Extrinsic Evaluation of Machine Translation Metrics

Extrinsic Evaluation of Machine Translation Metrics

Nikita Moghe and **Tom Sherborne** and **Mark Steedman** and **Alexandra Birch**
School of Informatics, University of Edinburgh
{nikita.moghe, tom.sherborne, a.birch}@ed.ac.uk , steedman@inf.ed.ac.uk

Abstract

Automatic machine translation (MT) metrics are widely used to distinguish the quality of machine translation systems across large test sets (i.e., system-level evaluation). However, it is unclear if automatic metrics can reliably distinguish good translations from bad at the sentence level (i.e., segment-level evaluation). We investigate how useful MT metrics are at detecting segment-level quality by correlating metrics with the translation utility for downstream tasks. We evaluate the segment-level performance of widespread MT metrics (chrF, COMET, BERTScore, *etc.*) on three downstream cross-lingual tasks (dialogue state tracking, question answering, and semantic parsing). For each task, we have access to a monolingual task-specific model and a translation model. We calculate the correlation between the metric's ability to predict a good/bad translation with the success/failure on the final task for machine-translated test sentences. Our experiments demonstrate that all metrics exhibit negligible correlation with the extrinsic evaluation of downstream outcomes. We also find that the scores provided by neural metrics are not interpretable, in large part due to having undefined ranges. We synthesise our analysis into recommendations for future MT metrics to produce labels rather than scores for more informative interaction between machine translation and multilingual language understanding.

1 Introduction

Although machine translation (MT) is typically seen as a standalone application, in recent years MT models have been more frequently deployed as a component of a complex NLP platform delivering multilingual capabilities such as cross-lingual information retrieval (Zhang et al., 2022) or automated multilingual customer support (Gerz et al., 2021). When an erroneous translation is generated by the MT systems, it may add new errors in the task pipeline leading to task failure and poor user experience. For example, consider the user's request in Chinese 剑桥有牙买加菜吗？ ("*Is there any good Jamaican food in Cambridge?*") machine-translated into English as *"Does Cambridge have a good meal in Jamaica?"*. The model will erroneously consider "Jamaica" as a location, instead of cuisine, and prompt the search engine to look up restaurants in Jamaica [1]. To avoid this *breakdown*, it is crucial to detect an incorrect translation before it causes further errors in the task pipeline.

One way to approach this *breakdown detection* is using segment-level scores provided by MT metrics. Recent MT metrics have demonstrated high correlation with human judgements at the system level for some language pairs (Ma et al., 2019). These metrics are potentially capable of identifying subtle differences between MT systems that emerge over a relatively large test corpus. These metrics are also evaluated on respective correlation with human judgements at the segment level, however, there is a considerable performance penalty (Ma et al., 2019; Freitag et al., 2021b). Segment-level evaluation of MT is indeed more difficult and even humans have low inter-annotator agreement on this task (Popović, 2021). Despite MT systems being a crucial intermediate step in several applications, characterising the behaviour of these metrics under task-oriented evaluation has not been explored.

In this work, we provide a complementary evaluation of MT metrics. We focus on the segment-level performance of metrics, and we evaluate their performance extrinsically, by correlating each with the outcome of downstream tasks with respective, reliable accuracy metrics. We assume access to a parallel task-oriented dataset, a task-specific monolingual model, and a translation model that can translate from the target language into the language of the monolingual model. We consider the *Translate-Test* setting — where at test time, the examples from the test language are translated to the

[1] Example from the Multi^{2}WoZ dataset (Hung et al., 2022)

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 13060–13078

task language for evaluation. We use the outcomes of this extrinsic task to construct a breakdown detection benchmark for the metrics.

We use dialogue state tracking, semantic parsing, and extractive question answering as our extrinsic tasks. We evaluate nine metrics consisting of string overlap metrics, embedding-based metrics, and metrics trained using scores from human evaluation of MT. Surprisingly, we find our setup challenging for all existing metrics; demonstrating poor capability in discerning good and bad translations across tasks. We present a comprehensive analysis of the failure of the metrics through quantitative and qualitative evaluation.

Our contributions are summarised as follows:
1) We derive a new **breakdown detection task**, for evaluating MT metrics, measuring how indicative segment-level scores are for downstream performance of an extrinsic cross-lingual task (Section 3). We evaluate nine metrics on three extrinsic tasks covering 39 unique language pairs. The task outputs, the breakdown detection labels, and metric outputs are publicly available. [2]
2) We show that segment-level scores, from these metrics, have **minimal correlation with extrinsic task performance** (Section 4.1). Our results indicate that these scores are uninformative at the segment level (Section 4.3) — clearly demonstrating a serious deficiency in the best contemporary MT metrics. In addition, we find variable task sensitivity to different MT errors (Section 4.2).
3) We propose **recommendations** on developing MT metrics to produce useful segment-level output by predicting labels instead of scores and suggest reusing existing post-editing datasets and explicit error annotations (See Section 5).

2 Related Work

Evaluation of machine translation has been of great research interest across different communities (Nakazawa et al., 2022; Fomicheva et al., 2021). Notably, the Conference on Machine Translation (WMT) has been organising annual shared tasks on automatic MT evaluation since 2006 (Koehn and Monz, 2006; Freitag et al., 2021b) that invites metric developers to evaluate their methods on outputs of several MT systems. Metric evaluation typically includes a correlation of the scores with human judgements collected for the respective translation outputs. But, designing such guidelines is challenging (Mathur et al., 2020a), leading to the development of several different methodologies and analyses over the years.

The human evaluation protocols include general guidelines for fluency, adequacy and/or comprehensibility (White et al., 1994) on continuous scales (Koehn and Monz, 2006; Graham et al., 2013) (direct assessments) or fine-grained annotations of MT errors (Freitag et al., 2021a,b) based on error ontology like Multidimensional Quality Metrics (MQM) (Lommel et al., 2014) or rank outputs from different MT systems for the same input (Vilar et al., 2007). Furthermore, the best way to compare MT scores with their corresponding judgements is also an open question (Callison-Burch et al., 2006; Bojar et al., 2014, 2017). The new metrics claim their effectiveness by comparing their performance with competitive metrics on the latest benchmark.

The progress and criticism of MT evaluation are generally documented in a metrics shared task overview (Callison-Burch et al., 2007). For example, Stanojević et al. (2015) highlighted the effectiveness of neural embedding-based metrics; Ma et al. (2019) show that metrics struggle on segment-level performance despite achieving impressive system-level correlation; Mathur et al. (2020b) investigate how different metrics behave under different domains. In addition to these overviews, Mathur et al. (2020a) show that meta-evaluation regimes were sensitive to outliers and minimal changes in evaluation metrics are insufficient to claim metric efficacy. Kocmi et al. (2021) conducted a comprehensive evaluation effort to identify which metric is best suited for pairwise ranking of MT systems. Guillou and Hardmeier (2018) look at a specific phenomenon of whether metrics are capable of evaluating translations involving pronominal anaphora. Recent works have also criticised individual metrics such as COMET (Amrhein and Sennrich, 2022) and BERTScore (Hanna and Bojar, 2021).

These works draw their conclusions based on some comparison with human judgement or on specific pitfalls of individual metrics. Our work focuses on the usability of the metrics as solely judged on their ability to predict downstream tasks where MT is an intermediate step (with a primary emphasis on segment-level performance). Task-based evaluation has been well studied (Jones and Galliers (1996); Laoudi et al. (2006); Zhang et al.

