Results of the WMT14 Metrics Shared Task

Matouš Macháček and **Ondřej Bojar**
Charles University in Prague, Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
machacekmatous@gmail.com and bojar@ufal.mff.cuni.cz

Abstract

This paper presents the results of the WMT14 Metrics Shared Task. We asked participants of this task to score the outputs of the MT systems involved in WMT14 Shared Translation Task. We collected scores of 23 metrics from 12 research groups. In addition to that we computed scores of 6 standard metrics (BLEU, 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 WMT14 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 Bojar et al. (2014).

In this task, we asked metrics developers to score the outputs of WMT14 Shared Translation Task (Bojar et al., 2014). We have collected the computed metrics' scores and use them to evaluate quality of the metrics.

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. Segment level correlation with a detailed discussion and a slight change in the calculation compared to the previous year is reported in Section 4.

[1] http://www.statmt.org/wmt13

2 Data

We used the translations of MT systems involved in WMT14 Shared Translation Task together with reference translations as the test set for the Metrics Task. This dataset consists of 110 systems' outputs and 10 reference translations in 10 translation directions (English from and into Czech, French, German, Hindi and Russian). For most of the translation directions each system's output and the reference translation contain 3003 sentences. For more details please see the WMT14 overview paper (Bojar et al., 2014).

2.1 Manual MT Quality Judgements

During the WMT14 Translation Task, a large scale manual annotation was conducted to compare the systems. We used these collected human judgements for the evalution 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 the outputs of five systems to which he or she was supposed to assign ranks. Ties were allowed.

These collected rank labels for each five-tuple of systems were then interpreted as 10 pairwise comparisons of systems and used to assign each system a score that reflects how high that system was usually ranked by the annotators. Please see the WMT14 overview paper for details on how this score is computed. You can also find inter- and intra-annotator agreement estimates there.

2.2 Participants of the Metrics Shared Task

Table 1 lists the participants of WMT14 Shared Metrics Task, along with their metrics. We have

Proceedings of the Ninth Workshop on Statistical Machine Translation, pages 293–301,
Baltimore, Maryland USA, June 26–27, 2014.

Metric	Participant
APAC	Hokkai-Gakuen University (Echizen'ya, 2014)
BEER	ILLC – University of Amsterdam (Stanojevic and Sima'an, 2014)
RED-*	Dublin City University (Wu and Yu, 2014)
DISCOTK-*	Qatar Computing Research Institute (Guzman et al., 2014)
ELEXR	University of Tehran (Mahmoudi et al., 2013)
LAYERED	Indian Institute of Technology, Bombay (Gautam and Bhattacharyya, 2014)
METEOR	Carnegie Mellon University (Denkowski and Lavie, 2014)
AMBER, BLEU-NRC	National Research Council of Canada (Chen and Cherry, 2014)
PARMESAN	Charles University in Prague (Barančíková, 2014)
TBLEU	Charles University in Prague (Libovický and Pecina, 2014)
UPC-IPA, UPC-STOUT	Technical University of Catalunya (Gonzàlez et al., 2014)
VERTA-W, VERTA-EQ	University of Barcelona (Comelles and Atserias, 2014)

Table 1: Participants of WMT14 Metrics Shared Task

collected 23 metrics from a total of 12 research groups.

In addition to that we have computed the following two groups of standard metrics as baselines:

- **Mteval.** The metrics BLEU (Papineni et al., 2002) and NIST (Doddington, 2002) were computed using the script `mteval-v13a.pl`[2] 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.

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

3 System-Level Metric Analysis

While the Spearman's ρ correlation coefficient was used as the main measure of system-level metrics' quality in the past, we have decided to use Pearson correlation coefficient as the main measure this year. At the end of this section we give reasons for this change.

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 metrics are sorted by average Pearson correlation coefficient across translation directions. 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 %).

As in previous years, a lot of metrics outperformed BLEU in system level correlation. In into-English directions, metric DISCOTK-PARTY-TUNED has the highest correlation in two language directions and it is also the best correlated metric on average according to both Pearson and Spearman's coefficients. The second best correlated metric on average (according to Pearson) is LAYERED which is also the single best metric in Hindi-to-English direction. Metrics REDSYS and REDSYSSENT are quite unstable, they win in French-to-English and Czech-to-English directions respectively but they perform very poorly in

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

other directions.

