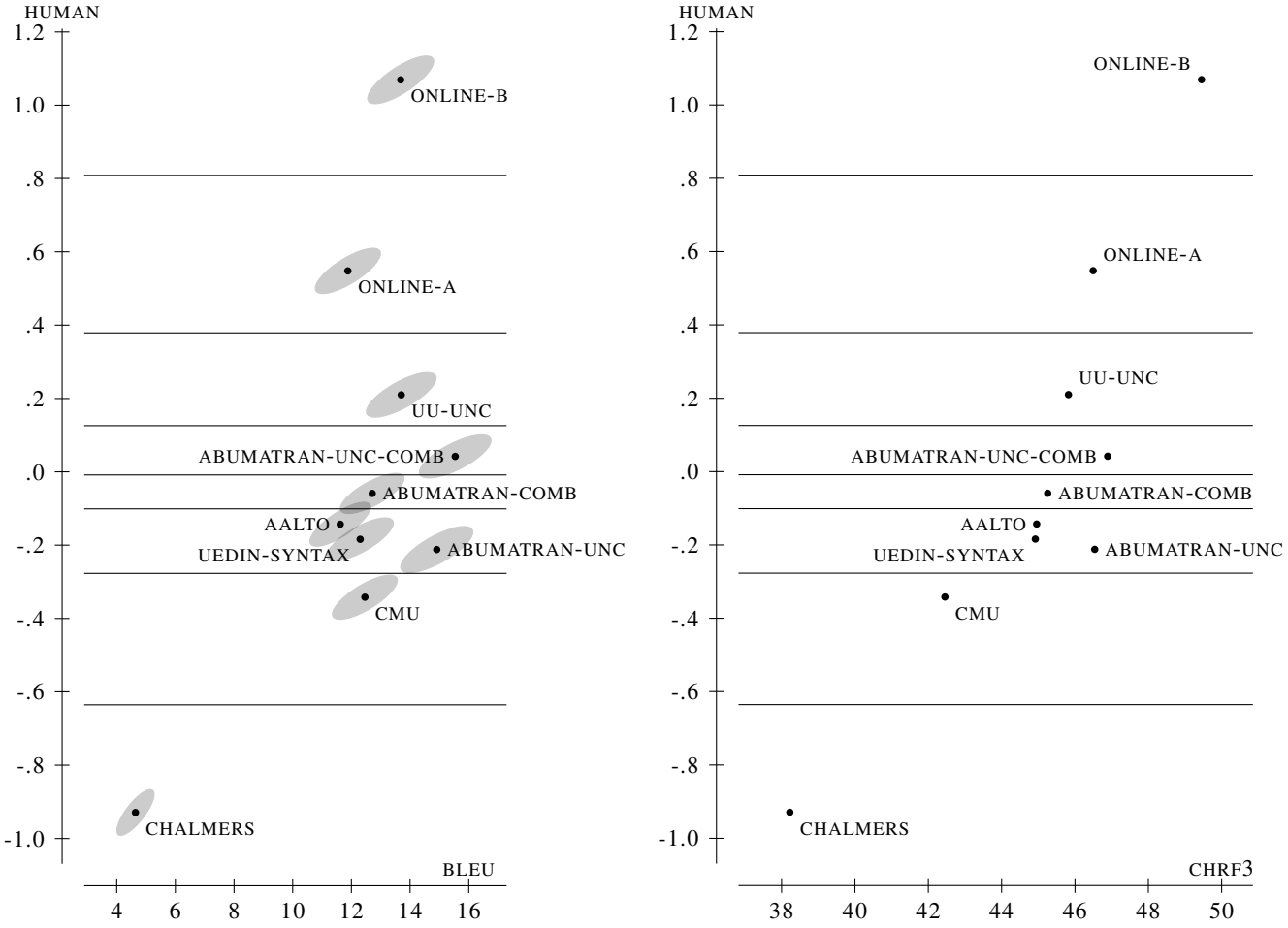




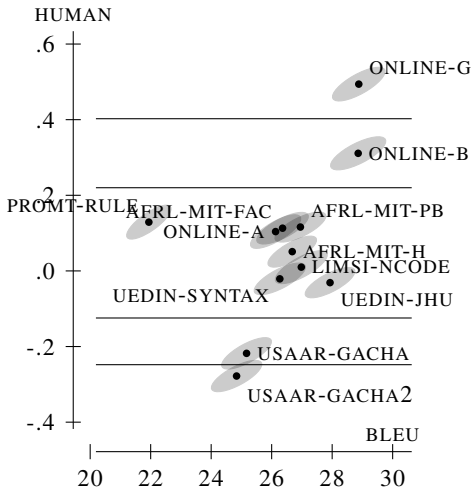
Finnish-English

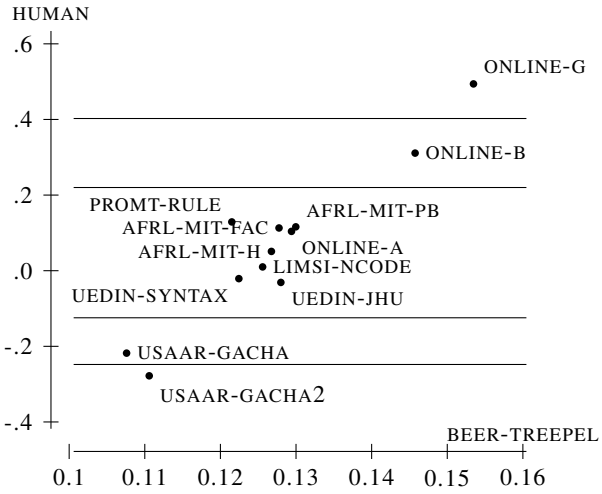


English-Finnish



Russian-English





English-Russian