[2] https://huggingface.co/datasets/uoe-nlp/extrinsic_mt_eval

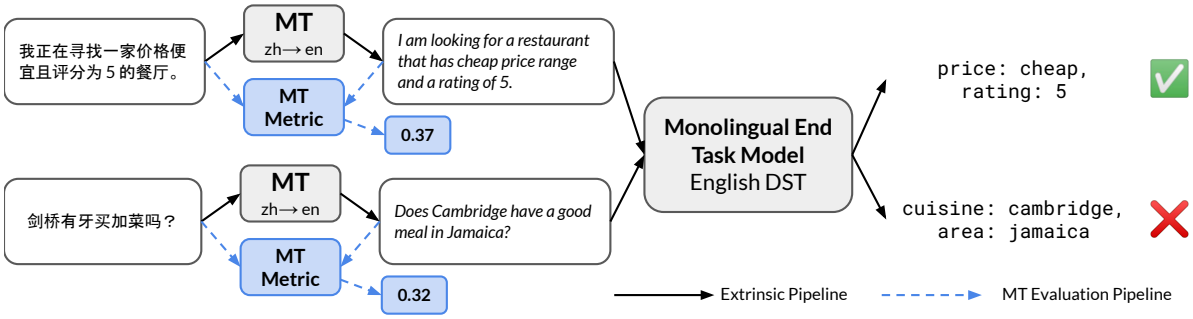

Figure 1: The meta-evaluation pipeline. The predictions for the extrinsic task in the test language (Chinese, ZH) are obtained using the *Translate-Test* setup — the test language is translated into the task language (English, EN) before passing to the task-specific model. The input sentence (ZH) and the corresponding translations (EN) are evaluated with a metric of interest. The metric is evaluated based on the correlation of its scores with the predictions of the end task.

(2022), *inter alia*) but limited to evaluating MT systems rather than MT metrics. Closer to our work is Scarton et al. (2019); Zouhar et al. (2021) which proposes MT evaluation as ranking translations based on the time to post-edit model outputs. We borrow the term of *breakdown detection* from Martinovski and Traum (2003) that proposes breakdown detection for dialogue systems to detect unnatural responses.

3 Methodology

Our aim is to determine how reliable MT metrics are for predicting success on downstream tasks. Our setup uses a monolingual model (e.g., a dialogue state tracker) trained on a *task language* and parallel test data from multiple languages. We use MT to translate a test sentence (from a *test language* to the *task language*) and then infer a label for this example using the monolingual model. If the model predicts a correct label for the parallel *task language* input but an incorrect label for the translated *test language* input, then we have observed a *breakdown* due to a material error in the translation pipeline. We then study if the metric could predict if the translation is suitable for the end task. We refer to Figure 1 for an illustration. We frequently use the terms *test language* and *task language* to avoid confusion with the usage of *source language* and *target language* in the traditional machine translation setup. In Figure 1, the task language is English and the test language is Chinese. We now describe our evaluation setup and the metrics under investigation.

3.1 Setup

For all the tasks described below, we first train a model for the respective tasks on the monolingual setup. We evaluate the task language examples on each task and capture the monolingual predictions of the model. We consider the *Translate-Test* paradigm (Hu et al., 2020), we translate the examples from each test language into the task language. The generated translations are then fed to the task-specific monolingual model. We use either (i) OPUS translation models (Tiedemann and Thottingal, 2020), (ii) M2M100 translation (Fan et al., 2021) or (iii) translations provided by the authors of respective datasets. Note that the examples across all the languages are parallel and we therefore always have access to the correct label for a translated sentence. We obtain the predictions for the translated data to construct a breakdown detection benchmark for the metrics.

We consider only the subset of examples in the test language which were correctly predicted in the task language to avoid errors that arise from extrinsic task complexity. Therefore, all incorrect extrinsic predictions for the test language in our setup arise from erroneous translation. This isolates the extrinsic task failure as the fault of *only* the MT system. We use these predictions to build a binary classification benchmark—all target language examples that are correctly predicted in the extrinsic task receive a positive label (no breakdown) while the incorrect predictions receive a negative label (breakdown).

We consider the example from the test language as *source*, the corresponding machine translation as *hypothesis* and the human reference from the task language as *reference*. Thus, in Figure 1, the source is 剑桥有牙买加菜吗？, the hypothesis is “Does Cambridge have a good meal in Jamaica", and the reference will be “Is there any good Jamaican food in Cambridge". These triples are then

scored by the respective metrics. After obtaining the segment-level scores for these triples, we define a threshold for the scores, thus turning metrics into classifiers. For example, if the threshold for the metric in Figure 1 is 0.5, it would mark both examples as bad translations. We plot a histogram over the scores with ten bins for every setup and select the interval with the highest performance on the development set as a threshold. The metrics are then evaluated on how well their predictions for a good/bad translation correlate with the breakdown detection labels.

3.2 Tasks

We choose tasks that contain outcomes belonging to a small set of labels, unlike natural language generation tasks which have a large solution space. This discrete nature of the outcomes allows us to quantify the performance of MT metrics based on standard classification metrics. The tasks also include varying types of textual units: utterances, sentences, questions, and paragraphs, allowing a comprehensive evaluation of the metrics.

3.2.1 Semantic Parsing (SP)

Semantic parsing transforms natural language utterances into logical forms to express utterance semantics in some machine-readable language. The original ATIS study (Hemphill et al., 1990) collected questions about flights in the USA with the corresponding SQL to answer respective questions from a relational database. We use the MultiATIS++SQL dataset from Sherborne and Lapata (2022) comprising gold parallel utterances in English, French, Portuguese, Spanish, German and Chinese (from Xu et al. (2020)) paired to executable SQL output logical forms (from Iyer et al. (2017)). The model follows Sherborne and Lapata (2023), as an encoder-decoder Transformer model based on mBART50 (Tang et al., 2021). The parser generates valid SQL queries and performance is measured as exact-match *denotation accuracy*—the proportion of output queries returning identical database results relative to gold SQL queries.

3.2.2 Extractive Question Answering (QA)

The task of extractive question answering is predicting a span of words from a paragraph corresponding to the question. We use the XQuAD dataset (Artetxe et al., 2020) for evaluating extractive question answering. The XQuAD dataset was obtained by professionally translating examples from the development set of English SQuAD dataset (Rajpurkar et al., 2016) into ten languages: Spanish, German, Greek, Russian, Turkish, Arabic, Vietnamese, Thai, Chinese, and Hindi. We use the publicly available question answering model that finetunes RoBERTa (Liu et al., 2019) on the SQuAD training set. We use the *Exact-Match* metric, i.e., the model's predicted answer span exactly matches the gold standard answer span; for the breakdown detection task. The metrics scores are produced for the question and the context. A translation is considered to be faulty if either of the scores falls below the chosen threshold for every metric.

3.2.3 Dialogue State Tracking (DST)

In the dialogue state tracking task, a model needs to map the user's goals and intents in a given conversation to a set of slots and values, known as a *dialogue state*, based on a pre-defined ontology. MultiWoZ 2.1 (Eric et al., 2020) is a popular dataset for examining the progress in dialogue state tracking which consists of multi-turn conversations in English spanning across 7 domains. We consider the Multi^{2}WoZ dataset (Hung et al., 2022) where the development and test set have been professionally translated into German, Russian, Chinese, and Arabic from the MultiWoZ 2.1 dataset. We use the dialogue state tracking model trained on the English dataset by Lee et al. (2019). We consider the *Joint Goal Accuracy* where the inferred label is correct only if the predicted dialogue state is exactly equal to the ground truth to provide labels for the breakdown task. We use oracle dialogue history and the metric scores are produced only for the current utterance spoken by the user.

3.3 Metrics

We describe the metrics based on their design principles: derived from the surface level token overlap, embedding similarity, and neural metrics trained using WMT data. We selected the following metrics as they are the most studied, frequently used, and display a varied mix of design principles.

3.3.1 Surface Level Overlap

BLEU (Papineni et al., 2002) is a string-matching metric that compares the token-level n-grams of the hypothesis with the reference translation. BLEU is computed as a precision score weighted by a brevity penalty. We use sentence-level BLEU in our experiments.

chrF (Popović, 2017) computes a character n-gram

F-score based on the overlap between the hypothesis and the reference.

3.3.2 Embedding Based

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

3.3.3 Trained on WMT Data

WMT organises an annual shared task on developing MT models for several categories in machine translation (Akhbardeh et al., 2021). Human evaluation of the translated outputs from the participating machine translation models is often used to determine the best-performing MT system. In recent years, this human evaluation has followed two protocols: (i) Direct Assessment (DA) (Graham et al., 2013): where the given translation is rated from 0 to 100 based on the perceived translation quality and (ii) Expert based evaluation where the translations are evaluated by professional translators with explicit error listing based on the Multidimensional Quality Metrics (MQM) ontology. MQM ontology consists of a hierarchy of errors and translations are penalised based on the severity of errors in this hierarchy. These human evaluations are then used as training data for building new MT metrics.

