ACES: Translation Accuracy Challenge Sets at WMT 2023

Chantal Amrhein[1,2*] and **Nikita Moghe**[3*] and **Liane Guillou**[4*]
[1]Textshuttle, Zurich
[2]Department of Computational Linguistics, University of Zurich
[3]School of Informatics, University of Edinburgh
[4]Department of Computer Science, RISE Research Institutes of Sweden
amrhein@textshuttle.com, nikita.moghe@ed.ac.uk, liane.guillou@ri.se

Abstract

We benchmark the performance of segment-level metrics submitted to WMT 2023 using the ACES Challenge Set (Amrhein et al., 2022). The challenge set consists of 36K examples representing challenges from 68 phenomena and covering 146 language pairs. The phenomena range from simple perturbations at the word/character level to more complex errors based on discourse and real-world knowledge. For each metric, we provide a detailed profile of performance over a range of error categories as well as an overall *ACES*-Score for quick comparison. We also measure the incremental performance of the metrics submitted to both WMT 2023 and 2022. We find that 1) there is no clear *winner* among the metrics submitted to WMT 2023, and 2) performance change between the 2023 and 2022 versions of the metrics is highly variable. Our recommendations are similar to those from WMT 2022. Metric developers should focus on: building ensembles of metrics from different design families, developing metrics that pay more attention to the source and rely less on surface-level overlap, and carefully determining the influence of multilingual embeddings on MT evaluation.

1 Introduction

Challenge sets are a useful tool in measuring the performance of systems or metrics on one or more specific phenomena of interest. They may be used to compare the performance of a range of *different* systems or to identify performance improvement/degradation between successive iterations of the *same* system.

Challenge sets exist for a range of natural language processing (NLP) tasks including Sentiment Analysis (Li et al., 2017; Mahler et al., 2017; Staliūnaitė and Bonfil, 2017), Natural Language Inference (McCoy and Linzen, 2019; Rocchietti et al., 2021), Question Answering (Ravichander et al., 2021), Machine Reading Comprehension (Khashabi et al., 2018), Machine Translation (MT) (King and Falkedal, 1990; Isabelle et al., 2017), and the more specific task of pronoun translation in MT (Guillou and Hardmeier, 2016).

The WMT 2021 Metrics shared task (Freitag et al., 2021) introduced the task of constructing contrastive challenge sets for the evaluation of MT metrics. Contrastive challenge sets aim to assess how well a given metric is able to discriminate between a *good* and *incorrect* translation of the *source* text. The provision of a *reference* translation allows for flexibility: it may be included for the assessment of reference-based (i.e. MT) metrics, or excluded for the assessment of reference-free (i.e. Quality Estimation (QE)) metrics.

We re-submitted ACES[1] (Amrhein et al., 2022), originally developed for the WMT 2022 challenge sets shared task (Freitag et al., 2022), to the corresponding shared task at WMT 2023. ACES largely focuses on translation accuracy errors and consists of 68 phenomena ranging from simple perturbations at the word/character level to more complex errors based on discourse and real-world knowledge. We report on both the performance of metrics submitted to WMT 2023, and on the incremental performance of those metrics that were submitted to both WMT 2022 and WMT 2023. We also repeat the analyses in Amrhein et al. (2022) for the WMT 2023 metrics to confirm whether the findings from WMT 2022 still hold.

Overall, we find similar trends to those observed last year. Again, we do not find one clear winner and whilst neural metrics tend to perform better than their non-neural counterparts, different categories of metrics exhibit different strengths and weaknesses. The major challenges identified in Amrhein et al. (2022) still hold: (i) reference-based metrics are still overly reliant on the reference

*Equal contribution by all authors.

[1]The ACES dataset is available at https://huggingface.co/datasets/nikitam/ACES

Proceedings of the Eighth Conference on Machine Translation (WMT), pages 695–712
December 6–7, 2023.

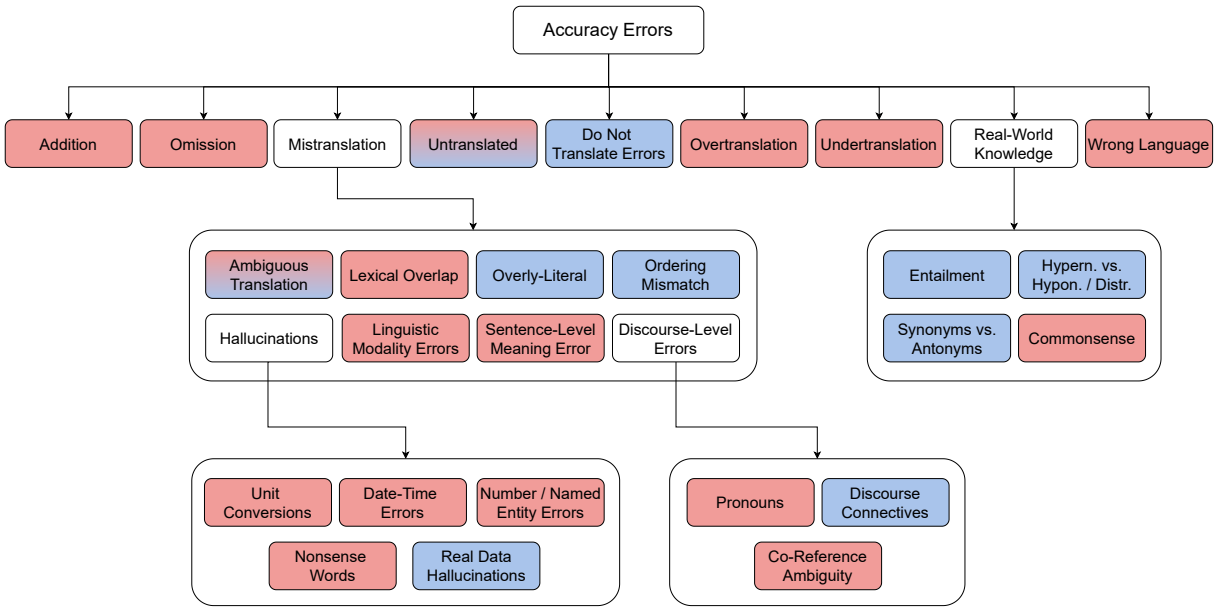

Figure 1: Diagram of the error categories on which our collection of challenge sets is based. Red means challenge sets are created automatically, and blue means challenge sets are created manually.

and do not pay enough attention to the source, (ii) reference-based metrics still rely on surface-level overlap, and (iii) the over-reliance on multilingual embeddings still persists – evidence from our analyses suggests that language agnostic representations present in the multilingual space may harm performance. Accordingly, our recommendations are also similar to those of last year. Metric developers should focus on: constructing ensembles of metrics with different design principles, developing metrics that also focus on information in the source, reducing dependence on surface-level overlap with the reference, and reassessing the impact of multilingual embeddings on MT evaluation.

With respect to incremental performance changes between metrics submitted to both 2022 and 2023, we observe mixed results. Whilst improvements are observed for some metrics, there is a degradation in performance for other metrics. However, even for those metrics for which an overall improvement was observed, this improvement was inconsistent across the top-level categories in ACES. Further, the performance even degraded for some categories.

2 ACES Overview

The Translation **A**ccuracy **C**hallen**E** Set (ACES) consists of 36,476 examples covering 146 language pairs and representing challenges from 68 linguistic phenomena. These phenomena are grouped into ten top-level categories: addition, omission, mistranslation, untranslated, do not translate errors, overtranslation, undertranslation, real-world knowledge, wrong language, and punctuation [2]. The mistranslation and real-world knowledge categories are further sub-categorised to include additional fine-grained categories. We illustrate the broad accuracy error categories in Fig 1 and give examples for each of the top-level categories in Appendix A.

The focus of ACES is on translation accuracy errors, reflecting the need to evaluate contemporary MT systems that are capable of producing fluent but potentially error-prone output. The selection of the top-level categories in the ACES error hierarchy is based on the *Accuracy* class in the Multidimensional Quality Metrics (MQM) ontology (Lommel et al., 2014), and extended to include translations defying *real-world knowledge* and translations in the *wrong language*. ACES includes a wide range of phenomena ranging from simple perturbations that involve the omission/addition of characters or tokens, to more complex scenarios involving mistranslations e.g. ambiguity or hallucinations in translation, untranslated elements of a sentence, discourse-level phenomena, and real-world knowledge.

Each ACES example consists of a *source* sentence, a *reference* translation, a *phenomenon label* indicating the error type, and two translation

[2]Note that although the focus of ACES is on *accuracy* errors, we also include a small set of *fluency* errors for the punctuation category.

