Results of the WMT13 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 WMT13 Metrics Shared Task. We asked participants of this task to score the outputs of the MT systems involved in WMT13 Shared Translation Task. We collected scores of 16 metrics from 8 research groups. In addition to that we computed scores of 5 standard metrics such as BLEU, WER, PER as baselines. Collected scores were evaluated in terms of system level correlation (how well each metric's scores correlate with WMT13 official human scores) and in terms of segment level correlation (how often a metric agrees with humans in comparing two translations of a particular sentence). This is a corrected version of January 20, 2014.

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 (Callison-Burch et al., 2012). This year, the Metrics Task was run by different organizers but the only visible change is hopefully that the results of the task are presented in a separate paper instead of the main WMT overview paper.

In this task, we asked metrics developers to score the outputs of WMT13 Shared Translation Task (Bojar et al., 2013). 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 is reported in Section 4.

2 Data

We used the translations of MT systems involved in WMT13 Shared Translation Task together with reference translations as the test set for the Metrics Task. This dataset consists of 135 systems' outputs and 6 reference translations in 10 translation directions (5 into English and 5 out of English). Each system's output and the reference translation contain 3000 sentences. For more details please see the WMT13 main overview paper (Bojar et al., 2013).

2.1 Manual MT Quality Judgements

During the WMT13 Translation Task a large scale manual annotation was conducted to compare the systems. We used these collected human judgements for evaluating 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. Only sentences with 30 or less words were ranked by humans.

These collected rank labels were then used to assign each system a score that reflects how high that system was usually ranked by the annotators. Please see the WMT13 main 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 Shared Task

Table 1 lists the participants of WMT13 Shared Metrics Task, along with their metrics. We have collected 16 metrics from a total of 8 research groups.

Metrics	Participant
METEOR	Carnegie Mellon University (Denkowski and Lavie, 2011)
LEPOR, NLEPOR	University of Macau (Han et al., 2013)
ACTA, ACTA5+6	Idiap Research Institute (Hajlaoui, 2013) (Hajlaoui and Popescu-Belis, 2013)
DEPREF-{ALIGN,EXACT}	Dublin City University (Wu et al., 2013)
SIMPBLEU-{RECALL,PREC}	University of Shefield (Song et al., 2013)
MEANT, UMEANT	Hong Kong University of Science and Technology (Lo and Wu, 2013)
TERRORCAT	German Research Center for Artificial Intelligence (Fishel, 2013)
LOGREGFSS, LOGREGNORM	DFKI (Avramidis and Popović, 2013)

Table 1: Participants of WMT13 Metrics Shared Task

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

- **Moses Scorer.** Metrics BLEU (Papineni et al., 2002), 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. In this paper we use the suffix *-MOSES to label these metrics.

- **Mteval.** Metrics BLEU (Papineni et al., 2002) and NIST (Doddington, 2002) were computed using the script `mteval-v13a.pl` [1] which is used in OpenMT Evaluation Campaign and includes its own tokenization. We use *-MTEVAL suffix to label these metrics. By default, `mteval` assumes the text is in ASCII, causing poor tokenization around curly quotes. We run `mteval` in both the default setting as well as with the flag `--international-tokenization` (marked *-INTL).

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

3 System-Level Metric Analysis

We measured the quality of system-level metrics' scores using the Spearman's rank correlation coefficient ρ. For each direction of translation we converted the official human scores into ranks. For each metric, we converted the metric's scores of systems in a given direction into ranks. Since there were no ties in the rankings, we used the simplified formula to compute the Spearman's ρ:

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} \qquad (1)$$

where d_i is the difference between the human rank and metric's rank for system i and n is number of systems. The possible values of ρ range between 1 (where all systems are ranked in the same order) and -1 (where the systems are ranked in the reverse order). A good metric produces rankings of systems similar to human rankings. Since we have normalized all metrics such that better translations get higher score we consider metrics with values of Spearman's ρ closer to 1 as better.