COMET metrics: Cross-lingual Optimized Metric for Evaluation of Translation (COMET) (Rei et al., 2020) uses a cross-lingual encoder (XLM-R (Conneau et al., 2020)) and pooling operations to predict score of the given translation. Representations for the source, hypothesis, and reference (obtained using the encoder) are combined and passed through a feedforward layer to predict a score. These metrics use a combination of WMT evaluation data across the years to produce different metrics. In all the variants, the MQM scores and DA scores are normalised to z-scores to reduce the effect of outlier annotations.

COMET-DA uses direct assessments from 2017 to 2019 as training data while **COMET-MQM** uses direct assessments from 2017 to 2021 as training data. This metric is then fine-tuned with MQM data from Freitag et al. (2021a).

UniTE metrics (Wan et al., 2022), Unified Translation Evaluation, is another neural translation metric that proposes a multi-task setup for the three strategies of evaluation: source-hypothesis, source-hypothesis-reference, and reference-hypothesis in a single model. The pre-training stage involves training the model with synthetic data constructed using a subset of WMT evaluation data. Fine-tuning uses novel attention mechanisms and aggregate loss functions to facilitate the multi-task setup.

All the above reference-based metrics have their corresponding reference-free versions which use the same training regimes but exclude encoding the reference. We refer to them as COMET-QE-DA, COMET-QE-MQM, and UniTE-QE respectively. COMET-QE-DA in this work uses DA scores from 2017 to 2020. We list the code sources of these metrics in Appendix B.

3.4 Metric Evaluation

The meta-evaluation for the above metrics uses the breakdown detection benchmark. As the class distribution changes depending on the task and the language pair, we require an evaluation that is robust to class imbalance. We consider using macro-F1 and Matthew's Correlation Coefficient (MCC) (Matthews, 1975) on the classification labels. The range of macro-F1 is from 0 to 1 with equal weight to positive and negative classes. We include MCC to interpret the MT metric's standalone performance for the given extrinsic task. The range of MCC is between -1 to 1. An MCC value near 0 indicates no correlation with the class distribution. Any MCC value between 0 and 0.3 indicates negligible correlation, 0.3 to 0.5 indicates low correlation.

4 Results

We report the aggregated results for semantic parsing, question answering, and dialogue state tracking in Table 1 with fine-grained results in Appendix D. We use a random baseline for comparison which assigns the positive and negative labels with equal probability.

4.1 Performance on Extrinsic Tasks

We find that almost all metrics perform above the random baseline on the macro-F1 metric. We use MCC to identify if this increase in macro-F1 makes the metric usable in the end task. Evaluating MCC, we find that all the metrics show negligible correlation across all three tasks. Contrary to trends where neural metrics are better than metrics based on surface overlap (Freitag et al., 2021b), we find this breakdown detection to be difficult irrespective

Metric	Semantic Parsing		Question Answering		Dialogue State Tracking	
	F1	MCC	F1	MCC	F1	MCC
Random	0.453	-0.034	0.496	0.008	0.493	0.008
BLEU	0.580	0.179	0.548	0.121	0.529	0.082
chrF	0.609	0.234	0.554	0.127	0.508	0.067
BERTScore	0.590	0.205	0.555	0.127	0.505	0.071
COMET-DA	0.606	0.228	0.562	0.137	0.608	0.244
COMET-MQM	0.556	0.132	0.387	0.027	0.597	0.204
UniTE	0.600	0.225	0.375	0.012	0.620	0.262
COMET-QE-DA	0.556	0.135	0.532	0.100	0.561	0.145
COMET-QE-MQM	0.597	0.211	0.457	0.033	0.523	0.094
UniTE-QE	0.567	0.155	0.388	0.032	0.587	0.192
Ensemble	0.620	0.251	0.577	0.168	0.618	0.248

Table 1: Performance of MT metrics on the classification task for extrinsic tasks Parsing (MultiATIS++SQL), Question Answering (XQuad) using an English-trained question answering system, and Dialogue State Tracking (Multi^{2}WoZ) using an English-trained state tracker. Reported Macro F1 scores and MCC scores quantify if the metric detects a breakdown for the extrinsic task. Metrics have a negligible correlation with the outcomes of the end task. MCC and F1 are average over respective language pairs

of the design of the metric. We also evaluate an ensemble with majority voting of the predictions from the top three metrics per task. Ensembling provides minimal gains suggesting that metrics are making similar mistakes despite varying properties of the metrics.

Comparing the reference-based versions of trained metrics (COMET-DA, COMET-MQM, UniTE) with their reference-free quality estimation (QE) equivalents, we observe that reference-based versions perform better, or are competitive to, their reference-free versions for the three tasks. We also note that references are unavailable when the systems are in production, hence reference-based metrics are unsuitable for realistic settings. We discuss alternative ways of obtaining references in Section 4.4.

Between the use of MQM-scores and DA-scores during fine-tuning COMET variants, we find that both COMET-QE-DA and COMET-DA are strictly better than COMET-QE-MQM and COMET-MQM for question answering and dialogue state tracking respectively, with no clear winner for semantic parsing (See Appendix D).

The results on per-language pair in Appendix D suggest that no specific language pairs stand out as easier/harder across tasks. As this performance is already poor, we cannot verify if neural metrics can generalise in evaluating language pairs unseen during training.

Case Study: We look at Semantic Parsing with an English-trained parser tested with Chinese inputs for our case study with the well-studied COMET-DA metric. We report the number of correct and incorrect predictions made by COMET-DA across ten equal ranges of scores in Figure 2. The bars labelled on the x-axis indicate the end-point of the interval i.e., the bar labelled -0.74 contains examples that were given scores between -1.00 and -0.74.

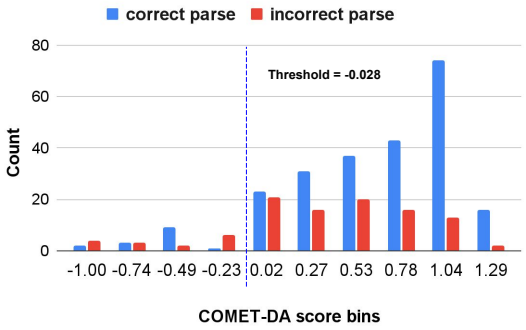

Figure 2: Graph of predictions by COMET-DA (threshold: -0.028), categorised by the metric scores in ten intervals. Task: Semantic Parsing with English parser and test language is Chinese. The bars indicate the count of examples with incorrect parses (red) and correct parses (blue) assigned the scores for the given ranges.

First, we highlight that the threshold is -0.028, counter-intuitively suggesting that even some correct translations receive a negative score. We expected the metric to fail in the regions around the threshold as those represent strongest confusion. For example, "周日下午从迈阿密飞往克利夫兰" is correctly translated as "Sunday afternoon from Miami to Cleveland" yet the metric assigns it a score of -0.1. However, the metric makes mis-

Task	Errors by the Extrinsic model	False Positive	False Negative
SP	25%	mistranslation (90%), omission(10%)	mistranslation (25.7%), fluency (20%), omission (5.7%), no error (48.6%)
QA	20%	mistranslation (60%), omission(8.6%), addition (5.7%), fluency (20%), undertranslation (2.9%), untranslated (2.9%)	mistranslation (18%), fluency (22%), addition (2%), no error (54%)
DST	5%	mistranslation (100%)	omission (26%), mistranslation (1%), no error (73%)

Table 2: The proportion of the different types of errors erroneously detected and undetected by COMET-DA for languages mentioned in Section 4.2. False positives and false negatives are computed by excluding the examples where the extrinsic task model was at fault.

takes throughout the bins. For example, "我需要预订一趟联合航空下周六的从辛辛那提飞往纽约市的航班" is translated as "I need to book a flight from Cincinnati to New York City next Saturday." and loses the crucial information of "United Airlines"; yet it is assigned a high score of 0.51. This demonstrates that the metric possesses a limited perception of a good or bad translation for the end task.