	Mistranslation - Overly literal (Idioms)
SRC (de):	Er hat versucht, mir die Spielregeln zu erklären, aber **ich verstand nur Bahnhof**.
REF (en):	He tried to explain the rules of the game to me, but **I did not understand them**.
✓:	He tried to explain the rules of the game to me, but **it was all Greek to me**.
✗:	He tried to explain the rules of the game to me, but **I only understood train station**.
	Real-world Knowledge - Commonsense
SRC (en):	Die Luft im Haus war kühler als in der Wohnung.
REF (de):	The air in the house was cooler than in the apartment.
✓:	The air in the house was cooler than in the apartment because **the apartment** had a broken air conditioner.
✗:	The air in the house was cooler than in the apartment because **the house** had a broken air conditioner.

Table 1: Examples from the Mistranslation and Real-world Knowledge categories in ACES. An example consists of a source sentence (SRC), reference (REF), good (✓) and incorrect (✗) translations, and a phenomenon label indicating the error type. en: English, de: German. Top: the German idiom "ich verstand nur Bahnhof" has been translated in an overly-literal way in the incorrect translation. Bottom: the incorrect translation contains an error in commonsense reasoning as to why the air in the house was cooler than in the apartment.

hypotheses: an *incorrect* translation containing an error relating to the phenomenon of interest, and a *good* translation. Several examples from ACES are presented in Table 1. In the top example, from the *Mistranslation* error category, the incorrect translation contains an *overly literally* translation of the German *idiom* "ich verstand nur Bahnhof" (corresponding to the English expression "it was all Greek to me"). In the bottom example, from the *Real-world Knowledge* error category, both the good and incorrect translations contain additional information not present in the source sentence, however, the incorrect translation contains an error in *commonsense* reasoning as to why the air in the house was cooler than the apartment.

ACES examples were constructed from pre-existing datasets, using a range of automatic, semi-automatic, and manual methods. A detailed description of each of the phenomena in ACES can be found in Amrhein et al. (2022).

3 Related Work

Challenge sets have been used for several tasks (Li et al. (2017); McCoy and Linzen (2019); Ravichander et al. (2021), *inter alia*) to investigate the behaviour of these tasks under a specific phenomenon rather than the standard test distribution (Popović and Castilho, 2019). Lately, with the success of neural metrics, the development of challenge sets for MT evaluation has promoted great interest in studying the strengths and weaknesses of these metrics. We summarise here recent work on challenge sets for MT metric evaluation.

DEMETR (Karpinska et al., 2022), which comprises 31K English examples translated from ten languages, was developed for evaluating MT metric sensitivity to a range of 35 different types of linguistic perturbations, belonging to semantic, syntactic, and morphological error categories. These were divided into minor, major, and critical errors according to the type of perturbation, similar to the grading of error categories to compute the weighted ACES-Score. As in ACES, example generation was carefully designed to form minimal pairs such that the perturbed translation only differs from the actual translation in one aspect. The application of DEMETR in evaluating a suite of baseline metrics revealed a similar pattern to the analyses in Amrhein et al. (2022) - that metric performance varies considerably across the different error categories, often with no clear *winner*. It is worth noting that DEMETR and ACES each have their respective advantages: all examples in DEMETR have been verified by human annotators; ACES provides broader coverage in terms of both languages and linguistic phenomena.

In addition to ACES, three other datasets were submitted to the WMT 2022 challenge sets shared task (Freitag et al., 2022): SMAUG (Alves et al., 2022), the HWTSC challenge set (Chen et al., 2022), and the DFKI challenge set (Avramidis and Macketanz, 2022). These datasets differ from ACES in terms of their size, and the languages and phenomena/categories they cover (see Table 2).

Both SMAUG and HWTSC contained five different phenomena each pertaining to a single category of critical error for meaning change. In comparison, the DFKI challenge set has over 100 linguistically motivated phenomena, organised into 14 categories. Whereas the aim of ACES was to provide a broad coverage of language pairs, the other datasets provide an in-depth focus on specific lan-

	Ex.	Lang. pairs	Phenomena	Categories
SMAUG	632	2	5	5
HWTSC	721	1	5	5
DFKI	19,347	1	>100	14
ACES	36,476	146	68	10

Table 2: Comparison of challenge sets for MT metric evaluation in terms of: **Ex**amples, **Lang**uage**-pairs**, Phenomena, and Categories.

guage pairs: SMAUG (pt↔en and es→en), DFKI (de↔en), and HWTSC (zh↔en). Whilst there is a clear overlap between the ACES phenomena and those in SMAUG and HWTSC, many of the phenomena in the DFKI dataset are complementary such that in the case of evaluating metrics for the German-English pair, metric developers might consider benchmarking on both datasets.

4 Metrics

We list below the metrics that participated in the 2023 challenge set shared task and the baselines provided by the organisers.

4.1 Baseline Metrics

BERTScore (Zhang et al., 2020) uses contextual embeddings from pre-trained language models to compute the cosine similarity between the tokens in the hypothesis and the reference translation. The resulting similarity matrix is used to compute precision, recall, and F1-scores.

BLEURT-20 (Sellam et al., 2020) is a BERT-based (Devlin et al., 2019) regression model, which is first trained on scores produced by automatic metrics/similarity of pairs of reference sentences and their corrupted counterparts. It is then fine-tuned on WMT human evaluation data to provide a score for a hypothesis given a reference translation.

BLEU (Papineni et al., 2002) compares the token-level n-grams in the hypothesis with those in the reference translation. It then computes a precision score weighted by a brevity penalty.

chrF (Popović, 2017) provides a character n-gram F-score by computing overlaps between the hypothesis and reference translation.

COMET-22 (Rei et al., 2022) is an ensemble between a vanilla COMET model (Rei et al., 2020) trained with Direct Assessment (DA) scores and a multitask model that is trained on regression (MQM regression) and sequence tagging (OK/BAD word identification from MQM span annotations). These models are ensembled together using a hyperparameter search that weights different features extracted from these two evaluation models and combines them into a single score. The vanilla COMET model is trained with DAs ranging from 2017 to 2020 while the Multitask model is trained using DAs ranging from 2017 to 2020 plus MQM annotations from 2020 (except for en-ru which uses TedTalk annotations from 2021).

COMET-Kiwi (Rei et al., 2022) ensembles two QE models similarly to COMET-22. The first model follows the classic Predictor-Estimator QE architecture where MT and source are encoded together. This model is trained on DAs ranging from 2017 to 2019 and then fine-tuned on DAs from MLQE-PE (the official DA from the QE shared task). The second model is the same multitask model used in the COMET-22 submission but without access to a reference translation. This means that this model is a multitask model trained on regression and sequence tagging. Both models are ensembled together using a hyperparameter search that weights different features extracted from these two QE models and combines them into a single score.

f200spBLEU (Goyal et al., 2022) computes BLEU over text tokenised with a single language-agnostic SentencePiece subword model. For the f200spBLEU version of spBLEU used in this year's shared task, the SentencePiece tokeniser (Kudo and Richardson, 2018) was trained using data from the FLORES-200 languages.

MS-COMET-22 (Kocmi et al., 2022) is built on top of the COMET (Rei et al., 2020) architecture. It is trained on a set of human judgments several times larger – covering 113 languages and 15 domains. Furthermore, the authors propose filtering out those human judgements with potentially low quality. MS-COMET-22 is a reference-based metric that receives the source, the MT hypothesis and the human reference as input.

Random-sysname is a random baseline. The metric takes the name of the system as the only parameter. It uses a discrete score. Segment-level scores follow a Gaussian distribution around mean value X (in the range 0-9) and a standard deviation of 2. The mean X is calculated from the name of the system as: $X = sha256(sysname)[0]\%10$

The idea behind this baseline is two-fold. Firstly, having a baseline showing how a random metric would perform could help to put scores into context (in particular, pairwise accuracy can create a perception of great performance while 50% is just a toss of a coin). Secondly, it could help to detect errors in metric meta-evaluations.

YiSi-1 (Lo, 2019) measures the semantic similarity between the hypothesis and the reference translation by using cosine similarity scores of multilingual representations at the lexical level. It optionally uses a semantic role labeller to obtain structural similarity. Finally, a weighted F-score based on structural and lexical similarity is used for scoring the hypothesis against the reference translation.

4.2 Metrics Submitted to WMT 2023

We list the descriptions of the metrics submitted to WMT 2023 by the metric developers and refer the reader to the relevant system description papers for further details.