We also computed empirical confidences of Spearman's ρ using bootstrap resampling. Since we did not have direct access to participants' metrics (we received only metrics' scores for the complete test sets without the ability to run them on new sampled test sets), we varied the "golden truth" by sampling from human judgments. We have bootstrapped 1000 new sets and used 95 % confidence level to compute confidence intervals.

The Spearman's ρ correlation coefficient is sometimes too harsh: If a metric disagrees with humans in ranking two systems of a very similar quality, the ρ coefficient penalizes this equally as if the systems were very distant in their quality. Aware of how uncertain the golden ranks are in general, we do not find the method very fair. We thus also computed three following correlation coefficients besides the Spearman's ρ:

- **Pearson's correlation coefficient.** This coefficient measures the strength of the linear relationship between metric's scores and human scores. In fact, Spearman's ρ is Pearson's correlation coefficient applied to ranks.

- **Correlation with systems' clusters.** In the Translation Task (Bojar et al., 2013), the manual scores are also presented as clusters of systems that can no longer be significantly distinguished from one another given

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

the available judgements. (Please see the WMT13 Overview paper for more details). We take this cluster information as a "rank with ties" for each system and calculate its Pearson's correlation coefficient with each metric's scores.

- **Correlation with systems' fuzzy ranks.** For a given system the fuzzy rank is computed as an average of ranks of all systems which are not significantly better or worse than the given system. The Pearson's correlation coefficient of a metric's scores and systems' fuzzy ranks is then computed.

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 Spearman's ρ correlation across translation directions. The best results in each direction are in bold.

As in previous years, a lot of metrics outperformed BLEU in system level correlation. The metric which has on average the strongest correlation in directions into English is METEOR. For the out of English direction, SIMPBLEU-RECALL has the highest system-level correlation. TERRORCAT achieved even a higher average correlation but it did not participate in all language pairs. The implementation of BLEU in `mteval` is slightly better than the one in Moses scorer (BLEU-MOSES). This confirms the known truth that tokenization and other minor implementation details can considerably influence a metric performance.

4 Segment-Level Metric Analysis

We measured the quality of metrics' segment-level scores using Kendall's τ rank correlation coefficient. For this we did not use the official WMT13 human scores but we worked with raw human judgements: For each translation direction we extracted all pairwise comparisons where one system's translation of a particular segment was judged to be (strictly) better than the other system's translation. Formally, this is a list of pairs (a, b) where a segment translation a was ranked better than translation b:

$$Pairs := \{(a, b) \mid r(a) < r(b)\} \tag{2}$$

where $r(\cdot)$ is human rank. For a given metric $m(\cdot)$, we then counted all concordant pairwise comparisons and all discordant pairwise comparisons. A concordant pair is a pair of two translations of the same segment in which the comparison of human ranks agree with the comparison of the metric's scores. A discordant pair is a pair in which the comparison of human ranks disagrees with the metric's comparison. Note that we totally ignore pairs where human ranks or metric's scores are tied. Formally:

$$Con := \{(a, b) \in Pairs \mid m(a) > m(b)\} \tag{3}$$

$$Dis := \{(a, b) \in Pairs \mid m(a) < m(b)\} \tag{4}$$

Finally the Kendall's τ is computed using the following formula:

$$\tau = \frac{|Con| - |Dis|}{|Con| + |Dis|} \tag{5}$$

The 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 higher τ are better.

The final Kendall's τs 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 the translation directions. Metrics which did not compute scores for systems in all directions are at the bottom of the tables.

You can see that in both categories, into and out of English, the strongest correlated segment-level metric is SIMPBLEU-RECALL.