We suspect this behaviour is due to the current framework of MT evaluation. The development of machine translation metrics largely caters towards the intrinsic task of evaluating the quality of a translated text in the target language. The severity of a translation error is dependent on the guidelines released by the organisers of the WMT metrics task or the design choices of the metric developers. Our findings agree with Zhang et al. (2022) that different downstream tasks will demonstrate varying levels of sensitivity to the same machine translation errors.

4.2 Qualitative Evaluation

To quantify detecting which translation errors are most crucial to the respective extrinsic tasks, we conduct a qualitative evaluation of the MT outputs and task predictions. We annotate 50 false positives and 50 false negatives for test languages Chinese (SP), Hindi (QA), and Russian (DST) respectively. The task language is English. We annotate the MT errors (if present) in these examples based on the MQM ontology. We tabulate these results in Table 2 using COMET-DA for these analyses.

Within the false negatives, a majority of the errors (>48%) are due to the metric's inability to detect translations containing synonyms or paraphrases of the references as valid translations. Further, omission errors detected by the metric are not crucial for DST as these translations often exclude pleasantries. Similarly, errors in fluency are not important for both DST and SP but they are crucial for QA as grammatical errors in questions produce incorrect answers. Mistranslation of named entities (NEs), especially which lie in the answer span, is a false negative for QA since QA models find the answer by focusing on the words in the context surrounding the NE rather than the error in that NE. Detecting mistranslation in NEs is crucial for both DST and SP as this error category dominates the false positives. A minor typo of *Lester* instead of *Leicester* marks the wrong location in the dialogue state which is often undetected by the metric. Addition and omission errors are also undetected for SP while mistranslation of reservation times is undetected for DST.

We also find that some of the erroneous predictions can be attributed to the failure of the extrinsic task model than the metric. For example, the MT model uses an alternative term of *direct* instead of *nonstop* while generating the translation for the reference "show me nonstop flights from montreal to orlando". The semantic parser fails to generalise despite being trained with mBART50 to ideally inherit some skill at disambiguiting semantically similar phrases. This error type accounts for 25% for SP, 20% for QA and 5% in DST of the total annotated errors. We give examples in Appendix C.

4.3 Finding the Threshold

Interpreting system-level scores provided by automatic metrics requires additional context such as the language pair of the machine translation model or another MT system for comparison [3]. In this classification setup, we rely on interpreting the segment-level score to determine whether the translation is suitable for the downstream task. We find that choosing the right threshold to identify trans-

[3] https://github.com/Unbabel/COMET/issues/18

Extrinsic Task	SP	QA	DST
BLEU	15.5 ± 08.8	16.1 ± 04.9	20.0 ± 0.00
chrF	44.0 ± 13.7	53.9 ± 07.8	30.7 ± 0.45
BERTScore	0.50 ± 0.21	0.54 ± 0.08	0.39 ± 0.21
COMET-DA	0.21 ± 0.35	0.30 ± 0.23	0.58 ± 0.08
COMET-MQM	0.03 ± 0.01	0.06 ± 0.01	0.02 ± 0.00
UniTE	0.04 ± 0.22	-0.40 ± 0.38	-0.01 ± 0.29
COMET-QE-DA	0.02 ± 0.07	0.02 ± 0.01	0.06 ± 0.01
COMET-QE-MQM	0.11 ± 0.01	0.00 ± 0.04	0.03 ± 0.00
UniTE-QE	-0.01 ± 0.22	-0.24 ± 0.13	0.11 ± 0.18

Table 3: Mean and Standard Deviation of the best threshold on the development set for all the language pairs in the respective extrinsic tasks. The thresholds are inconsistent across language pairs and tasks for both bounded and unbounded metrics.

lations requiring correction is not straightforward. Our current method to obtain a threshold relies on validating candidate thresholds on the development set and selecting an option with the best F1 score. These different thresholds are obtained by plotting a histogram of scores with ten bins per task and language pair.

We report the mean and standard deviation of best thresholds for every language pair for every metric in Table 3. Surprisingly, the thresholds are inconsistent and biased for bounded metrics: BLEU (0–100), chrF (0–100), and BERTScore (0–1). The standard deviations across the table indicate that the threshold varies greatly across language pairs. We find that thresholds of these metrics are also not transferable across tasks. COMET metrics, except COMET-DA, have lower standard deviations. By design, the range of COMET metrics in this work is unbounded. However, as discussed in the theoretical range of COMET metrics [4], empirically, the range for COMET-MQM lies between -0.2 to 0.2, questioning whether lower standard deviation is an indicator of threshold consistency. Some language pairs within the COMET metrics have negative thresholds. We also find that some of the use cases under the UniTE metrics have a mean negative threshold, indicating that good translations can have negative UniTE scores. Similar to Marie (2022), we suggest that the notion of negative scores for good translations, only for certain language pairs, is counter-intuitive as most NLP metrics tend to produce positive scores.

Thus, we find that both bounded and unbounded metrics discussed here do not provide segment-level scores whose range can be interpreted meaningfully across tasks and language pairs.

[4] https://unbabel.github.io/COMET/html/faqs.html

Metric	SP	QA	DST
BLEU	0.003	0.013	0.050
chrF	0.018	0.021	0.055
BERTScore	0.028	0.065	0.036
COMET-DA	0.071	0.085	0.083
COMET-MQM	0.080	0.019	0.116
UniTE	0.225	0.056	0.193

Table 4: MCC scores of reference based metrics with pseudo references when gold references are unavailable at test time. Performance is worse than metrics with oracle references and reference-free metrics (Table 1)

4.4 Reference-based Metrics in an Online Setting

In an online setting, we do not have access to references at test time. To test the effectiveness of reference-based methods here, we consider translating the translation back into the test language. For example, for an *en* parser, the test language ti_{zh} is translated into mt_{en} and then translated back to Chinese as mt_{zh}. The metrics now consider mt_{en} as source, mt_{zh} as hypothesis, and ti_{zh} as the reference. We generate these new translations using the mBART50 translation model (Tang et al., 2021) and report the results in Table 4.

Compared to the results in Table 1, there is a further drop in performance across all the tasks and metrics. The metrics also perform worse than their reference-free counterparts. The second translation is likely to add additional errors to the existing translation. This cascading of errors confuses the metric and it can mark a perfectly useful translation as a breakdown. The only exception is that of the UniTE metric which has comparable performance (but overall poor) due to its multi-task setup.

5 Recommendations

Our experiments suggest that evaluating MT metrics on the segment level for extrinsic tasks has considerable room for improvement. We propose recommendations based on our observations:

Prefer MQM for Human Evaluation of MT outputs: We reinforce the proposal of using the MQM scoring scheme with expert annotators for evaluating MT outputs in line with Freitag et al. (2021a). As seen in Section 4.2, different tasks have varying tolerance to different MT errors. With explicit errors marked per MT output, future classifiers can be trained on a subset of human evaluation

data containing errors most relevant to the downstream application.

MT Metrics Could Produce Labels over Scores: The observations from Section 4.2 and Section 4.3 suggest that interpreting the quality of the produced MT translation based on a number is unreliable and difficult. We recommend exploring whether segment-level MT evaluation can be approached as an error classification task instead of regression. Specifically, whether the words in the source/hypothesis can be tagged with explicit error labels. Resorting to MQM-like human evaluation will result in a rich repository of human evaluation based on an ontology of errors and erroneous spans marked across the source and hypothesis (Freitag et al., 2021a). Similarly, the post-editing datasets (Scarton et al. (2019); Fomicheva et al. (2022) , *inter alia*) also provide a starting point. An interesting exploration in this direction are the works by Perrella et al. (2022); Rei et al. (2022) that treat MT evaluation as a sequence-tagging problem by labelling the errors in an example. Such metrics can also be used for intrinsic evaluation by assigning weights to the labels and producing a weighted score.