Embed_Llama relies on pretrained Llama 2 embeddings, without any finetuning, to transform sentences into a vector space that establishes connections between geometric and semantic proximities. This metric draws inspiration from Word2vec and utilizes cosine distance for the purpose of estimating similarity or dissimilarity between sentences.

MetricX-23 and MetricX-23-QE are learned reference-based and reference-free (respectively) regression metrics based on the mT5 encoder-decoder language model. They further finetune the mT5-XXL checkpoint on direct assessment data from 2015-2020 and MQM data from 2020 to 2021 as well as synthetic data.

Tokengram_F is an F-score-based evaluation metric that is heavily inspired by chrF++. By replacing word-grams with token-grams obtained from contemporary tokenization algorithms, tokengram_F captures similarities between words sharing the same semantic roots and thus obtains more accurate ratings.

Partokengram_F we did not receive a description of this metric.

XCOMET is a new COMET-base model that is trained to identify errors in sentences along with a final quality score and thus leads to an explainable neural metric. The metric is optimised towards regression and error span detection simultaneously. The same model may be used both with references (XCOMET) and without references (XCOMET-QE). The models are built using XLM-R XL and XXL, thus XCOMET-XL has 3.5B parameters and XCOMET-XXL has 10.7B parameters. The metric is trained in stages where it first sees DAs and then is fine-tuned with MQM. XCOMET-ENSEMBLE is an ensemble between 1 XL and 2 XXL checkpoints that result from these training stages.

XLsim is a supervised reference-based metric that regresses on human scores provided by WMT (2017-2022). Using a cross-lingual language model (XLM-RoBERTa (Conneau et al., 2020)), a supervised model is trained using a Siamese network architecture with CosineSimilarityLoss. **XLSIMQE** is the reference-free variant of this metric.

Cometoid22 is a reference-free metric created using knowledge distillation from reference-based metrics. First, using COMET-22 as a teacher metric, the MT outputs submitted to the WMT News/General Translation task since 2009 are scored. Next, a student metric, called Cometoid22, is trained to mimic the teacher scores without using reference translation. The student metric has the same architecture as COMET-QE, and is initialised with pre-trained weights from the multilingual language model InfoXLM. Three variants were submitted: cometoid22-wmt21,22,23, where the suffix indicates the training data cut-off year.

COMETKiwi-XL and COMETKiwi-XXL use the same COMETKiwi model architecture from WMT 2022 but replace InfoXLM with XLM-R XL and XXL (for COMETKIWI-XL and COMETKIWI-XXL respectively).

KG-BERTScore incorporates a multilingual knowledge graph into BERTScore and generates the final evaluation score by linearly combining the results of KGScore and BERTScore. In contrast to last year, COMET-QE is used to calculate BERTScore.

GEMBA-MQM is an LLM-enabled metric for error quality span marking. It uses three-shot prompting with a GPT-4 model. In contrast to EAPrompt (Lu et al., 2023), it does not require language-specific examples and requires only a single prompt.

5 Results

5.1 Phenomena-level Results

As in Amrhein et al. (2022) we begin by providing a broad overview of metric performance on the different phenomena categories, before conducting more detailed analyses (see Section 5.3). We restrict the overview to the metrics which provide a) segment-level scores and b) scores for all language pairs and directions in ACES. Out of the metrics that participated, 33 fulfil these criteria: 10 baselines, 11 reference-based, and 12 reference-free metrics.

We first compute the Kendall's tau-like correlation scores[3] (Freitag et al., 2021, 2022) for all of the ACES examples. This metric measures the number of times a metric scores the good translation above the incorrect translation (concordant) and equal to or lower than the incorrect translation (discordant):

$$\tau = \frac{concordant - discordant}{concordant + discordant}$$

We then report the average score over all examples belonging to each of the nine top-level accuracy categories in ACES, plus the fluency category *punctuation* (see Table 3). In addition, we compute the ACES-Score, a weighted combination of the top-level categories, which allows us to identify high-level performance trends of the metrics (see Equation 1). Note that the ACES-Score ranges from -29.1 (all phenomena have a correlation of -1) to 29.1 (all phenomena have a correlation of +1).

$$\text{ACES} = sum \left\{ \begin{array}{r} 5 * \tau_{\text{addition}} \\ 5 * \tau_{\text{omission}} \\ 5 * \tau_{\text{mistranslation}} \\ 1 * \tau_{\text{untranslated}} \\ 1 * \tau_{\text{do not translate}} \\ 5 * \tau_{\text{overtranslation}} \\ 5 * \tau_{\text{undertranslation}} \\ 1 * \tau_{\text{real-world knowledge}} \\ 1 * \tau_{\text{wrong language}} \\ 0.1 * \tau_{\text{punctuation}} \end{array} \right\} \quad (1)$$

Overall, the best-performing metrics submitted to this year's shared task, according to the ACES-Score, are COMETKIWI (a reference-free baseline metric), and KG-BERTSCORE (a reference-free metric). BLEU remains one of the worst-performing metrics, with only the random baseline, RANDOM-SYSNAME, achieving a lower ACES-Score. XCOMET-ENSEMBLE is the top ranking among the reference-based metrics. We caution that we developed ACES to investigate strengths and weaknesses of metrics on a phenomena level – hence, we advise the reader not to draw any conclusions based solely on the ACES-Score.

As observed in Amrhein et al. (2022) the performance of the metrics is highly variable, with no clear *winner* in terms of performance across all of the top-level ACES categories. We also observe similar trends in terms of the most challenging categories (*addition*, *undertranslation*, *real-world knowledge*, and *wrong language*). We find that, unlike last year, some metrics perform similarly to or worse than the baseline metrics. In particular, EMBED_LLAMA and GEMBA-MQM which are designed using Large Language Models (LLMs), struggle with this challenge set. This suggests that we need better design strategies in using the rich representations from LLMs for MT evaluation. In general, we find that reference-free metrics perform on par or better than reference-based metrics.

In terms of performance across the top-level categories, we also observe variation in the performance of metrics belonging to the baseline, reference-based, and reference-free groups. The reference-free group exhibits overall stronger performance compared with the other groups, but in particular for the *mistranslation*, *overtranslation*, *undertranslation*, and *real-world knowledge* categories.