Except METEOR, none of the participants took part in the last year metrics task. We can therefore compare current and last year results only for METEOR and baseline metrics. METEOR, the last year winner, performs generally well in some directions but it horribly suffers when evaluating translations from non-Latin script (Russian and especially Hindi). For the baseline metrics the results are quite similar across the years. In both years BLEU performs best among baseline metrics, closely followed by CDER. NIST is in the middle of the list in both years. The remaining baseline metrics TER, WER and PER perform much worse.

The results into German are markedly lower and have broader confidence intervals than the results in other directions. This could be explained by a very high number (18) of participating systems of similar quality. Both human judgements and automatic metrics are negatively affected by these circumstances. To preserve the reliability of overall metrics' performance across languages, we decided to exclude English-to-German direction from the average Pearson and Spearman's correlation coefficients.

In other out-of-English directions, the best correlated metric on average according to Pearson coefficient is NIST, even though it does not win in any single direction. CDER is the second best according to Pearson and the best metric according to Spearman's. Again it does not win in any single direction. The metrics PER and WER are quite unstable. Each of them wins in two directions but performs very badly in others.

Compared to the last year results, the order of metrics participating in both years is quite similar: NIST and CDER performed very well both years, followed by BLEU. The metrics TER and WER are again at the end of the list. An interesting change is that PER perform much better this year.

3.1 Reasons for Pearson correlation coefficient

In the translation task, there are often similar systems with human scores very close to each other. It can therefore easily happen that even a good metric compares two similar systems differently from humans. We believe that the penalty incurred by the metric for such a swap should somehow reflect that the systems were hard to separate.

Since the Spearman's ρ converts both human and metric scores to ranks and therefore disregards the absolute differences in the scores, it does exactly what we feel is not fair. The Pearson correlation coefficient does not suffer from this problem. We are aware of the fact that Pearson correlation coefficient also reflects whether the relation between manual and automatic scores is linear (as opposed to e.g. quadratic). We don't think this would be negatively affecting any of the metrics since overall, the systems are of a comparable quality and the metrics are likely to behave linearly in this small range of scores.

Moreover, the general agreement to adopt Pearson instead of Spearman's correlation coefficient was already apparent during the WMT12 workshop. This change just did not get through for WMT13.

4 Segment-Level Metric Analysis

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) suggests that a genuine pairwise comparison is likely to lead to more stable results.

In the past, slightly different variations of Kendall's τ computation were used in the Metrics Tasks. Also some of the participants have noticed a problem with ties in the WMT13 method. Therefore, we discuss several possible variants in detail in this paper.

4.1 Notation for Kendall's τ computation

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. In the original Kendall's τ, comparisons with human or metric ties are considered neither concordant nor discordant. However in the past, Metrics