Add Diverse References During Training: From Section 4.2, we find that both the neural metric and the task-specific model are not robust to paraphrases. We also recommend the inclusion of diverse references through automatic paraphrasing (Bawden et al., 2020) or data augmentation during the training of neural metrics.

6 Conclusion

We propose a method for evaluating MT metrics which is reliable at the segment-level and does not depend on human judgements by using correlation MT metrics with the success of extrinsic downstream tasks. We evaluated nine different metrics on the ability to detect errors in generated translations when machine translation is used as an intermediate step for three extrinsic tasks: Semantic Parsing, Question Answering, and Dialogue State Tracking. We find that segment-level scores provided by all the metrics show negligible correlation with the success/failure outcomes of the end task across different language pairs. We attribute this result to segment scores produced by these metrics being uninformative and that different extrinsic tasks demonstrate different levels of sensitivity to different MT errors. We propose recommendations to predict error types instead of error scores to facilitate the use of MT metrics in downstream tasks.

7 Limitations

As seen in Section 4.2, sometimes the metrics are unnecessarily penalised due to errors made by the end task models. Filtering these cases would require checking every example in every task manually. We hope our results can provide conclusive trends to the metric developers focusing on segment-level MT evaluation.

We included three tasks to cover different types of errors in machine translations and different types of contexts in which an online MT metric is required. Naturally, this regime can be extended to other datasets, other tasks, and other languages (Ruder et al., 2021; Doddapaneni et al., 2022). Further, our tasks used stricter evaluation metrics such as exact match. Incorporating information from partially correct outputs is not trivial and will be hopefully addressed in the future. We have covered 37 language pairs across the tasks which majorly use English as one of the languages. Most of the language pairs in this study are high-resource languages. Similarly, the examples in multilingual datasets are likely to exhibit *translationese* - unnatural artefacts from the task language present in the test language during manual translation; which tend to overestimate the performance of the various tasks (Majewska et al., 2023; Freitag et al., 2020). We hope to explore the effect of translationese on MT evaluation (Graham et al., 2020) and extrinsic tasks in future. The choice of metrics in this work is not exhaustive and is dependent on the availability and ease of use of the metric provided by the authors.

8 Ethics Statement

This work uses datasets, models, and metrics that are publicly available. Although the scope of this work does not allow us to have an in-depth discussion of biases associated with metrics (Amrhein et al., 2022), we caution the readers of drawbacks of metrics that cause unfair evaluation to marginalised subpopulations which are discovered or yet to be discovered. We will release the translations, metrics scores, and corresponding task outputs for reproducibility.

9 Acknowledgements

We thank Barry Haddow for providing us with valuable feedback on setting up this work. We thank Arushi Goel and the attendees at the MT Marathon 2022 for discussions about this work. We thank Ankita Vinay Moghe, Nikolay Bogoychev, and Chantal Amrhein for their comments on the earlier drafts. We thank the anonymous reviewers for their helpful 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). We also thank Huawei for their support (Moghe). Sherborne gratefully acknowledges the support of the UK Engineering and Physical Sciences Research Council (grant EP/W002876/1).

References

Farhad Akhbardeh, Arkady Arkhangorodsky, Magdalena Biesialska, Ondřej Bojar, Rajen Chatterjee, Vishrav Chaudhary, Marta R. Costa-jussa, Cristina España-Bonet, Angela Fan, Christian Federmann, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Leonie Harter, Kenneth Heafield, Christopher Homan, Matthias Huck, Kwabena Amponsah-Kaakyire, Jungo Kasai, Daniel Khashabi, Kevin Knight, Tom Kocmi, Philipp Koehn, Nicholas Lourie, Christof Monz, Makoto Morishita, Masaaki Nagata, Ajay Nagesh, Toshiaki Nakazawa, Matteo Negri, Santanu Pal, Allahsera Auguste Tapo, Marco Turchi, Valentin Vydrin, and Marcos Zampieri. 2021. Findings of the 2021 conference on machine translation (WMT21). In *Proceedings of the Sixth Conference on Machine Translation*, pages 1–88, Online. 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*, pages 479–513, Abu Dhabi. Association for Computational Linguistics.

Chantal Amrhein and Rico Sennrich. 2022. Identifying weaknesses in machine translation metrics through minimum Bayes risk decoding: A case study for COMET. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1125–1141, Online only. Association for Computational Linguistics.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637, Online. Association for Computational Linguistics.

Rachel Bawden, Biao Zhang, Lisa Yankovskaya, Andre Tättar, and Matt Post. 2020. A study in improving BLEU reference coverage with diverse automatic paraphrasing. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 918–932, Online. Association for Computational Linguistics.

Ondřej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, Radu Soricut, Lucia Specia, and Aleš Tamchyna. 2014. Findings of the 2014 workshop on statistical machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 12–58, Baltimore, Maryland, USA. Association for Computational Linguistics.

Ondřej Bojar, Yvette Graham, and Amir Kamran. 2017. Results of the WMT17 metrics shared task. In *Proceedings of the Second Conference on Machine Translation*, pages 489–513, Copenhagen, Denmark. Association for Computational Linguistics.

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2007. (meta-) evaluation of machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 136–158, Prague, Czech Republic. Association for Computational Linguistics.

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Re-evaluating the role of Bleu in machine translation research. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 249–256, Trento, Italy. 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.

Sumanth Doddapaneni, Rahul Aralikatte, Gowtham Ramesh, Shreya Goyal, Mitesh M. Khapra, Anoop Kunchukuttan, and Pratyush Kumar. 2022. Indicxtreme: A multi-task benchmark for evaluating indic languages.

Mihail Eric, Rahul Goel, Shachi Paul, Abhishek Sethi, Sanchit Agarwal, Shuyang Gao, Adarsh Kumar, Anuj Goyal, Peter Ku, and Dilek Hakkani-Tur. 2020. MultiWOZ 2.1: A consolidated multi-domain dialogue dataset with state corrections and state tracking baselines. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 422–428,

Marseille, France. European Language Resources Association.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Michael Auli, and Armand Joulin. 2021. Beyond english-centric multilingual machine translation. *J. Mach. Learn. Res.*, 22:107:1–107:48.

Marina Fomicheva, Piyawat Lertvittayakumjorn, Wei Zhao, Steffen Eger, and Yang Gao. 2021. The Eval4NLP shared task on explainable quality estimation: Overview and results. In *Proceedings of the 2nd Workshop on Evaluation and Comparison of NLP Systems*, pages 165–178, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Marina Fomicheva, Shuo Sun, Erick Fonseca, Chrysoula Zerva, Frédéric Blain, Vishrav Chaudhary, Francisco Guzmán, Nina Lopatina, Lucia Specia, and André F. T. Martins. 2022. MLQE-PE: A multilingual quality estimation and post-editing dataset. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4963–4974, Marseille, France. European Language Resources Association.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021a. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.

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, George Foster, Alon Lavie, and Ondřej Bojar. 2021b. 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.

Daniela Gerz, Pei-Hao Su, Razvan Kusztos, Avishek Mondal, Michał Lis, Eshan Singhal, Nikola Mrkšić, Tsung-Hsien Wen, and Ivan Vulić. 2021. Multilingual and cross-lingual intent detection from spoken data. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7468–7475, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2013. Continuous measurement scales in human evaluation of machine translation. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 33–41, Sofia, Bulgaria. Association for Computational Linguistics.

Yvette Graham, Barry Haddow, and Philipp Koehn. 2020. Statistical power and translationese in machine translation evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 72–81, Online. Association for Computational Linguistics.

Liane Guillou and Christian Hardmeier. 2018. Automatic reference-based evaluation of pronoun translation misses the point. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4797–4802, Brussels, Belgium. Association for Computational Linguistics.

Michael Hanna and Ondřej Bojar. 2021. A fine-grained analysis of BERTScore. In *Proceedings of the Sixth Conference on Machine Translation*, pages 507–517, Online. Association for Computational Linguistics.