[3]Evaluation scripts are available here: https://github.com/EdinburghNLP/ACES

	addition	omission	mistranslation	untranslated	do not translate	overtranslation	undertranslation	real-world knowledge	wrong language	punctuation	ACES-Score
Examples	*999*	*999*	*24457*	*1300*	*100*	*1000*	*1000*	*2948*	*2000*	*1673*	
BERTscore	**0.872**	0.754	0.318	0.771	0.940	-0.186	-0.288	0.030	0.551	**0.844**	9.722
BLEU	0.742	0.427	-0.227	0.353	0.580	-0.838	-0.856	-0.768	0.660	0.704	-2.862
BLEURT-20	0.435	0.812	0.427	0.743	0.860	0.202	0.014	0.388	0.536	0.708	12.048
chrF	0.644	0.784	0.162	**0.781**	**0.960**	-0.696	-0.592	-0.294	0.693	0.773	3.728
COMET-22	0.295	0.822	0.402	0.718	0.820	0.502	0.258	0.382	0.078	0.673	13.458
CometKiwi	0.536	**0.918**	0.614	-0.105	0.520	0.766	0.604	0.577	-0.307	0.765	**17.951**
f200spBLEU	0.666	0.584	-0.082	0.680	0.920	-0.752	-0.794	-0.394	0.657	0.708	0.041
MS-COMET-QE-22	-0.179	0.674	0.440	0.394	0.300	0.524	0.382	0.262	-0.195	0.632	10.027
Random-sysname	-0.117	-0.117	-0.116	-0.083	-0.100	-0.118	-0.152	-0.245	-0.113	-0.074	-3.648
YiSi-1	0.766	0.868	0.354	0.720	0.940	-0.062	-0.076	0.110	0.421	0.763	11.517
eBLEU	0.674	0.682	0.197	0.739	0.880	-0.662	-0.684	-0.042	**0.771**	0.270	3.406
embed_llama	0.211	0.457	0.016	0.503	0.400	-0.170	-0.492	-0.165	0.154	0.476	1.054
MetricX-23	-0.027	0.568	0.578	0.473	0.800	0.790	0.586	0.766	-0.486	0.636	14.091
MetricX-23-b	-0.135	0.622	0.572	0.613	0.860	0.772	0.568	0.749	-0.444	0.532	13.826
MetricX-23-c	-0.015	0.794	0.617	0.611	0.800	0.740	0.526	**0.783**	-0.629	0.527	14.929
partokengram_F	0.087	0.191	-0.034	0.310	0.140	-0.042	-0.028	0.032	0.508	0.171	1.878
tokengram_F	0.698	0.758	0.160	0.779	**0.960**	-0.732	-0.632	-0.273	0.687	0.830	3.492
XCOMET-Ensemble	0.311	0.786	0.663	0.379	0.780	**0.794**	0.612	0.708	-0.423	0.595	17.336
XCOMET-XL	0.169	0.542	0.570	0.222	0.800	0.656	0.464	0.582	-0.367	0.220	13.264
XCOMET-XXL	-0.119	0.413	0.547	0.234	0.600	0.736	0.568	0.508	-0.507	0.509	11.610
XLsim	0.429	0.618	0.153	0.643	0.820	-0.210	-0.290	-0.044	0.392	0.753	5.386
cometoid22-wmt21	-0.339	0.658	0.493	-0.076	0.280	0.670	0.566	0.362	-0.454	0.608	10.409
cometoid22-wmt22	-0.301	0.674	0.493	-0.119	0.280	0.686	0.538	0.340	-0.472	0.599	10.534
cometoid22-wmt23	-0.253	0.702	0.502	-0.046	0.420	0.750	0.590	0.362	-0.319	0.557	11.926
CometKiwi-XL	0.239	0.828	0.624	0.239	0.440	0.762	0.560	0.563	-0.380	0.630	15.988
CometKiwi-XXL	0.361	0.828	0.653	0.414	0.320	0.774	0.560	0.683	-0.537	0.503	16.809
GEMBA-MQM	0.037	0.281	0.153	0.094	0.140	0.466	0.276	0.268	-0.150	0.015	6.419
KG-BERTScore	0.538	0.912	0.585	-0.206	0.700	0.772	0.606	0.594	-0.307	0.654	17.906
MetricX-23-QE	0.045	0.678	0.654	0.379	0.460	0.772	0.612	0.654	-0.702	0.226	14.614
MetricX-23-QE-b	0.027	0.760	0.663	0.489	0.480	0.758	**0.620**	0.647	-0.673	0.256	15.106
MetricX-23-QE-c	-0.115	0.664	**0.721**	0.384	0.340	0.726	0.618	0.753	-0.712	0.375	13.873
XCOMET-QE-Ensemble	0.277	0.754	0.644	0.181	0.720	0.764	0.582	0.626	-0.519	0.449	16.156
XLsimQE	0.205	0.383	0.087	-0.694	0.940	0.454	0.352	0.042	0.307	0.671	8.070
Average	0.232	0.639	0.382	0.349	0.609	0.314	0.187	0.289	-0.069	0.532	10.002

Table 3: Average Kendall's tau-like correlation results for the ACES top-level categories and ACES-Scores (final column). Metrics are grouped into baseline (top), and participating reference-based (middle) and reference-free (bottom) metrics. Note that *Average* is an average over averages. Best results are highlighted in green.

5.2 Mistranslation Results

Next, we drill down to the fine-grained categories of the largest ACES category: *mistranslation*. We present metric performance for the sub-level categories (*discourse*, *hallucination*, and *other*) in Table 4. The *discourse* sub-category includes errors involving the mistranslation of discourse-level phenomena such as pronouns and discourse connectives. *Hallucination* includes errors at the word level that could occur due to hallucination by an MT model, for example, the use of wrong units, dates, times, numbers or named entities, as well as hallucinations at the subword level that result in nonsensical words. The *other* cub-category covers all other categories of mistranslation errors including overly-literal translations (see example in Table 1) and the introduction of ambiguities in the translation output. Again, as in 2022, we find that performance on the different sub-categories is highly variable, with no clear *winner* among the metrics. We also make similar observations to those in Amrhein et al. (2022), that the hallucination phenomena are generally more challenging than discourse-level phenomena; performance on the hallucination sub-category is poor overall.

5.3 Analysis

We repeat the analyses we performed in Amrhein et al. (2022) for the metrics submitted to WMT 2023 to confirm whether our findings from WMT 2022 still hold. We highlight similar observations to those from WMT 2022 and summarise our insights below.

5.3.1 How sensitive are metrics to the source?

Finding: Reference-based metrics tend to ignore the source.

In the ACES *Mistranslation - Ambiguous Translations* category, examples were designed in such a way that given an ambiguous reference the correct translation candidate can only be identified through the source sentence (See an example in Table 9). We leverage this property to present an analysis aimed at discovering how important the source is for different metrics. We exclude from the analysis all metrics that a) do not take the source and b) do not cover all language pairs. This leaves us with 22 metrics: seven reference-based metrics, fourteen reference-free metrics, and the RANDOM-SYSNAME baseline. In Table 5 we present results for the *Ambiguity - Discourse Connectives* (for the ambiguous English discourse connective "since" which can have either causal or temporal meaning), and *Ambiguity - Occupation Names Gender* (male and female) challenge sets.

	disco.	**halluci.**	**other**
Examples	***3698***	***10270***	***10489***
BERTscore	0.563	-0.062	0.361
BLEU	-0.042	-0.418	-0.250
BLEURT-20	0.695	0.141	0.398
chrF	0.406	-0.138	0.160
COMET-22	0.657	0.113	0.383
CometKiwi	0.779	0.465	0.580
f200spBLEU	0.095	-0.190	-0.150
MS-COMET-QE-22	0.631	0.240	0.417
Random-sysname	-0.117	-0.122	-0.111
YiSi-1	0.608	0.017	0.366
eBLEU	0.374	-0.166	0.282
embed_llama	-0.089	-0.140	0.189
MetricX-23	0.757	0.663	0.393
MetricX-23-b	0.749	0.656	0.390
MetricX-23-c	0.694	**0.755**	0.477
partokengram_F	-0.062	-0.101	0.027
tokengram_F	0.396	-0.132	0.157
XCOMET-Ensemble	**0.791**	0.566	0.626
XCOMET-XL	0.706	0.482	0.521
XCOMET-XXL	0.609	0.540	0.504
XLsim	0.217	-0.066	0.236
cometoid22-wmt21	0.782	0.286	0.400
cometoid22-wmt22	0.748	0.290	0.423
cometoid22-wmt23	0.758	0.223	0.478
CometKiwi-XL	0.752	0.501	0.602
CometKiwi-XXL	0.735	0.535	0.661
GEMBA-MQM	0.076	0.291	0.127
KG-BERTScore	0.685	0.466	0.580
MetricX-23-QE	0.728	0.604	0.628
MetricX-23-QE-b	0.694	0.617	0.666
MetricX-23-QE-c	0.747	0.659	**0.739**
XCOMET-QE-Ensemble	0.702	0.558	0.651
XLsimQE	0.053	0.050	0.134
Average	0.511	0.248	0.365

Table 4: Average Kendall's tau-like correlation results for the sub-level categories in mistranslation: **disco**urse-level, **halluci**nation, and **other** errors. Metrics are grouped into baseline (top), and participating reference-based (middle) and reference-free (bottom) metrics. Note that *Average* is an average over averages. Best results are highlighted in green.

In addition, we measure the correlation gain when metrics receive access to disambiguation information via the source – for this we use the *Real-world Knowledge - Commonsense* challenge set i.e. a scenario in which the source contains disambiguation information (See an example in Table 1). In Table 6 we observe that the correlation gain is lower for the majority of the reference-based metric correlation scores compared with the reference-free metric correlation scores, when access to the subordinate clause is provided via the source.