Correlation coefficient	**Pearson Correlation Coefficient**						**Spearman's**
Direction	**fr-en**	**de-en**	**hi-en**	**cs-en**	**ru-en**	**Average**	**Average**
Considered Systems	8	13	9	5	13		
DISCOTK-PARTY-TUNED	$.977 \pm .009$	$\mathbf{.943 \pm .020}$	$.956 \pm .007$	$.975 \pm .031$	$\mathbf{.870 \pm .022}$	$\mathbf{.944 \pm .018}$	$\mathbf{.912 \pm .043}$
LAYERED	$.973 \pm .009$	$.893 \pm .026$	$\mathbf{.976 \pm .006}$	$.941 \pm .045$	$.854 \pm .023$	$.927 \pm .022$	$.894 \pm .047$
DISCOTK-PARTY	$.970 \pm .010$	$.921 \pm .024$	$.862 \pm .015$	$.983 \pm .025$	$.856 \pm .023$	$.918 \pm .019$	$.856 \pm .046$
UPC-STOUT	$.968 \pm .010$	$.915 \pm .025$	$.898 \pm .013$	$.948 \pm .040$	$.837 \pm .024$	$.913 \pm .022$	≀ $.901 \pm .045$
VERTA-W	$.959 \pm .011$	$.867 \pm .029$	$.920 \pm .011$	$.934 \pm .050$	$.848 \pm .024$	$.906 \pm .025$	$.868 \pm .045$
VERTA-EQ	$.959 \pm .011$	$.854 \pm .031$	$.927 \pm .010$	$.938 \pm .048$	$.842 \pm .024$	$.904 \pm .025$	$.857 \pm .046$
TBLEU	$.952 \pm .012$	$.832 \pm .034$	$.954 \pm .007$	$.957 \pm .040$	$.803 \pm .027$	$.900 \pm .024$	$.841 \pm .056$
BLEU_NRC	$.953 \pm .012$	$.823 \pm .035$	$.959 \pm .007$	$.946 \pm .044$	$.787 \pm .028$	$.894 \pm .025$	≀ $.855 \pm .056$
BLEU	$.952 \pm .012$	$.832 \pm .034$	$.956 \pm .007$	$.909 \pm .054$	$.789 \pm .027$	$.888 \pm .027$	$.833 \pm .058$
UPC-IPA	$.966 \pm .010$	$.895 \pm .027$	$.914 \pm .010$	$.824 \pm .073$	$.812 \pm .026$	$.882 \pm .029$	≀ $.858 \pm .044$
CDER	$.954 \pm .012$	$.823 \pm .034$	$.826 \pm .016$	$.965 \pm .035$	$.802 \pm .027$	$.874 \pm .025$	$.807 \pm .050$
APAC	$.963 \pm .010$	$.817 \pm .034$	$.790 \pm .016$	$.982 \pm .026$	$.816 \pm .026$	$.874 \pm .022$	$.807 \pm .049$
REDSYS	$\mathbf{.981 \pm .008}$	$.898 \pm .026$	$.676 \pm .022$	$.989 \pm .021$	$.814 \pm .026$	$.872 \pm .021$	$.786 \pm .047$
REDSYSSENT	$.980 \pm .008$	$.910 \pm .024$	$.644 \pm .023$	$\mathbf{.993 \pm .018}$	$.807 \pm .027$	$.867 \pm .020$	$.771 \pm .043$
NIST	$.955 \pm .011$	$.811 \pm .035$	$.784 \pm .016$	$.983 \pm .025$	$.800 \pm .027$	$.867 \pm .023$	≀ $.824 \pm .055$
DISCOTK-LIGHT	$.965 \pm .011$	$.935 \pm .022$	$.557 \pm .025$	$.954 \pm .038$	$.791 \pm .027$	$.840 \pm .024$	$.774 \pm .046$
METEOR	$.975 \pm .009$	$.927 \pm .022$	$.457 \pm .027$	$.980 \pm .029$	$.805 \pm .026$	$.829 \pm .023$	≀ $.788 \pm .046$
TER	$.952 \pm .012$	$.775 \pm .038$	$.618 \pm .021$	$.976 \pm .031$	$.809 \pm .027$	$.826 \pm .026$	$.746 \pm .057$
WER	$.952 \pm .012$	$.762 \pm .038$	$.610 \pm .021$	$.974 \pm .033$	$.809 \pm .027$	$.821 \pm .026$	$.736 \pm .058$
AMBER	$.948 \pm .012$	$.910 \pm .026$	$.506 \pm .026$	$.744 \pm .095$	$.797 \pm .027$	$.781 \pm .037$	$.728 \pm .051$
PER	$.946 \pm .013$	$.867 \pm .031$	$.411 \pm .025$	$.883 \pm .063$	$.799 \pm .028$	$.781 \pm .032$	$.698 \pm .047$
ELEXR	$.971 \pm .009$	$.857 \pm .031$	$.535 \pm .026$	$.945 \pm .044$	$-.404 \pm .045$	$.581 \pm .031$	$.652 \pm .046$