Charles T. Hemphill, John J. Godfrey, and George R. Doddington. 1990. The ATIS spoken language systems pilot corpus. In *Speech and Natural Language: Proceedings of a Workshop Held at Hidden Valley, Pennsylvania, June 24-27,1990*.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. XTREME: A massively multilingual multitask benchmark for evaluating cross-lingual generalisation. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 4411–4421. PMLR.

Chia-Chien Hung, Anne Lauscher, Ivan Vulić, Simone Ponzetto, and Goran Glavaš. 2022. Multi2WOZ: A robust multilingual dataset and conversational pretraining for task-oriented dialog. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3687–3703, Seattle, United States. Association for Computational Linguistics.

Srinivasan Iyer, Ioannis Konstas, Alvin Cheung, Jayant Krishnamurthy, and Luke Zettlemoyer. 2017. Learning a neural semantic parser from user feedback. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 963–973, Vancouver, Canada. Association for Computational Linguistics.

Karen Sparck Jones and Julia Rose Galliers, editors. 1996. *Evaluating Natural Language Processing Systems, An Analysis and Review*, volume 1083 of *Lecture Notes in Computer Science*. Springer.

Tom Kocmi, Christian Federmann, Roman Grundkiewicz, Marcin Junczys-Dowmunt, Hitokazu Matsushita, and Arul Menezes. 2021. To ship or not to ship: An extensive evaluation of automatic metrics for machine translation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 478–494, Online. Association for Computational Linguistics.

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

Jamal Laoudi, Calandra R. Tate, and Clare R. Voss. 2006. Task-based MT evaluation: From who/when/where extraction to event understanding. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA).

Hwaran Lee, Jinsik Lee, and Tae-Yoon Kim. 2019. SUMBT: Slot-utterance matching for universal and scalable belief tracking. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5478–5483, Florence, Italy. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *ArXiv preprint*, abs/1907.11692.

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.

Qingsong Ma, Johnny Wei, Ondřej Bojar, and Yvette Graham. 2019. Results of the WMT19 metrics shared task: Segment-level and strong MT systems pose big challenges. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 62–90, Florence, Italy. Association for Computational Linguistics.

Olga Majewska, Evgeniia Razumovskaia, Edoardo M. Ponti, Ivan Vulić, and Anna Korhonen. 2023. Cross-Lingual Dialogue Dataset Creation via Outline-Based Generation. *Transactions of the Association for Computational Linguistics*, 11:139–156.

Benjamin Marie. 2022. An Automatic Evaluation of the WMT22 General Machine Translation Task.

Bilyana Martinovski and David Traum. 2003. The Error Is the Clue: Breakdown In Human-Machine Interaction. In *Proceedings of ISCA Tutorial and Research Workshop International Speech Communication Association*, Switzerland.

Nitika Mathur, Timothy Baldwin, and Trevor Cohn. 2020a. Tangled up in BLEU: Reevaluating the evaluation of automatic machine translation evaluation metrics. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4984–4997, Online. Association for Computational Linguistics.

Nitika Mathur, Johnny Wei, Markus Freitag, Qingsong Ma, and Ondřej Bojar. 2020b. Results of the WMT20 metrics shared task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 688–725, Online. Association for Computational Linguistics.

Brian W. Matthews. 1975. Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochimica et Biophysica Acta (BBA) - Protein Structure*, 405(2):442–451.

Toshiaki Nakazawa, Hideya Mino, Isao Goto, Raj Dabre, Shohei Higashiyama, Shantipriya Parida, Anoop Kunchukuttan, Makoto Morishita, Ondřej Bojar, Chenhui Chu, Akiko Eriguchi, Kaori Abe, Yusuke Oda, and Sadao Kurohashi. 2022. Overview of the 9th workshop on Asian translation. In *Proceedings of the 9th Workshop on Asian Translation*, pages 1–36, Gyeongju, Republic of Korea. International Conference on Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Stefano Perrella, Lorenzo Proietti, Alessandro Scirè, Niccolò Campolungo, and Roberto Navigli. 2022. MaTESe: Machine Translation Evaluation as a Sequence Tagging Problem. In *Proceedings of the Seventh Conference on Machine Translation*, pages 569–577, Abu Dhabi. 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. 2021. Agree to disagree: Analysis of inter-annotator disagreements in human evaluation of machine translation output. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 234–243, Online. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. 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*, pages 578–585, Abu Dhabi. 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.

Sebastian Ruder, Noah Constant, Jan Botha, Aditya Siddhant, Orhan Firat, Jinlan Fu, Pengfei Liu, Junjie Hu, Dan Garrette, Graham Neubig, and Melvin Johnson. 2021. XTREME-R: Towards more challenging and nuanced multilingual evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10215–10245, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Scarton Scarton, Mikel L. Forcada, Miquel Esplà-Gomis, and Lucia Specia. 2019. Estimating post-editing effort: a study on human judgements, task-based and reference-based metrics of MT quality. In *Proceedings of the 16th International Conference on Spoken Language Translation*, Hong Kong. Association for Computational Linguistics.

Tom Sherborne and Mirella Lapata. 2022. Zero-shot cross-lingual semantic parsing. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4134–4153, Dublin, Ireland. Association for Computational Linguistics.

Tom Sherborne and Mirella Lapata. 2023. Meta-Learning a Cross-lingual Manifold for Semantic Parsing. *Transactions of the Association for Computational Linguistics*, 11:49–67.

Miloš Stanojević, Amir Kamran, Philipp Koehn, and Ondřej Bojar. 2015. Results of the WMT15 metrics shared task. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 256–273, Lisbon, Portugal. Association for Computational Linguistics.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2021. Multilingual translation from denoising pre-training. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3450–3466, Online. Association for Computational Linguistics.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT – building open translation services for the world. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 479–480, Lisboa, Portugal. European Association for Machine Translation.

David Vilar, Gregor Leusch, Hermann Ney, and Rafael E. Banchs. 2007. Human evaluation of machine translation through binary system comparisons. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 96–103, Prague, Czech Republic. Association for Computational Linguistics.

Yu Wan, Dayiheng Liu, Baosong Yang, Haibo Zhang, Boxing Chen, Derek Wong, and Lidia Chao. 2022. UniTE: Unified translation evaluation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8117–8127, Dublin, Ireland. Association for Computational Linguistics.

John S. White, Theresa A. O'Connell, and Francis E. O'Mara. 1994. The ARPA MT evaluation methodologies: Evolution, lessons, and future approaches. In *Proceedings of the First Conference of the Association for Machine Translation in the Americas*, Columbia, Maryland, USA.

Weijia Xu, Batool Haider, and Saab Mansour. 2020. End-to-end slot alignment and recognition for cross-lingual NLU. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5052–5063, Online. Association for Computational Linguistics.

Hang Zhang, Liling Tan, and Amita Misra. 2022. Evaluating machine translation in cross-lingual E-commerce search. In *Proceedings of the 15th biennial conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 322–334, Orlando, USA. Association for Machine Translation in the Americas.

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.

Vilém Zouhar, Martin Popel, Ondřej Bojar, and Aleš Tamchyna. 2021. Neural machine translation quality and post-editing performance. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10204–10214, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Code	Language	Code	Language
en	English	el	Greek
de	German	es	Spanish
zh	Mandarin Chinese	hi	Hindi
fr	French	th	Thai
ar	Arabic	tr	Turkish
ru	Russian	vi	Vietnamese
pt	Portuguese		

Table 5: Language codes of languages used in this work

A Language Codes

Please find the language codes in Table 5.

B Implementation Details

We provide the implementation details of metrics and models in Table 6. All models are publicly available and required no training from our side. The metrics BERTScore, COMET family and UniTE family can run on both GPU and CPU. If run on GPU, the metrics run under 5 minutes for a given task and given language pair. No hyperparameters are required. We follow the standard train-dev-test split as released by the authors for DST (Hung et al., 2022) and SP (Sherborne and Lapata, 2022). As no development set is available for the XQuAD dataset, we use the first 200 examples as development set to choose the threshold but report the performance on the full test set.

C Errors of COMET-DA

The proportion of errors from Section 4.2 are listed in Table 2. We also provide error examples in Figure 3.