	since		female		male		
	causal	**temp.**	**anti.**	**pro.**	**anti.**	**pro.**	**AVG**
Examples	*106*	*106*	*1000*	*806*	*806*	*1000*	*3824*
Random-sysname	-0.075	-0.019	-0.146	-0.156	-0.109	-0.154	-0.110
COMET-22	-0.868	0.887	-0.254	0.591	-0.467	0.432	0.053
MetricX-23	-1.000	**1.000**	-0.864	-0.062	0.062	0.870	0.001
MetricX-23-b	-1.000	**1.000**	-0.790	0.112	-0.092	0.780	0.002
MetricX-23-c	-0.849	0.849	-0.998	-0.581	**0.576**	**0.996**	-0.001
XCOMET-Ensemble	-0.585	0.981	**0.852**	0.948	0.273	0.922	**0.565**
XCOMET-XL	-0.698	0.906	0.456	**0.960**	-0.330	0.698	0.332
XCOMET-XXL	-0.868	0.925	0.372	0.675	0.541	0.918	0.427
cometoid22-wmt21	-0.698	0.868	0.580	0.950	-0.787	0.022	0.156
cometoid22-wmt22	-0.623	0.868	0.456	0.851	-0.444	0.442	0.258
cometoid22-wmt23	-0.566	0.925	0.342	0.851	0.117	0.844	0.419
CometKiwi	0.075	**1.000**	**0.990**	**0.998**	-0.171	0.440	0.555
CometKiwi-XL	0.075	0.925	0.952	0.990	0.380	0.892	**0.702**
CometKiwi-XXL	**0.132**	0.943	0.932	0.995	0.241	0.796	0.673
GEMBA-MQM	-0.604	0.736	0.722	0.320	-0.762	-0.692	-0.047
KG-BERTScore	0.075	**1.000**	**0.990**	**0.998**	-0.171	0.440	0.555
MS-COMET-QE-22	-0.283	0.811	-0.194	0.323	0.243	0.692	0.265
MetricX-23-QE	-0.472	0.736	0.974	0.995	0.117	0.816	0.528
MetricX-23-QE-b	-0.566	0.868	0.968	0.995	0.722	**0.968**	0.659
MetricX-23-QE-c	-0.302	0.774	0.968	**0.998**	**0.911**	0.866	**0.702**
XCOMET-QE-Ensemble	-0.208	0.925	0.930	0.975	0.546	0.912	0.680
XLsimQE	0.245	-0.113	0.208	0.350	-0.256	-0.170	0.044

Table 5: Results on the challenge sets where the good translation can only be identified through the source sentence. Upper block: reference-based metrics, lower block: reference-free metrics. Best results for each phenomenon and each group of models is marked in bold and green and the average over all can be seen in the last column.

In line with last year's findings, we report that reference-based metrics still lag behind reference-free metrics in terms of their correlation on challenge sets that can only be disambiguated by looking at the source. This indicates that reference-based metrics still rely too much on the reference translation. We conclude that our initial finding from 2022 still holds: that reference-based metrics tend to ignore relevant information in the source. One exception is XCOMET-ENSEMBLE, a reference-based metric that reaches similar correlations and correlation gains as some of the mid-performing reference-free metrics. We suspect that by training the same model to exhibit reference-based and reference-free behaviour, the model learns to utilise the information from the source in addition to the reference, when provided.

5.3.2 How much do metrics rely on surface overlap with the reference?

Finding: Reference-based metrics still rely on reference overlap.

Surface-level metrics are often too reliant on overlap with the reference. We aim to discover whether neural reference-based metrics submitted to the 2023 shared task are able to avoid this problem. Using the *Hallucination - Numbers and Named Entities* challenge set we compared how well reference-based and reference-free metrics[4] on average can identify *number* and *named entity* mismatches. In these challenge sets, we perform both word-level and character-level edits (i.e. substitutions) to simulate the hallucination behaviour. In order to thoroughly understand the behaviour of

[4]Excluding surface-level overlap metrics (BLEU, CHRF, FP200SPBLEU, PARTOKENGRAM_F, TOKENGRAM_F).

	corr. gain
Random-sysname	-0.052
COMET-22	0.042
MetricX-23	0.004
MetricX-23-b	-0.002
MetricX-23-c	0.008
XCOMET-Ensemble	**0.162**
XCOMET-XL	0.110
XCOMET-XXL	0.016
cometoid22-wmt21	0.120
cometoid22-wmt22	0.124
cometoid22-wmt23	0.138
CometKiwi	0.454
CometKiwi-XL	0.148
CometKiwi-XXL	0.108
GEMBA-MQM	**1.107**
KG-BERTScore	0.436
MS-COMET-QE-22	0.198
MetricX-23-QE	0.272
MetricX-23-QE-b	0.296
MetricX-23-QE-c	0.142
XCOMET-QE-Ensemble	0.112
XLsimQE	0.184

Table 6: Results on the *real-world knowledge commonsense challenge set* with reference-based metrics in the upper block and reference-free metrics in the lower block. The numbers are computed as the difference between the correlation with the subordinate clause in the source and the correlation without the subordinate clause in the source. Largest gains are bolded.

metrics under such hallucination errors, we introduced three levels. The first, easiest level follows Freitag et al. (2021) and applies a change to an alternative translation to form an incorrect translation. The second level uses an alternative translation that is lexically very similar to the reference as the good translation and applies a change to the reference to form an incorrect translation. The third, and hardest level, uses an alternative translation that is lexically very different from the reference as the good translation and applies a change to the reference to form an incorrect translation. In this way, our challenge set tests whether number and named entity differences can still be detected as the surface similarity between the two translation candidates decreases and the surface similarity between the incorrect translation and the reference increases. See an example of the different levels below as taken from the dataset paper -

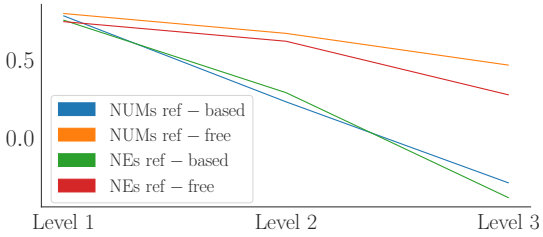

Figure 2: Decrease in correlation for reference-based and reference-free metrics on the named entity and number hallucination challenge sets.

SRC (es):	Sin embargo, Michael Jackson, Prince y **Madonna** fueron influencias para el álbum.
REF (en):	Michael Jackson, Prince and **Madonna** were, however, influences on the album.
Level-1 ✓:	However, Michael Jackson, Prince, and **Madonna** were influences on the album.
Level-1 ✗:	However, Michael Jackson, Prince, and **Garza** were influences on the album.
Level-2 ✓:	However, Michael Jackson, Prince, and **Madonna** were influences on the album.
Level-2 ✗:	Michael Jackson, Prince and **Garza** were, however, influences on the album.
Level-3 ✓:	The record was influenced by **Madonna**, Prince, and Michael Jackson though.
Level-3 ✗:	Michael Jackson, Prince and **Garza** were, however, influences on the album.

We take the average correlation for all reference-based and reference-free metrics that cover all languages. We then plot the decrease in correlation with increasing surface-level similarity of the incorrect translation to the reference (Figure 2). As in the corresponding analysis of the WMT 2022 metrics, we observe that, on average, reference-based metrics have a much steeper decrease in correlation than the reference-free metrics as the two translation candidates become more and more lexically diverse and the surface overlap between the incorrect translation and the reference increases. This indicates that reference-based metrics may prefer a) an incorrect translation in cases where it is lexically similar to the reference but contains a severe error over b) a good translation that shares little overlap with the reference.

We also observe a clear effect of surface-level overlap between the reference and the hypothesis

	reference-based	reference-free
hallucination	-0.32 ± 0.15	+0.06 ± 0.06
overly-literal	-0.22 ± 0.14	0.00 ± 0.03
untranslated	-0.44 ± 0.11	-0.03 ± 0.06

Table 7: Average correlation difference and standard deviation between the challenge sets with reference-copied good translations and the challenge sets with the synonymous good translations.

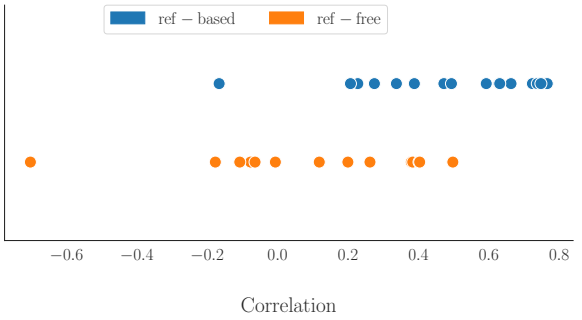

Figure 3: Correlation of reference-based metrics (blue) and reference-free metrics (orange) on the sentence-level untranslated test challenge set.

on three challenge sets for which we have different versions of the good translation, where the error was corrected with: a) the corresponding *correct token* from the reference and b) *a synonym for the correct token* from the reference. In Table 7, we can see a much larger difference in correlation between the challenge sets with reference-copied good translations and the challenge sets with the synonymous good translations, for the reference-based metrics as compared to the reference-free metrics. That is, it is much easier for reference-based metrics to identify mistranslations when the good translation matches a term in the reference compared with when a synonym is used. Furthermore, when the incorrect translation *shares a high degree of lexical overlap with the reference but does not have the same meaning* (as in the *Mistranslation - Lexical Overlap* challenge set based on adversarial paraphrase from PAWS-X (Yang et al., 2019)), the reference-based metrics only reach a correlation of 0.05 ± 0.16 on average. In contrast, the reference-free metrics reach a correlation of 0.27 ± 0.16.