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

Correlation coefficient	**Pearson Correlation Coefficient**						**Spearman's**
Direction	**en-fr**	**en-hi**	**en-cs**	**en-ru**	**Average**	**en-de**	**Average**
Considered Systems	13	12	10	9		18	(excl. en-de)
NIST	.941 ± .022	.981 ± .006	.985 ± .006	.927 ± .012	**.959 ± .012**	.200 ± .046	**.850 ± .030**
CDER	.949 ± .020	.949 ± .010	.982 ± .006	.938 ± .011	.955 ± .012	.278 ± .045	.840 ± .036
AMBER	.928 ± .023	**.990 ± .004**	.972 ± .008	.926 ± .012	.954 ± .012	.241 ± .045	.817 ± .041
METEOR	.941 ± .021	.975 ± .007	.976 ± .007	.923 ± .013	.954 ± .012	.263 ± .045	.806 ± .039
BLEU	.937 ± .022	.973 ± .007	.976 ± .007	.915 ± .013	.950 ± .012	.216 ± .046	≀ .809 ± .036
PER	.936 ± .023	.931 ± .011	**.988 ± .005**	**.941 ± .011**	.949 ± .013	.190 ± .047	≀ .823 ± .037
APAC	.950 ± .020	.940 ± .011	.973 ± .008	.929 ± .012	.948 ± .013	.346 ± .044	.799 ± .041
TBLEU	.932 ± .023	.968 ± .008	.973 ± .008	.912 ± .013	.946 ± .013	.239 ± .046	≀ .805 ± .039
BLEU_NRC	.933 ± .022	.971 ± .007	.974 ± .008	.901 ± .014	.945 ± .013	.205 ± .046	≀ .809 ± .039
ELEXR	.885 ± .029	.962 ± .009	.979 ± .007	.938 ± .011	.941 ± .014	.260 ± .044	.768 ± .036
TER	.954 ± .019	.829 ± .017	.978 ± .007	.931 ± .012	.923 ± .014	.324 ± .045	.745 ± .035
WER	**.960 ± .018**	.516 ± .026	.976 ± .007	.932 ± .011	.846 ± .016	**.357 ± .045**	.696 ± .037
PARMESAN	n/a	n/a	.962 ± .009	n/a	.962 ± .009	n/a	.915 ± .048
UPC-IPA	.940 ± .021	n/a	.969 ± .008	.921 ± .013	.943 ± .014	.285 ± .045	.785 ± .050
REDSYSSENT	.941 ± .021	n/a	n/a	n/a	.941 ± .021	.208 ± .045	≀ .962 ± .038
REDSYS	.940 ± .021	n/a	n/a	n/a	.940 ± .021	.208 ± .045	.962 ± .038
UPC-STOUT	.940 ± .021	n/a	.938 ± .011	.919 ± .013	.933 ± .015	.301 ± .044	.713 ± .040

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 Spearman's ρ average is out of sequence compared to the main Pearson average.

Tasks (Callison-Burch et al. (2012) and earlier), comparisons with human ties were considered as discordant.

To easily see which pairs are counted as concordant and which as discordant, we have developed the following tabular notation. This is for example the WMT12 method:

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

Given such a matrix $C_{h,m}$ where $h, m \in \{<, =, >\}$[3] and a metric we compute the Kendall's τ 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:

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

4.2 Discussion on Kendall's τ computation

In 2013, we thought that metric ties should not be penalized and we decided to excluded them like the human ties. We will denote this method as WMT13:

		Metric		
WMT13		$<$	$=$	$>$
Human	$<$	1	X	-1
	$=$	X	X	X
	$>$	-1	X	1

It turned out, however, that it was not a good idea: metrics could game the scoring by avoiding hard cases and assigning lots of ties. A natural solution is to count the metrics ties also in denominator to avoid the problem. We will denote this variant as WMT14:

		Metric		
WMT14		$<$	$=$	$>$
Human	$<$	1	0	-1
	$=$	X	X	X
	$>$	-1	0	1

The WMT14 variant does not allow for gaming the scoring like the WMT13 variant does. Compared to WMT12 method, WMT14 does not penalize ties.

We were also considering to get human ties involved. The most natural variant would be the following variant denoted as HTIES:

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

Unfortunately this method allows for gaming the scoring as well. The least risky choice for metrics in hard cases would be to assign a tie because it cannot worsen the Kendall's τ and there is quite a high chance that the human rank is also a tie. Metrics could be therefore tuned to predict ties often but such metrics are not very useful. For example, the simplistic metric which assigns the same score to all candidates (and therefore all pairs would be tied by the metric) would get the score equal to the proportion of ties in all human comparisons. It would become one of the best performing metrics in WMT13 even though it is not informative at all.

We have decided to use WMT14 variant as the main evaluation measure this year, however, we are also reporting average scores computed by other variants.

4.3 Kendall's τ results

The final Kendall's τ results are shown in Table 4 for directions into English and in Table 5 for directions out of English. Each row in the tables contains correlations of a metric in given directions. The metrics are sorted by average correlation across translation directions. The highest correlation in each column is in bold. The tables also contain average Kendall's τ computed by other variants including the variant WMT13 used last year. Metrics which did not compute scores in all directions are at the bottom of the tables. The