D Task-specific results

We now list the results across every language pair for all the tasks in Tables tables 7 to 11.

Method	Code	Notes
	Metrics	
chrF	https://github.com/mjpost/sacrebleu	Signature: "nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.1.0"
BLEU	https://github.com/mjpost/sacrebleu	Signature: "nrefs:1|case:mixed|eff:yes|nc:6|nw:0|space:no|version:2.1.0"
BERTScore	https://github.com/Tiiiger/bert_score	Model: xlm-roberta-large
COMET-DA	https://github.com/Unbabel/COMET	Model: wmt20-comet-da
COMET-MQM		Model: wmt21-comet-mqm
COMET-QE-DA		Model: wmt21-comet-qe-da
COMET-QE-MQM		Model: wmt21-comet-qe-mqm
UniTE	https://github.com/NLP2CT/UniTE	Model: UniTE-MUP, hparams.src_ref.yaml
UniTE-QE		Model: UniTE-MUP, hparams.src.yaml
	Extrinsic Task Models	
SP	https://github.com/tomsherborne/zx-parse	
DST	https://github.com/thu-coai/ConvLab-2	
QA	https://huggingface.co/csarron/roberta-base-squad-v1	

Table 6: Metric repositories and versions

Task	MT error	Prediction	input	reference	hypothesis	gold task output	translated task output
SP	mistranslation	No Breakdown	哪些航空公司在 多伦多 和 圣地亚哥 之间飞行	which airlines fly between toronto and san diego	Which airlines fly between Toronto and Santiago?	SELECT DISTINCT airline_1 … city1.city_name = 'TORONTO' … city_2 . city_name = 'SAN DIEGO' ;	SELECT DISTINCT airline_1 ...city1.city_name = 'TORONTO'; (city_2 is excluded)
DST	mistranslation	No Breakdown	Я ищу такси из Yu Garden, которое прибудет к 14:30.	I am looking for a taxi from yu garden arriving by 14:30	I'm looking for a taxi from Yu Garden, which will arrive by 2:30.	['taxi-departure-yu garden', 'taxi-arriveby-14:30']	['taxi-departure-yu garden', 'taxi-arriveby-02:30']
QA	fluency	No Breakdown	विस्तारित महानगरीय क्षेत्र कितने हैं?	How many extended metropolitan areas are there?	How much are the extended metropolitan areas?	two	exceed five million in population.
QA	mistranslation	Breakdown	एनर्जीप्रोजेक्ट AB कहाँ स्थित है?	Where is Energiprojekt AB based?	Where is Energyproject AB located?	Sweden	Sweden
SP	none	Breakdown	查询从 底特律 飞往 多伦多 的航班	get flights from detroit to toronto	Query flights from Detroit to Toronto.	SELECT DISTINCT flight_1 .. city1.city_name = 'DETROIT'... city2.city_name = 'TORONTO';	SELECT DISTINCT flight_1 .. city1.city_name = 'DETROIT'... city2.city_name = 'TORONTO';
DST	none	Breakdown	Да. Забронируйте на 3 человека.	yes. book for 3 people.	Yeah, make a reservation for three people.	['train_book-people-3']	['train_book-people-3']
QA	none	Breakdown	वारसॉ हमेशा से किस प्रकार का शहर रहा है?	What type of city has Warsaw been for as long as it's been a city?	What kind of city has Warsaw always been?	multi-cultural	multi-cultural

Figure 3: Examples of errors made by COMET-DA

Language Good / Bad	zh 1465 / 1796		de 2162 / 1099		ar 1744 / 1517		ru 1517 / 1744	
Method	F1	MCC	F1	MCC	F1	MCC	F1	MCC
Random	0.449	-0.013	0.417	0.018	0.429	-0.018	0.454	0.004
BLEU	0.511	0.079	0.541	0.091	0.540	0.083	0.527	0.076
chrF	0.518	0.078	0.496	0.033	0.499	0.071	0.52	0.086
BERTScore	0.438	0.000	0.519	0.068	0.546	0.136	0.518	0.080
COMET-DA	0.611	0.248	0.581	0.181	0.664	0.328	0.579	0.220
COMET-MQM	0.594	0.201	0.574	0.165	0.625	0.255	0.598	0.196
UniTE	0.642	0.285	0.572	0.164	0.653	0.346	0.614	0.255
COMET-QE-DA	0.558	0.119	0.489	0.03	0.569	0.141	0.476	0.088
COMET-QE-MQM	0.545	0.132	0.552	0.106	0.574	0.195	0.574	0.148
UniTE-QE	0.566	0.183	0.552	0.114	0.628	0.258	0.603	0.215

Table 7: MT metrics for extrinsic Dialogue State Tracking (Multi^{2}WoZ) using an English-trained state tracker. Good/Bad are the number of examples in the respective labels (Not breakdown/Breakdown) for the classification task. Reported Macro F1 scores and MCC scores quantify if the metric detects a breakdown for the extrinsic task. Metrics have negligible correlation with the outcomes of the end task.

Language Good / Bad	ar 592 / 264	de 696 / 169	el 701/ 170	es 721 / 152	hi 631 / 241	ru 701 / 173	th 539 / 323	tr 443 / 389	vi 616 / 251	zh 606 / 266
Random	0.023	-0.002	-0.002	0.017	0.001	-0.002	-0.002	0.028	-0.051	-0.045
BLEU	0.135	0.048	0.142	0.098	0.162	0.125	0.128	0.097	0.108	0.171
chrF	0.160	0.083	0.172	0.092	0.202	0.106	0.162	0.000	0.173	0.119
BERTScore	0.139	0.076	0.173	0.051	0.209	0.131	0.121	0.046	0.173	0.148
COMET-DA	0.193	0.122	0.194	0.086	0.187	0.111	0.125	0.108	0.124	0.120
COMET-MQM	0.096	0.011	0.025	0.017	0.062	-0.023	-0.001	-0.050	0.079	0.054
UniTE	0.068	-0.031	-0.002	-0.014	0.043	0.047	-0.006	0.056	-0.017	-0.023
COMET-QE-DA	0.178	0.084	0.142	0.068	0.125	0.115	0.066	0.049	0.063	0.110
COMET-QE-MQM	0.099	0.050	-0.013	0.025	0.090	-0.025	0.041	-0.077	0.068	0.070
UniTE-QE	0.065	-0.031	0.012	-0.008	0.035	0.069	0.073	0.056	-0.009	-0.069

Table 8: MCC values for different metrics for extrinsic task of Extractive Question Answering (XQuaD dataset) where the model is trained on English. Good/Bad are the number of examples in the respective labels (Not breakdown/Breakdown) for the classification task. Metrics have poor performance on the classification task as a majority report MCC < 0.3

Method	ar	de	el	es	hi	ru	th	tr	vi	zh
Good / Bad	592 / 264	696 / 169	701/ 170	721 / 152	631 / 241	701 / 173	539 / 323	443 / 389	616 / 251	606 / 266
Random	0.508	0.525	0.512	0.492	0.489	0.505	0.490	0.468	0.473	0.498
BLEU	0.549	0.515	0.564	0.543	0.571	0.562	0.556	0.487	0.549	0.585
chrF	0.579	0.541	0.575	0.546	0.595	0.545	0.567	0.480	0.557	0.554
BERTScore	0.569	0.538	0.586	0.523	0.604	0.528	0.561	0.523	0.580	0.535
COMET-DA	0.596	0.560	0.571	0.543	0.593	0.543	0.561	0.549	0.562	0.540
COMET-MQM	0.535	0.351	0.307	0.225	0.361	0.365	0.330	0.429	0.509	0.453
UniTE	0.370	0.479	0.343	0.314	0.308	0.519	0.366	0.438	0.282	0.326
COMET-QE-DA	0.575	0.534	0.559	0.530	0.550	0.544	0.532	0.474	0.530	0.495
COMET-QE-MQM	0.549	0.510	0.416	0.473	0.420	0.384	0.356	0.459	0.509	0.492
UniTE-QE	0.356	0.217	0.344	0.363	0.322	0.534	0.525	0.416	0.281	0.523