We again conclude that although state-of-the-art reference-based MT evaluation metrics are no longer solely reliant on surface-level overlap, it still has a considerable influence on their predictions.

5.3.3 Do multilingual embeddings help design better metrics?

Finding: Multilingual embeddings can be harmful with poor design.

We are interested in the extent to which the representations in neural MT evaluation metrics, which are trained on multilingual models, are language-dependent. For this analysis, we investigated the effect of alignment of multilingual embeddings (including LLMs) on the evaluation task through the *wrong-language* and *untranslated - full sentences* phenomena for those metrics that provided scores for examples in all language pairs. In the *wrong-language* phenomenon, the incorrect translation contains a high-quality translation of the source in a similar language to the target language while the good translation is the machine translation output of the source sentence in the target language. In the challenge set for *untranslated - full sentences*, the incorrect translation is a copy of the source sentence and the good translation is machine translation output in the target language. Multilingual embeddings learn cross-lingual representations by reducing the language-specific properties during pretraining (Wu and Dredze, 2019). We hypothesised that making representations language agnostic may harm MT evaluation in cases where translations are extremely poor, such that they remain untranslated or hallucinate from a similar language.

In Figure 3 we plot the correlations for all reference-based and reference-free metrics. Overall, we observe that several metrics from 2023 have much better correlation scores than 2022 indicating that newer models have developed strategies to avoid learning language-agnostic representations. In particular, we find that many of the reference-free metrics submitted to the 2023 shared task have improved on the *untranslated - full sentences* category (though a few reference-free metrics from 2022 had performance closer to 1, which is not the case with the 2023 metrics). This is a welcome change as we expect these metrics to perform a more faithful evaluation when many of the words remain untranslated in the hypothesis, especially in the lower resource setting. Whilst some reference-free metrics struggle considerably on this challenge set and almost always prefer the copied source to the real translation, reference-based metrics generally exhibit good correlation i.e. they can identify the copied source quite easily. As reference-based metrics tend to ignore the source, the scores are likely based on the similarity between the reference and the MT output. This is evident from their poor

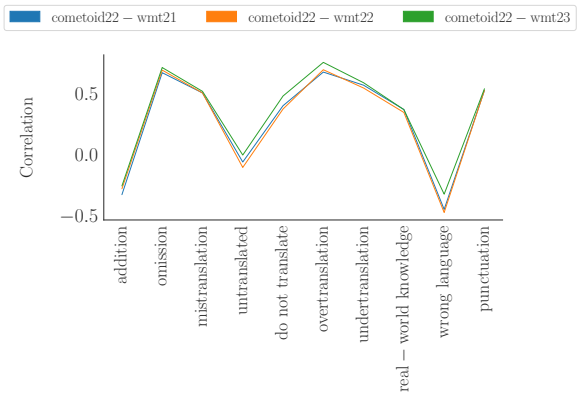

Figure 4: Correlations for different top-level phenomena categories with different models trained on successively more data.

performance on the *wrong - language* category (see Table 3). This suggests that language-agnostic representations present in the multilingual space may harm performance.

5.4 Training data size effects

One submission this year, namely COMETOID22, submitted three different reference-free metric versions, each trained on successively more data. This allows us to investigate the effects of the metric training data size[5] on the performance on ACES. (Note that we cannot draw any conclusions about the training data size of the pretraining models that are used.) In Figure 4 we can see the effect of training data size on the performance on the top-level phenomena categories. COMETOID22-WMT23, the model that has seen the most data, outperforms the other two metrics on almost all top-level categories. The correlation gain is especially pronounced for the *untranslated*, *do not translate* (content in the source is erroneously translated into the target language), *overtranslation* (the target translation contains more specific information than the source) and *wrong language* categories (see Table 9 for examples for each of the phenomena). For clearer insights as to where the performance gain comes from, we would need to analyse the training data in depth. However, it is evident from these results that more training data is beneficial for metric development. In the next section, we look at metric score changes over metric implementation cycles - where likely more than just the training data changed.

[5]Note that for COMETOID22 this is not human judgement labelled data but rather pseudo labelled data where labels come from the reference-based COMET-22 model.

5.5 Changes over one year

We compare the results of metrics submitted by the same teams last year and this year in Table 8.

We report changes in performance in terms of deltas, computed by subtracting the 2022 score from the 2023 score. We do this for the following pairs of metrics: KG-BERTSCORE (2022) and KG-BERTSCORE (2023); COMETKIWI (2022) paired with COMETKIWI-XL (2023) and COMETKIWI-XXL (2023); COMET-22 (2022) paired with XCOMET-ENSEMBLE (2023), XCOMET-XL (2023) and XCOMET-XXL (2023).

We observe that KG-BERTSCORE has improved over its performance of last year. From the description provided by the metric developers, the main difference is that the 2023 version of KG-BERTSCORE metric uses COMET-QE instead of BERTScore (Zhang et al., 2020) to compute the similarity between the source and the hypothesis. Whilst we might therefore attribute the increase in performance to this change, a more systematic comparison of the two metric versions would be required to confirm whether this is the only contributing factor.

The metrics in the COMETKIWI family exhibit: a slight drop in performance (COMETKIWI-XL) and a similar performance to that of last year (COMETKIWI-XXL). The difference can be attributed to changing the underlying encoder, XLM-R XL and XLM-R XXL (Goyal et al., 2021) respectively, and the use of additional fine-tuning data made available this year. We have seen that the addition of more training data helps in Section 5.4. Considering that there is no improvement in the performance, we question if an increase in the underlying model capacity of the encoder alone is useful for obtaining better MT evaluation.

Performance change for the XCOMET family is variable: there is a performance increase for XCOMET-ENSEMBLE (compared to COMET-22), for XCOMET-XL the increase is smaller, and the performance of XCOMET-XXL is degraded. The XCOMET family is designed to provide both a quality score and an error span. Considering that the metric also provides an explanation of the scores without hurting the performance, this is indeed a positive change. Finally, it is worth noting that for *all* metrics in Table 8 a change in performance is observed for almost all ACES categories, for all metrics.

	COMETKiwi		KG-BERTScore	XCOMET		
	-XL	-XXL		-Ensemble	-XL	-XXL
addition	-0.120	-0.004	-0.251	0.595	0.455	0.142
omission	-0.004	-0.002	0.103	0.118	-0.126	-0.254
mistranslation	-0.005	0.013	0.077	0.126	0.038	0.005
untranslated	0.000	0.142	0.266	-0.181	-0.342	-0.362
do not translate	-0.395	-0.553	0.000	0.053	0.079	-0.105
overtranslation	0.027	0.035	0.119	0.073	-0.067	0.017
undertranslation	-0.019	-0.021	0.077	0.014	-0.132	-0.025
real-world knowledge	-0.020	0.100	0.107	0.003	-0.123	-0.198
wrong language	-0.014	-0.173	-0.618	-0.296	-0.232	-0.395
punctuation	-0.037	0.004	0.264	0.206	-0.144	0.006
ACES-Score	-1.04	-0.38	0.40	4.23	0.21	-1.64

Table 8: Comparison of average Kendall's tau-like correlation: delta calculated as 2023 score minus 2022 score.

Whilst it is not possible to draw conclusions or make predictions about the future of metric development based solely on the observations from two consecutive metrics shared tasks, we highlight several high-level changes. Firstly, we note the participation of many more COMET-based metrics in 2023, compared with 2022. This is presumably based on the success of COMET at previous shared tasks and its adoption within the MT community. We find that three metrics from 2022 are now used as baseline metrics namely COMET-22, COMETKIWI, and MS-COMET-QE-22. In contrast to the submissions in 2022, we find some new metrics that use lexical overlap through text matching or embeddings (TOKENGRAM_F, PARTOKENGRAM_F, and EBLEU). However, their performance trend is similar to other surface overlap metrics. This year has also seen submissions based on large language models (EMBED_LLAMA and GEMBA-MQM). As seen in Section 2, their moderate performance indicates the need for more effective approaches. Additionally, we note an overall increase between 2022 and 2023 in the number of metrics submitted to WMT that a) provide segment-level scores and b) provide scores for all language pairs and directions in ACES. There were 37 segment-level metrics at WMT 2022, 24 of which covered the language pairs and directions in ACES, compared with 47 and 33, respectively in 2023. This suggests that the interest in metric development remains high, and could be increasing.