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

Direction	fr-en	de-en	hi-en	cs-en	ru-en	Avg	Averages of other variants of Kendall's τ		
Extracted-pairs	26090	25260	20900	21130	34460		WMT12	WMT13	HTIES
DISCOTK-PARTY-TUNED	**.433 ± .012**	**.380 ± .013**	.434 ± .013	**.328 ± .015**	**.355 ± .011**	**.386 ± .013**	**.386 ± .013**	.386 ± .013	.306 ± .010
BEER	.417 ± .013	.337 ± .014	**.438 ± .013**	.284 ± .016	.333 ± .011	.362 ± .013	.358 ± .013	.363 ± .013	≀**.318 ± .011**
REDCOMBSENT	.406 ± .012	.338 ± .014	.417 ± .013	.284 ± .015	.336 ± .011	.356 ± .013	.346 ± .013	.360 ± .013	.317 ± .011
REDCOMBSYSSENT	.408 ± .012	.338 ± .014	.416 ± .013	.282 ± .014	.336 ± .011	.356 ± .013	.346 ± .013	.359 ± .013	.316 ± .010
METEOR	.406 ± .012	.334 ± .014	.420 ± .013	.282 ± .015	.329 ± .010	.354 ± .013	.341 ± .013	.359 ± .013	≀.317 ± .010
REDSYSSENT	.404 ± .012	.338 ± .014	.386 ± .014	.283 ± .015	.321 ± .010	.346 ± .013	.335 ± .013	.350 ± .013	.309 ± .010
REDSENT	.403 ± .012	.336 ± .014	.383 ± .014	.283 ± .015	.323 ± .011	.345 ± .013	.334 ± .013	.349 ± .013	.308 ± .010
UPC-IPA	.412 ± .012	.340 ± .014	.368 ± .014	.274 ± .015	.316 ± .011	.342 ± .013	≀.340 ± .014	.343 ± .014	.300 ± .011
UPC-STOUT	.403 ± .012	.345 ± .014	.352 ± .014	.275 ± .015	.317 ± .011	.338 ± .013	.336 ± .013	.339 ± .013	.294 ± .011
VERTA-W	.399 ± .013	.321 ± .015	.386 ± .014	.263 ± .015	.315 ± .011	.337 ± .014	.320 ± .014	≀.342 ± .014	≀.304 ± .011
VERTA-EQ	.407 ± .013	.315 ± .014	.384 ± .013	.263 ± .015	.312 ± .011	.336 ± .013	≀.323 ± .013	.341 ± .013	.302 ± .011
DISCOTK-PARTY	.395 ± .013	.334 ± .014	.362 ± .013	.264 ± .016	.305 ± .011	.332 ± .013	≀.332 ± .013	.332 ± .013	.263 ± .011
AMBER	.367 ± .013	.313 ± .014	.362 ± .013	.246 ± .016	.294 ± .011	.316 ± .013	.302 ± .013	.321 ± .014	≀.286 ± .011
BLEU_NRC	.382 ± .013	.272 ± .014	.322 ± .014	.226 ± .016	.269 ± .011	.294 ± .013	.267 ± .014	.303 ± .014	.271 ± .011
SENTBLEU	.378 ± .013	.271 ± .014	.300 ± .013	.213 ± .016	.263 ± .011	.285 ± .013	.258 ± .014	.293 ± .014	.264 ± .011
APAC	.364 ± .012	.271 ± .014	.288 ± .014	.198 ± .016	.276 ± .011	.279 ± .013	.243 ± .014	.290 ± .014	.261 ± .011
DISCOTK-LIGHT	.311 ± .014	.224 ± .015	.238 ± .013	.187 ± .016	.209 ± .011	.234 ± .014	.234 ± .014	.234 ± .014	.184 ± .011
DISCOTK-LIGHT-KOOL	.005 ± .001	.001 ± .000	.000 ± .000	.002 ± .001	.001 ± .000	.002 ± .001	−.996 ± .001	≀**.676 ± .256**	≀.211 ± .005

Table 4: Segment-level Kendall's τ correlations of automatic evaluation metrics and the official WMT human judgements when translating into English. The last three columns contain average Kendall's τ computed by other variants. The symbol "≀" indicates where the averages of other variants are out of sequence compared to the WMT14 variant.