4.1 Details on Kendall's τ

The computation of Kendall's τ has slightly changed this year. In WMT12 Metrics Task (Callison-Burch et al., 2012), the concordant pairs were defined exactly as we do (Equation 3) but the discordant pairs were defined differently: pairs in which one system was ranked better by the human annotator but in which the metric predicted a tie were considered also as discordant:

$$Dis := \{(a, b) \in Pairs \mid m(a) \leq m(b)\} \tag{6}$$

We feel that for two translations a and b of a segment, where a is ranked better by humans, a metric

Correlation coefficient	Spearman's ρ Correlation Coefficient						Pearson's	Clusters	Fuzzy Ranks
Directions	**fr-en**	**de-en**	**es-en**	**cs-en**	**ru-en**	**Average**	**Average**	**Average**	**Average**
Considered systems	12	22	11	10	17				
Meteor	.984 ± .014	.961 ± .020	**.979** ± .024	.964 ± .027	.789 ± .040	**.935** ± .012	**.950**	**.924**	**.936**
DepRef-align	**.995** ± .011	**.966** ± .018	.965 ± .031	.964 ± .023	.768 ± .041	.931 ± .012	.926	.909	.924
UMEANT	.989 ± .011	.946 ± .018	.958 ± .028	**.973** ± .032	.775 ± .037	.928 ± .012	.909	.903	≀ .930
MEANT	.973 ± .014	.926 ± .021	.944 ± .038	**.973** ± .032	.765 ± .038	.916 ± .013	.901	.891	.918
SemPOS	.938 ± .014	.919 ± .028	.930 ± .031	.955 ± .018	**.823** ± .037	.913 ± .012	≀ .934	≀ .894	.901
DepRef-exact	.984 ± .011	.961 ± .017	.937 ± .038	.936 ± .027	.744 ± .046	.912 ± .015	≀ .924	≀ .892	.901
SimpBLEU-recall	.978 ± .014	.936 ± .020	.923 ± .052	.909 ± .027	.798 ± .043	.909 ± .017	≀ .923	.874	.886
BLEU-mteval-intl	.989 ± .014	.902 ± .017	.895 ± .049	.936 ± .032	.695 ± .042	.883 ± .015	.866	.843	.874
BLEU-mteval	.989 ± .014	.895 ± .020	.888 ± .045	.936 ± .032	.670 ± .041	.876 ± .015	.854	.835	.865
BLEU-moses	.993 ± .014	.902 ± .017	.879 ± .051	.936 ± .036	.651 ± .041	.872 ± .016	≀ .856	.826	.861
CDER-moses	**.995** ± .014	.877 ± .017	.888 ± .049	.927 ± .036	.659 ± .045	.869 ± .017	≀ .877	≀ .831	.859
SimpBLEU-prec	.989 ± .008	.846 ± .020	.832 ± .059	.918 ± .023	.704 ± .042	.858 ± .017	≀ .871	.815	.847
nLEPOR	.945 ± .022	.949 ± .025	.825 ± .056	.845 ± .041	.705 ± .043	.854 ± .018	≀ .867	.804	≀ .853
LEPOR v3.100	.945 ± .019	.934 ± .027	.748 ± .077	.800 ± .036	.779 ± .041	.841 ± .020	≀ .869	.780	≀ .850
NIST-mteval	.951 ± .019	.875 ± .022	.769 ± .077	.891 ± .027	.649 ± .045	.827 ± .020	.852	.774	.824
NIST-mteval-intl	.951 ± .019	.875 ± .022	.762 ± .077	.882 ± .032	.658 ± .045	.826 ± .021	≀ .856	.774	≀ .826
TER-moses	.951 ± .019	.833 ± .023	.825 ± .077	.800 ± .036	.581 ± .045	.798 ± .021	.803	.733	.797
WER-moses	.951 ± .019	.672 ± .026	.797 ± .070	.755 ± .041	.591 ± .042	.753 ± .020	.785	.682	.749
PER-moses	.852 ± .027	.858 ± .025	.357 ± .091	.697 ± .043	.677 ± .040	.688 ± .024	.757	.637	.706
TerrorCat	.984 ± .011	.961 ± .023	.972 ± .028	n/a	n/a	**.972** ± .012	**.977**	**.958**	**.959**