From our analyses in Section 5.3, we also draw similar conclusions to Amrhein et al. (2022) with the exception of reference-free metrics improving at the *Untranslated - Full Sentences* task. Despite the success of LLMs across various tasks (Brown et al., 2020), leveraging them to evaluate translated outputs still requires some improved design strategies. All these observations suggest that evaluating MT outputs is indeed a hard problem (Neubig, 2022). While we do have a good suite of metrics to provide a proxy for evaluation, there are indeed several interesting challenges that need to be tackled before we find an ideal evaluation regime. And even then, we need to continuously monitor this to ensure that we do not optimise towards metric weaknesses that we have not yet discovered.

5.6 Recommendations

We provide the same recommendations as last year:

No metric to rule them all. There is no consistent winning metric across all categories (see Table 3). We recommend developing evaluation methods that combine different design strategies for robust evaluation. We also recommend innovation in the ensemble building as simple strategies like majority voting do not lead to significant improvement (Moghe et al., 2023). We find that some submissions in this year's shared task already contain ensembles (XCOMET-ENSEMBLE, XCOMET-QE-ENSEMBLE) which suggests that our recommendations are in line with the efforts of the community.

The source matters. The trend where reference-based metrics tend to disregard information in the source is also persistent, as seen in Section 5.3. We also observe that reference-free metrics are highly competitive with reference-based metrics as seen in Table 3 and also in Freitag et al. (2022); Zerva et al. (2022), *inter alia*. Furthermore, as references are often not perfect themselves (Freitag et al., 2020), it is ideal to develop evaluation regimens that focus more on the information in the source sentence than

the references.

Surface overlap still prevails. Neural metrics were introduced to overcome surface-level overlap present in the string-based metrics. However, the results in Section 5.3.2 suggest that neural metrics tend to focus more on lexical overlap than semantic content. We thus recommend including paraphrases in the training regime as well as designing loss functions that explicitly discourage surface-level overlap.
Lastly, simple strategies to model language-specific information in the metrics could also improve the robustness of the metrics to language pair attacks.

6 Conclusion

We re-submitted the ACES Challenge Set to WMT2023 to identify the strengths and weaknesses of the metrics submitted to this year's shared task. Overall, we find similar trends to that of last year. While neural metrics tend to be better, different categories of metrics have different strengths, and we do not find one clear winner. With respect to the metrics that were resubmitted with some design changes, we find that these design changes have variable outcomes with a performance drop in some cases. The major challenges of (i) metrics not paying enough attention to the source, (ii) reference-based metrics still relying on surface-level overlap, and (iii) over-reliance on multilingual embeddings still persist. Hence, our recommendations are also similar to that of last year: build ensembles of different design families, encourage development that better utilises information in the source, include diverse training examples to reduce the influence of surface-level overlap, and carefully determine the influence of multilingual embeddings/LLMs on MT evaluation.

Limitations

When comparing the results of the baseline metrics common to the 2022 and 2023 metrics shared tasks, we observed differences in the scores returned for a small subset (2,659; approx 7%) of the ACES examples. A subsequent investigation suggested that differences in the pre-processing steps by the shared task organisers in 2022 and 2023 may have led to the differences; we further conjecture that differences in handling the double quotes present in some of the ACES examples may be one of the main causes. Regardless of the source of the differences, we highlight that care should be taken when pre-processing the ACES dataset prior to benchmarking metric performance, especially when the aim is to draw comparisons with results reported in previous work. However, we note that this issue is not specific to ACES, but may potentially affect any text-based dataset. With the exception of the comparison of results from 2022 and 2023 in Section 5.5, for which we used the subset of 33,817 examples which were unaffected by pre-processing differences, all other results reported in this paper use the full set of 36,476 ACES examples. We also note that ideally, incorrect processing of double quotes by a metric should not lead to a difference in scores especially when dealing with semantic errors.

As we re-submitted the same version of the ACES dataset to WMT 2023, the same biases described in Amrhein et al. (2022) remain: 1) there is greater coverage in terms of phenomena and number of examples for some language pairs (particularly en-de and en-fr), 2) more examples are provided for categories for which examples may be generated automatically, compared to those that required manual construction/filtering, 3) errors present in the datasets used to construct the examples may have propagated through into ACES, 4) the focus of the ACES is on accuracy errors; the inclusion and evaluation of fluency errors remains a direction for future work.

ACES consists of examples that target a range of linguistic phenomena, which are then arranged in a hierarchy of error categories. In order to provide metric profiles over this range of error categories we require segment-level scores. We therefore report only results for those metrics submitted to WMT 2023 that provide segment-level scores; metrics that provide only system-level outputs are excluded. Further, we excluded those metrics that did not provide scores for all of the language pairs in ACES from the results and analyses in this paper.

The 2023 WMT metric shared task evaluated metrics at the paragraph level for English-German. Currently, ACES is not able to capture document-level metric performance. We hope such challenge sets will become available in the near future to be able to track metric improvements beyond the sentence level.

Ethics Statement

As described in Amrhein et al. (2022) some examples within the ACES challenge set exhibit biases.

However, this is necessary in order to expose the limitations of existing metrics. The challenge set is already publicly available.

Acknowledgements

We thank the organisers of the WMT 2023 Metrics task for organising the Challenge Sets shared task, and the shared task participants for scoring our challenge sets with their systems. We thank Nikolay Bogoychev and the anonymous reviewers for their insightful comments and suggestions. This work was supported in part by the UKRI Centre for Doctoral Training in Natural Language Processing, funded by the UKRI (grant EP/S022481/1) and the University of Edinburgh (Moghe), and by the Swiss National Science Foundation (project MUTAMUR; no. 176727) (Amrhein). We also thank Huawei (Moghe) and the RISE Center for Applied AI (Guillou) for their support.

References

Duarte Alves, Ricardo Rei, Ana C Farinha, José G. C. de Souza, and André F. T. Martins. 2022. Robust MT evaluation with sentence-level multilingual augmentation. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 469–478, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Chantal Amrhein, Nikita Moghe, and Liane Guillou. 2022. ACES: Translation accuracy challenge sets for evaluating machine translation metrics. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 479–513, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Eleftherios Avramidis and Vivien Macketanz. 2022. Linguistically motivated evaluation of machine translation metrics based on a challenge set. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 514–529, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *CoRR*, abs/2005.14165.

Xiaoyu Chen, Daimeng Wei, Hengchao Shang, Zongyao Li, Zhanglin Wu, Zhengzhe Yu, Ting Zhu, Mengli Zhu, Ning Xie, Lizhi Lei, Shimin Tao, Hao Yang, and Ying Qin. 2022. Exploring robustness of machine translation metrics: A study of twenty-two automatic metrics in the WMT22 metric task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 530–540, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Markus Freitag, David Grangier, and Isaac Caswell. 2020. BLEU might be guilty but references are not innocent. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 61–71, Online. Association for Computational Linguistics.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André F. T. Martins. 2022. Results of WMT22 metrics shared task: Stop using BLEU – neural metrics are better and more robust. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 46–68, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, George Foster, Alon Lavie, and Ondřej Bojar. 2021. Results of the WMT21 metrics shared task: Evaluating metrics with expert-based human evaluations on TED and news domain. In *Proceedings of the Sixth Conference on Machine Translation*, pages 733–774, Online. Association for Computational Linguistics.

Naman Goyal, Jingfei Du, Myle Ott, Giri Anantharaman, and Alexis Conneau. 2021. Larger-scale transformers for multilingual masked language modeling. *CoRR*, abs/2105.00572.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán,

and Angela Fan. 2022. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Liane Guillou and Christian Hardmeier. 2016. PROTEST: A test suite for evaluating pronouns in machine translation. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 636–643, Portorož, Slovenia. European Language Resources Association (ELRA).

Pierre Isabelle, Colin Cherry, and George Foster. 2017. A challenge set approach to evaluating machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2486–2496, Copenhagen, Denmark. Association for Computational Linguistics.

Marzena Karpinska, Nishant Raj, Katherine Thai, Yixiao Song, Ankita Gupta, and Mohit Iyyer. 2022. DEMETR: Diagnosing evaluation metrics for translation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9540–9561, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Daniel Khashabi, Snigdha Chaturvedi, Michael Roth, Shyam Upadhyay, and Dan Roth. 2018. Looking beyond the surface: A challenge set for reading comprehension over multiple sentences. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 252–262, New Orleans, Louisiana. Association for Computational Linguistics.

Margaret King and Kirsten Falkedal. 1990. Using test suites in evaluation of machine translation systems. In *COLING 1990 Volume 2: Papers presented to the 13th International Conference on Computational Linguistics*.