Direction	en-fr	en-de	en-hi	en-cs	en-ru	Avg	Averages of other variants of Kendall's τ		
Extracted-pairs	33350	54660	28120	55900	28960		WMT12	WMT13	HTIES
BEER	.292 ± .012	**.268 ± .009**	.250 ± .013	**.344 ± .009**	**.440 ± .013**	**.319 ± .011**	**.314 ± .011**	**.320 ± .011**	.272 ± .009
METEOR	.280 ± .012	.238 ± .009	.264 ± .012	.318 ± .009	.427 ± .012	.306 ± .011	.283 ± .011	.313 ± .011	≀**.273 ± .008**
AMBER	.264 ± .012	.227 ± .009	**.286 ± .012**	.302 ± .009	.397 ± .013	.295 ± .011	.269 ± .011	.303 ± .011	.266 ± .009
BLEU_NRC	.261 ± .012	.202 ± .009	.234 ± .013	.297 ± .009	.391 ± .012	.277 ± .011	.235 ± .011	.289 ± .011	.256 ± .009
APAC	.253 ± .012	.210 ± .008	.203 ± .012	.292 ± .009	.388 ± .013	.269 ± .011	.217 ± .011	.285 ± .011	.252 ± .008
SENTBLEU	.256 ± .012	.191 ± .009	.227 ± .012	.290 ± .009	.381 ± .013	.269 ± .011	≀.232 ± .011	.280 ± .011	.246 ± .009
UPC-STOUT	.279 ± .011	.234 ± .008	n/a	.282 ± .009	.425 ± .013	.305 ± .011	.300 ± .010	.306 ± .011	.256 ± .008
UPC-IPA	.264 ± .012	.227 ± .009	n/a	.298 ± .009	.426 ± .013	.304 ± .011	.292 ± .011	≀.308 ± .011	≀.259 ± .008
REDSENT	**.293 ± .012**	.242 ± .009	n/a	n/a	n/a	.267 ± .010	.246 ± .010	.273 ± .011	.257 ± .008
REDCOMBSYSSENT	.291 ± .012	.244 ± .009	n/a	n/a	n/a	.267 ± .010	≀.249 ± .010	.272 ± .010	.256 ± .008
REDCOMBSENT	.290 ± .012	.242 ± .009	n/a	n/a	n/a	.266 ± .010	.248 ± .010	.271 ± .011	.256 ± .008
REDSYSSENT	.290 ± .012	.239 ± .008	n/a	n/a	n/a	.264 ± .010	.235 ± .010	≀.273 ± .010	≀.257 ± .008

Table 5: Segment-level Kendall's τ correlations of automatic evaluation metrics and the official WMT human judgements when translating out of English. The last three columns contain average Kendall's τ computed by other variants. The symbol "≀" indicates where the averages of other variants are out of sequence compared to the WMT14 variant.

possible values of τ range between -1 (a metric always predicted a different order than humans did) and 1 (a metric always predicted the same order as humans). Metrics with a higher τ are better.

We also computed empirical confidence intervals of Kendall's τ using bootstrap resampling. We varied the "golden truth" by sampling from human judgments. We have generated 1000 new sets and report the average of the upper and lower 2.5 % empirical bound, which corresponds to the 95 % confidence interval.

In directions into English (Table 4), the strongest correlated segment-level metric on average is DISCOTK-PARTY-TUNED followed by BEER. Unlike the system level correlation, the results are much more stable here. DISCOTK-PARTY-TUNED has the highest correlation in 4 of 5 language directions. Generally, the ranking of metrics is almost the same in each direction.

The only two metrics which also participated in last year metrics task are METEOR and SENTBLEU. In both years, METEOR performed quite well unlike SENTBLEU which was outperformed by most of the metrics.

The metric DISCOTK-LIGHT-KOOL is worth mentioning. It is deliberately designed to assign the same score for all systems for most of the segments. It obtained scores very close to zero (i.e. totally uninformative) in WMT14 variant. In WMT13 thought it reached the highest score.

In directions out of English (Table 5), the metric with highest correlation on average across all directions is BEER, followed by METEOR.

5 Conclusion

In this paper, we summarized the results of the WMT14 Metrics Shared Task, which assesses the quality of various automatic machine translation metrics. As in previous years, human judgements collected in WMT14 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).

This year, neither the system-level nor the segment-level scores are directly comparable to the previous years. The system-level scores are affected by the change of the underlying interpretation of the collected judgements in the main translation task evaluation as well as our choice of Pearson coefficient instead of Spearman's rank correlation. The segment-level scores are affected by the different handling of ties this year. Despite somewhat sacrificing the year-to-year comparability, we believe all changes are towards a fairer evaluation and thus better in the long term.