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

Correlation coefficient	Spearman's ρ Correlation Coefficient						Pearson's	Clusters	Fuzzy Ranks
Directions	en-fr	en-de	en-es	en-cs	en-ru	Average	Average	Average	Average
Considered systems	14	14	12	11	12				
SimpBLEU-recall	.924 ± .022	**.925** ± .020	.830 ± .047	.867 ± .031	.710 ± .053	**.851** ± .018	.844	.856	**.849**
LEPOR v3.100	.904 ± .034	.900 ± .027	.841 ± .049	.748 ± .056	**.855** ± .048	.850 ± .020	≀ **.854**	.833	.844
NIST-mteval-intl	**.929** ± .032	.846 ± .029	.797 ± .060	.902 ± .045	.771 ± .048	.849 ± .020	.808	≀ **.863**	≀ .845
CDER-moses	.921 ± .029	.867 ± .029	**.857** ± .058	.888 ± .024	.701 ± .059	.847 ± .019	.796	≀ .861	.843
nLEPOR	.919 ± .028	.904 ± .027	.852 ± .049	.818 ± .045	.727 ± .064	.844 ± .021	≀ .849	≀ .846	.840
NIST-mteval	.914 ± .034	.825 ± .030	.780 ± .066	.916 ± .031	.723 ± .048	.832 ± .021	.794	≀ .851	.828
SimpBLEU-prec	.909 ± .026	.879 ± .025	.780 ± .071	.881 ± .035	.697 ± .051	.829 ± .020	≀ .840	≀ .852	.827
Meteor	.924 ± .027	.879 ± .030	.780 ± .060	**.937** ± .024	.569 ± .066	.818 ± .022	≀ .806	.825	.814
BLEU-mteval-intl	.917 ± .033	.832 ± .030	.764 ± .071	.895 ± .028	.657 ± .062	.813 ± .022	≀ .802	.821	.808
BLEU-mteval	.895 ± .037	.786 ± .034	.764 ± .071	.895 ± .028	.631 ± .053	.794 ± .022	≀ .799	.809	.790
TER-moses	.912 ± .038	.854 ± .032	.753 ± .066	.860 ± .059	.538 ± .068	.783 ± .023	.746	.806	.778
BLEU-moses	.897 ± .034	.786 ± .034	.759 ± .078	.895 ± .028	.574 ± .057	.782 ± .022	≀ .802	.792	≀ .779
WER-moses	.914 ± .034	.825 ± .034	.714 ± .077	.860 ± .056	.552 ± .066	.773 ± .024	.737	≀ .796	.766
PER-moses	.873 ± .040	.686 ± .045	.775 ± .047	.797 ± .049	.591 ± .062	.744 ± .024	≀ .758	.747	.739
TerrorCat	**.929** ± .022	**.946** ± .018	**.912** ± .041	n/a	n/a	**.929** ± .017	**.952**	**.933**	**.923**
SemPOS	n/a	n/a	n/a	.699 ± .045	n/a	.699 ± .045	.717	.615	.696
ACTa5 ± 6	.809 ± .046	-.526 ± .034	n/a	n/a	n/a	.141 ± .029	.166	.196	.176
ACTa	.809 ± .046	-.526 ± .034	n/a	n/a	n/a	.141 ± .029	.166	.196	.176

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 other averages are out of sequence compared to the main Spearman's ρ average.

Directions	**fr-en**	**de-en**	**es-en**	**cs-en**	**ru-en**	**Average**
Extracted pairs	80741	128668	67832	85469	151422	
SimpBLEU-recall	**.303**	**.318**	**.388**	.260	.234	**.301**
Meteor	.264	.293	.324	**.265**	**.239**	.277
DepRef-align	.257	.267	.312	.228	.200	.253
DepRef-exact	.258	.263	.307	.227	.195	.250
SimpBLEU-prec	.238	.236	.287	.208	.174	.229
nLEPOR	.225	.240	.281	.176	.172	.219
sentBLEU-moses	.229	.218	.266	.197	.170	.216
LEPOR v3.100	.235	.221	.236	.187	.177	.211
UMEANT	.161	.166	.202	.160	.108	.160
MEANT	.158	.160	.202	.164	.109	.159
TerrorCat	.249	.298	.313	n/a	n/a	.287
logregFSS-33	n/a	.272	n/a	n/a	n/a	.272
logregFSS-24	n/a	.270	n/a	n/a	n/a	.270