Tom Kocmi, Hitokazu Matsushita, and Christian Federmann. 2022. MS-COMET: More and better human judgements improve metric performance. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 541–548, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Yitong Li, Trevor Cohn, and Timothy Baldwin. 2017. BIBI system description: Building with CNNs and breaking with deep reinforcement learning. In *Proceedings of the First Workshop on Building Linguistically Generalizable NLP Systems*, pages 27–32, Copenhagen, Denmark. Association for Computational Linguistics.

Chi-kiu Lo. 2019. YiSi - a unified semantic MT quality evaluation and estimation metric for languages with different levels of available resources. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 507–513, Florence, Italy. Association for Computational Linguistics.

Arle Lommel, Aljoscha Burchardt, and Hans Uszkoreit. 2014. Multidimensional quality metrics (mqm): A framework for declaring and describing translation quality metrics. *Tradumàtica: tecnologies de la traducció*, 0:455–463.

Qingyu Lu, Baopu Qiu, Liang Ding, Kanjian Zhang, Tom Kocmi, and Dacheng Tao. 2023. Error analysis prompting enables human-like translation evaluation in large language models: A case study on chatgpt.

Taylor Mahler, Willy Cheung, Micha Elsner, David King, Marie-Catherine de Marneffe, Cory Shain, Symon Stevens-Guille, and Michael White. 2017. Breaking NLP: Using morphosyntax, semantics, pragmatics and world knowledge to fool sentiment analysis systems. In *Proceedings of the First Workshop on Building Linguistically Generalizable NLP Systems*, pages 33–39, Copenhagen, Denmark. Association for Computational Linguistics.

Richard T McCoy and Tal Linzen. 2019. Non-entailed subsequences as a challenge for natural language inference. *Proceedings of the Society for Computation in Linguistics (SCiL)*, pages 358–360.

Nikita Moghe, Tom Sherborne, Mark Steedman, and Alexandra Birch. 2023. Extrinsic evaluation of machine translation metrics. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13060–13078, Toronto, Canada. Association for Computational Linguistics.

Graham Neubig. 2022. Is my nlp model working? the answer is harder than you think.

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 of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

Maja Popović and Sheila Castilho. 2019. Challenge test sets for MT evaluation. In *Proceedings of Machine Translation Summit XVII: Tutorial Abstracts*, Dublin, Ireland. European Association for Machine Translation.

Abhilasha Ravichander, Siddharth Dalmia, Maria Ryskina, Florian Metze, Eduard Hovy, and Alan W Black. 2021. NoiseQA: Challenge set evaluation for user-centric question answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2976–2992, Online. Association for Computational Linguistics.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Guido Rocchietti, Flavia Achena, Giuseppe Marziano, Sara Salaris, and Alessandro Lenci. 2021. Fancy: A diagnostic data-set for nli models. In *Proceedings of the Eighth Italian Conference on Computational Linguistics (CLiC-it)*.

Thibault Sellam, Amy Pu, Hyung Won Chung, Sebastian Gehrmann, Qijun Tan, Markus Freitag, Dipanjan Das, and Ankur Parikh. 2020. Learning to evaluate translation beyond English: BLEURT submissions to the WMT metrics 2020 shared task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 921–927, Online. Association for Computational Linguistics.

Ieva Staliūnaitė and Ben Bonfil. 2017. Breaking sentiment analysis of movie reviews. In *Proceedings of the First Workshop on Building Linguistically Generalizable NLP Systems*, pages 61–64, Copenhagen, Denmark. Association for Computational Linguistics.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 833–844, Hong Kong, China. Association for Computational Linguistics.

Yinfei Yang, Yuan Zhang, Chris Tar, and Jason Baldridge. 2019. PAWS-X: A cross-lingual adversarial dataset for paraphrase identification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3687–3692, Hong Kong, China. Association for Computational Linguistics.

Chrysoula Zerva, Frédéric Blain, Ricardo Rei, Piyawat Lertvittayakumjorn, José G. C. de Souza, Steffen Eger, Diptesh Kanojia, Duarte Alves, Constantin Orăsan, Marina Fomicheva, André F. T. Martins, and Lucia Specia. 2022. Findings of the WMT 2022 shared task on quality estimation. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 69–99, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

A Examples from ACES

We shall now list one example from every top-level category in Table 9. We reuse most of the examples mentioned in the original paper under the respective categories.

	Addition *target includes content not present in the source*
SRC (de):	In den letzten 20 Jahren ist die Auswahl in Uptown Charlotte exponentiell gewachsen.
REF (en):	In the past 20 years, the amount in Uptown Charlotte has grown exponentially.
✓:	Over the past 20 years, the selection in Uptown Charlotte has grown exponentially.
✗:	Over the past 20 years, the selection of **child-friendly options** in Uptown Charlotte has grown exponentially.
	Omission *errors where content is missing from the translation that is present in the source*
SRC (fr):	Une tornade est un tourbillon d'air à basse-pression en forme de colonne, l'air alentour est aspiré vers l'intérieur et le haut.
REF (en):	A tornado is a **spinning column** of very low-pressure air, which sucks the surrounding air inward and upward.
✓:	A tornado is a **column-shaped** low-pressure air turbine, the air around it is sucked inside and up.
✗:	A tornado is a low-pressure air turbine, the air around it is sucked inside and up.
	Untranslated - Word Level *errors occurring when a text segment that was intended for translation is left untranslated in the target content*
SRC (fr):	À l'origine, l'émission mettait en scène des **comédiens de doublage** amateurs, originaires de l'est du Texas.
REF (de):	Die Sendung hatte ursprünglich lokale Amateur**synchronsprecher** aus Ost-Texas.
✓ (copy):	Ursprünglich spielte die Show mit Amateur**synchronsprechern** aus dem Osten von Texas.
✓ (syn.):	Ursprünglich spielte die Show mit Amateur-**Synchron-Schauspielern** aus dem Osten von Texas.
✗:	Ursprünglich spielte die Show mit Amateur-**Doubling-Schauspielern** aus dem Osten von Texas.
	Mistranslation - Ambiguous Translation *an unambiguous source text is translated ambiguously*
SRC (de):	Der Manager feuerte **die** Bäcker**in**.
REF (en):	The manager fired the baker.
✓:	The manager fired the **female** baker.
✗:	The manager fired the **male** baker.
	Do Not Translate *content in the source that should be copied to the output in the source language, but was mistakenly translated into the target language.*
SRC (en):	Dance was one of the inspirations for the exodus - song **"The Toxic Waltz"**, from their 1989 album "Fabulous Disaster".
REF (de):	Dance war eine der Inspirationen für das Exodus-Lied **„The Toxic Waltz"** von ihrem 1989er Album „Fabulous Disaster".
✓:	Der Tanz war eine der Inspirationen für den Exodus-Song **„The Toxic Waltz"**, von ihrem 1989er Album „Fabulous Disaster".
✗:	Der Tanz war eine der Inspirationen für den Exodus-Song **„Der Toxische Walzer"**, von ihrem 1989er Album „Fabulous Disaster".
	Undertranslation *erroneous translation has a meaning that is more generic than the source*
SRC (de):	Bob und Ted waren Brüder. Ted ist der **Sohn** von John.
REF (en):	Bob and Ted were brothers. Ted is John's **son**.
✓:	Bob and Ted were brothers, and Ted is John's **son**.
✗:	Bob and Ted were brothers. Ted is John's **male offspring**.
	Overtranslation *erroneous translation has a meaning that is more specific than the source*
SRC (ja):	その 40 分の映画はアノーがアラン・ゴダードと協力して脚本を書いた。
REF (en):	The 40-minute **film** was written by Annaud with Alain Godard.
✓:	The 40-minute **film** was written by Annaud along with Alain Godard.
✗:	he 40-minute **cinema verite** was written by Annaud with Alain Godard.
	Real-world Knowledge - Textual Entailment *meaning of the source/reference is entailed by the "good" translation*
SRC (de):	Ein Mann **wurde ermordet**.
REF (en):	A man **was murdered**.
✓:	A man **died**.
✗:	A man **was attacked**.
	Wrong Language *incorrect translation is a perfect translation in a related language*
SRC (en):	Cell comes from the Latin word cella which means small room.
REF (es):	El término célula deriva de la palabra latina cella, que quiere decir «cuarto pequeño».
✓ (es):	La célula viene de la palabra latina cella que significa habitación pequeña.
✗ (ca):	Cèl·lula ve de la paraula llatina cella, que vol dir habitació petita.

Table 9: Examples from each top-level accuracy error category in ACES. An example consists of a source sentence (SRC), reference (REF), good (✓) and incorrect (✗) translations, language pair, and a phenomenon label. We also provide a description of the relevant phenomenon. en: English, de: German, fr: French, ja: Japanese, es: Spanish, ca: Catalan