As in previous years, segment-level correlations are much lower than system-level ones, reaching at most Kendall's τ of 0.45 for the best performing metric in its best language pair. So there is quite some research work to be done. We are happy to see that many new metrics emerged this year, which also underlines the importance of the Metrics Shared Task.

Acknowledgements

This work was supported by the grant FP7-ICT-2011-7-288487 (MosesCore) of the European Union. We are grateful to Jacob Devlin and also Preslav Nakov for pointing out the issue of rewarding ties and for further discussion.

References

Barančíková, P. (2014). Parmesan: Improving Meteor by More Fine-grained Paraphrasing. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, Baltimore, USA. Association for Computational Linguistics.

Bojar, O., Buck, C., Federmann, C., Haddow, B., Koehn, P., Leveling, J., Monz, C., Pecina, P., Post, M., Saint-Amand, H., Soricut, R., Specia, L., and Tamchyna, A. (2014). Findings of the 2014 workshop on statistical machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*.

Callison-Burch, C., Koehn, P., Monz, C., Post, M., Soricut, R., and Specia, L. (2012). Findings of the 2012 workshop on statistical machine translation. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 10–51, Montréal, Canada. Association for Computational Linguistics.

Chen, B. and Cherry, C. (2014). A Systematic Comparison of Smoothing Techniques for Sentence-Level BLEU. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, Baltimore, USA. Association for Computational Linguistics.

Comelles, E. and Atserias, J. (2014). VERTa participation in the WMT14 Metrics Task. In *Proceedings of the Ninth Workshop on Statistical*

Machine Translation, Baltimore, USA. Association for Computational Linguistics.

Denkowski, M. and Lavie, A. (2014). Meteor Universal: Language Specific Translation Evaluation for Any Target Language. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, Baltimore, USA. Association for Computational Linguistics.

Doddington, G. (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.

Echizen'ya, H. (2014). Application of Prize based on Sentence Length in Chunk-based Automatic Evaluation of Machine Translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, Baltimore, USA. Association for Computational Linguistics.

Gautam, S. and Bhattacharyya, P. (2014). LAYERED: Description of Metric for Machine Translation Evaluation in WMT14 Metrics Task. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, Baltimore, USA. Association for Computational Linguistics.

Gonzàlez, M., Barrón-Cedeño, A., and Màrquez, L. (2014). IPA and STOUT: Leveraging Linguistic and Source-based Features for Machine Translation Evaluation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, Baltimore, USA. Association for Computational Linguistics.

Guzman, F., Joty, S., Màrquez, L., and Nakov, P. (2014). DiscoTK: Using Discourse Structure for Machine Translation Evaluation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, Baltimore, USA. Association for Computational Linguistics.

Koehn, P. and Monz, C. (2006). Manual and automatic evaluation of machine translation between european languages. In *Proceedings on the Workshop on Statistical Machine Translation*, pages 102–121, New York City. Association for Computational Linguistics.

Leusch, G., Ueffing, N., and Ney, H. (2006). CDER: Efficient MT Evaluation Using Block Movements. In *In Proceedings of EACL*, pages 241–248.

Libovický, J. and Pecina, P. (2014). Tolerant BLEU: a Submission to the WMT14 Metrics Task. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, Baltimore, USA. Association for Computational Linguistics.

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

Mahmoudi, A., Faili, H., Dehghan, M., and Maleki, J. (2013). ELEXR: Automatic Evaluation of Machine Translation Using Lexical Relationships. In Castro, F., Gelbukh, A., and González, M., editors, *Advances in Artificial Intelligence and Its Applications*, volume 8265 of *Lecture Notes in Computer Science*, pages 394–405. Springer Berlin Heidelberg.

Papineni, K., Roukos, S., Ward, T., and jing Zhu, W. (2002). BLEU: a method for automatic evaluation of machine translation. pages 311–318.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., and Makhoul, J. (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.

Stanojevic, M. and Sima'an, K. (2014). BEER: A Smooth Sentence Level Evaluation Metric with Rich Ingredients. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, Baltimore, USA. Association for Computational Linguistics.

Vazquez-Alvarez, Y. and Huckvale, M. (2002). The reliability of the itu-t p.85 standard for the evaluation of text-to-speech systems. In Hansen, J. H. L. and Pellom, B. L., editors, *INTERSPEECH*. ISCA.

Wu, X. and Yu, H. (2014). RED, The DCU Submission of Metrics Tasks. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, Baltimore, USA. Association for Computational Linguistics.