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

Directions	**en-fr**	**en-de**	**en-es**	**en-cs**	**en-ru**	**Average**
Extracted pairs	100783	77286	60464	102842	87323	
SimpBLEU-recall	**.261**	**.254**	**.231**	**.192**	**.245**	**.236**
Meteor	.236	.203	.175	.160	.203	.195
SimpBLEU-prec	.219	.197	.187	.148	.175	.185
nLEPOR	.200	.199	.163	.139	.188	.178
sentBLEU-moses	.214	.177	.171	.139	.173	.175
LEPOR v3.100	.206	.179	.178	.084	.205	.170
TerrorCat	.207	.238	.186	n/a	n/a	.210
logregNorm-411	n/a	n/a	.135	n/a	n/a	.135
logregNormSoft-431	n/a	n/a	.033	n/a	n/a	.033

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

which produces equal scores for both translations should not be penalized as much as a metric which strongly disagrees with humans. The method we used this year does not harm metrics which often estimate two segments as equally good.

5 Conclusion

We carried out WMT13 Metrics Shared Task in which we assessed the quality of various automatic machine translation metrics. We used the human judgements as collected for WMT13 Translation Task to compute system-level and segment-level correlations with human scores.

While most of the metrics correlate very well on the system-level, the segment-level correlations are still rather poor. It was shown again this year that a lot of metrics outperform BLEU, hopefully one of them will attract a wider use at last.

Acknowledgements

This work was supported by the grants P406/11/1499 of the Grant Agency of the Czech Republic and FP7-ICT-2011-7-288487 (MosesCore) of the European Union. We would like to thank Jacob Devlin for spotting an obscure off-by-one error which affected sentence-level results.

References

Eleftherios Avramidis and Maja Popović. 2013. Machine learning methods for comparative and time-oriented Quality Estimation of Machine Translation output. In *Proceedings of the Eight Workshop on Statistical Machine Translation*.

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 Eight Workshop on Statistical Machine Translation*.

Chris Callison-Burch, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 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, June. Association for Computational Linguistics.

Michael Denkowski and Alon Lavie. 2011. Meteor 1.3: Automatic Metric for Reliable Optimization and Evaluation of Machine Translation Systems. In *Proceedings of the EMNLP 2011 Workshop on Statistical Machine Translation*.

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.

Mark Fishel. 2013. Ranking Translations using Error Analysis and Quality Estimation. In *Proceedings of the Eight Workshop on Statistical Machine Translation*.

Najeh Hajlaoui and Andrei Popescu-Belis. 2013. Assessing the accuracy of discourse connective translations: Validation of an automatic metric. In *14th International Conference on Intelligent Text Processing and Computational Linguistics*, page 12. University of the Aegean, Springer, March.

Najeh Hajlaoui. 2013. Are ACT's scores increasing with better translation quality. In *Proceedings of the Eight Workshop on Statistical Machine Translation*.

Aaron Li-Feng Han, Derek F. Wong, Lidia S. Chao, Yi Lu, Liangye He, Yiming Wang, and Jiaji Zhou. 2013. A Description of Tunable Machine Translation Evaluation Systems in WMT13 Metrics Task. In *Proceedings of the Eight Workshop on Statistical Machine Translation*.

Gregor Leusch, Nicola Ueffing, and Hermann Ney. 2006. Cder: Efficient mt evaluation using block movements. In *In Proceedings of EACL*, pages 241–248.

Chi-Kiu Lo and Dekai Wu. 2013. MEANT @ WMT2013 metrics evaluation. In *Proceedings of the Eight Workshop on Statistical Machine Translation*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. pages 311–318.

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.

Xingyi Song, Trevor Cohn, and Lucia Specia. 2013. BLEU deconstructed: Designing a better MT evaluation metric. March.

Xiaofeng Wu, Hui Yu, and Qun Liu. 2013. DCU Participation in WMT2013 Metrics Task. In *Proceedings of the Eight Workshop on Statistical Machine Translation*.