Table 9: macro F1 scores for different metrics for extrinsic task of Extractive Question Answering (XQuAD dataset) where the model is trained on English. Good/Bad are the number of examples in the respective labels (Not breakdown/Breakdown) for the classification task.

src	tgt	Random	BLEU	chrF	BERTScore	COMET-DA	COMET-MQM	UniTE	COMET-QE-DA	COMET-QE-MQM	UniTE-QE
	de	0.465	0.492	0.500	0.45	0.436	0.465	0.469	0.511	0.474	0.481
	fr	0.440	0.487	0.519	0.467	0.473	0.491	0.525	0.489	0.525	0.509
en	pt	0.466	0.676	0.659	0.614	0.555	0.609	0.4525	0.527	0.500	0.588
	es	0.463	0.599	0.566	0.564	0.630	0.614	0.626	0.546	0.535	0.574
	zh	0.429	0.574	0.570	0.582	0.590	0.577	0.586	0.516	0.513	0.490
	en	0.490	0.611	0.598	0.623	0.624	0.637	0.629	0.556	0.620	0.673
	fr	0.409	0.523	0.539	0.515	0.595	0.613	0.608	0.592	0.522	0.536
de	pt	0.462	0.592	0.641	0.638	0.684	0.683	0.619	0.645	0.619	0.580
	es	0.479	0.605	0.621	0.569	0.666	0.631	0.684	0.596	0.576	0.621
	zh	0.468	0.614	0.670	0.571	0.614	0.553	0.581	0.524	0.532	0.554
	en	0.489	0.595	0.590	0.607	0.630	0.606	0.628	0.597	0.574	0.588
	de	0.385	0.518	0.616	0.587	0.541	0.570	0.546	0.503	0.476	0.542
fr	pt	0.472	0.620	0.620	0.565	0.543	0.583	0.538	0.549	0.534	0.520
	es	0.492	0.462	0.613	0.512	0.627	0.648	0.574	0.594	0.568	0.573
	zh	0.384	0.641	0.702	0.666	0.667	0.658	0.661	0.521	0.502	0.575
	en	0.476	0.629	0.676	0.681	0.685	0.655	0.705	0.695	0.654	0.526
	de	0.438	0.550	0.575	0.577	0.586	0.594	0.481	0.608	0.569	0.501
pt	fr	0.458	0.546	0.603	0.488	0.599	0.495	0.574	0.574	0.545	0.645
	es	0.491	0.640	0.646	0.634	0.639	0.639	0.459	0.562	0.586	0.509
	zh	0.403	0.610	0.690	0.551	0.580	0.511	0.621	0.621	0.492	0.591
	en	0.455	0.530	0.561	0.566	0.605	0.601	0.600	0.544	0.564	0.529
	de	0.455	0.530	0.546	0.587	0.540	0.521	0.584	0.49	0.486	0.513
es	fr	0.453	0.542	0.531	0.606	0.564	0.568	0.584	0.569	0.560	0.556
	pt	0.500	0.506	0.561	0.579	0.554	0.564	0.529	0.561	0.566	0.581
	zh	0.374	0.562	0.644	0.562	0.627	0.587	0.687	0.524	0.478	0.662
	en	0.455	0.530	0.561	0.566	0.605	0.601	0.600	0.544	0.564	0.529
	de	0.455	0.530	0.546	0.587	0.540	0.521	0.584	0.490	0.486	0.513
es	fr	0.453	0.542	0.531	0.606	0.564	0.568	0.584	0.569	0.560	0.556
	pt	0.500	0.506	0.561	0.579	0.554	0.564	0.529	0.561	0.566	0.581
	zh	0.374	0.562	0.644	0.562	0.627	0.587	0.687	0.524	0.478	0.662

Table 10: MT Metric performance on F1 for extrinsic semantic parsing (MultiATIS++SQL) with the parser trained in src language.

src	tgt	Random	BLEU	chrF	BERTScore	COMET-DA	COMET-MQM	UniTE	COMET-QE-DA	COMET-QE-MQM	UniTE-QE
	de	0.012	0.008	0.016	-0.096	-0.122	-0.000	-0.06	0.025	-0.021	-0.027
	fr	-0.043	-0.024	0.039	-0.066	-0.020	-0.001	0.050	-0.021	-0.021	0.017
en	pt	-0.067	0.353	0.328	0.231	0.201	0.228	0.114	0.089	0.209	0.187
	es	0.002	0.203	0.133	0.152	0.279	0.229	0.252	0.110	0.107	0.166
	zh	-0.090	0.152	0.146	0.173	0.187	0.188	0.172	0.060	0.035	0.078
	en	-0.003	0.226	0.210	0.251	0.263	0.328	0.303	0.161	0.250	0.349
	fr	-0.007	0.046	0.078	0.033	0.196	0.226	0.243	0.185	0.044	0.078
de	pt	-0.070	0.184	0.300	0.312	0.394	0.406	0.302	0.331	0.295	0.206
	es	-0.035	0.230	0.242	0.200	0.332	0.264	0.370	0.206	0.181	0.256
	zh	-0.063	0.241	0.340	0.150	0.242	0.124	0.258	0.054	0.088	0.112
	en	0.006	0.194	0.182	0.220	0.269	0.229	0.262	0.195	0.148	0.178
	de	-0.087	0.099	0.237	0.180	0.105	0.155	0.125	0.026	-0.043	0.086
fr	pt	-0.023	0.242	0.240	0.177	0.133	0.170	0.117	0.100	0.115	0.106
	es	-0.015	0.053	0.233	0.118	0.283	0.300	0.151	0.229	0.177	0.153
	zh	-0.116	0.311	0.413	0.373	0.365	0.347	0.390	0.143	0.051	0.248
	en	0.013	0.315	0.365	0.378	0.372	0.320	0.414	0.402	0.310	0.175
	de	-0.093	0.112	0.181	0.159	0.188	0.190	0.216	0.183	0.150	0.007
pt	fr	0.013	0.100	0.222	0.061	0.218	0.030	0.155	0.053	0.090	0.291
	es	0.009	0.286	0.293	0.278	0.278	0.288	0.142	0.076	0.243	0.025
	zh	0.061	0.221	0.449	0.253	0.161	0.048	0.242	0.000	-0.011	0.212
	en	-0.063	0.080	0.179	0.136	0.214	0.208	0.200	0.095	0.128	0.058
	de	-0.075	0.092	0.169	0.175	0.082	0.047	0.186	-0.013	-0.024	0.033
es	fr	-0.065	0.140	0.118	0.214	0.129	0.140	0.196	0.150	0.124	0.112
	pt	0.014	0.012	0.144	0.169	0.148	0.143	0.110	0.160	0.133	0.166
	zh	-0.005	0.148	0.289	0.154	0.254	0.173	0.393	0.102	0.000	0.363
	en	-0.034	0.283	0.218	0.252	0.302	0.290	0.333	0.264	0.324	0.232
	de	0.008	0.260	0.274	0.302	0.314	0.347	0.273	0.139	0.199	0.169
zh	fr	-0.045	0.204	0.238	0.343	0.330	0.247	0.328	0.222	0.259	0.287
	pt	-0.130	0.264	0.357	0.430	0.327	0.295	0.307	0.171	0.205	0.134
	es	-0.015	0.340	0.375	0.446	0.407	0.417	0.213	0.139	0.229	0.211

Table 11: MT Metric performance on MCC for the classification task with extrinsic semantic parsing (MultiATIS++SQL) with the parser trained in src language.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
7

☑ A2. Did you discuss any potential risks of your work?
8

☑ A3. Do the abstract and introduction summarize the paper's main claims?
1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

3

☑ B1. Did you cite the creators of artifacts you used?
3

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
8

☐ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Not applicable. 3

☐ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Not applicable. 8

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
3

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Appendix

C ☑ Did you run computational experiments?

4

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Appendix

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
3

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
4

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Appendix

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Not applicable. Left blank.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Not applicable. Left blank.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Not applicable. Left blank.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Not applicable. Left blank.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Not applicable. Left blank.