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 translation qualities of individual machine translation systems. However, it is unclear whether automatic metrics produce reliable sentence level scores for downstream tasks that use machine translation as an intermediate step. We evaluate the segment-level performance of nine MT metrics (chrF, COMET, BERTScore, *etc.*) on three downstream cross-lingual tasks (dialogue state tracking, question answering, and semantic parsing). For each task, we only have access to a monolingual task-specific 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 the *Translate-test* setup. Our experiments demonstrate that all metrics exhibit negligible correlation with downstream outcomes. We also find that the scores provided by neural metrics are not interpretable mostly because of undefined ranges. Our analysis suggests that future MT metrics be designed to produce error labels rather than scores to facilitate extrinsic evaluation.

1 Introduction

Machine translation (MT) systems are being widely deployed as a stand-alone service or used as an application programming interface in complex tasks 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 pipeline of the complex task leading to a poor user experience. For example, consider the user's request in Chinese 有牙加菜？ (*Is there any good Jamaican food in Cambridge?*) is machine translated into English as *Does Cambridge have a good meal in Jamaica?*. An intent detection model will thus consider "Jamaica" as a location instead of cuisine and prompt the search engine to look up restaurants in Jamaica [1]. To avoid this cascading of errors, it is crucial to detect an incorrect translation before it is passed into the next stage.

One way to approach this problem is to use segment level scores provided by MT metrics. There has been great progress in the development of automatic MT metrics with some metrics demonstrating > 0.9 correlation on the system level for some language pairs (Ma et al., 2019). Despite MT systems being a crucial intermediate step in several applications, the behaviour of these metrics under task-oriented evaluation is less explored.

In this work, we provide a complementary evaluation of MT metrics where the scores from these metrics are used to determine if that translation will not add to new errors in the downstream task. We consider the *Translate-Test* setting. 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. At test time, the target language input is translated into the source language and then executed on a task-specific model. We use the outcome of this extrinsic task to construct a binary classification benchmark for the metrics. Within this new dataset, we exclude all the examples where the examples in the source language failed on the extrinsic task. The metrics are then evaluated on this new classification task.

We use dialogue state tracking, semantic parsing, and extractive question answering as our respective extrinsic tasks. We evaluate nine metrics consisting of string overlap metrics, embedding based metrics, and metrics trained using scores obtained through human evaluation of MT outputs. Surprisingly, we find that this setup is hard for the existing metrics as the metrics show poor performance on the classification task. We thoroughly analyse the failure

[1] Example taken from the Multi^{2}WoZ dataset

of the metrics through quantitative and qualitative evaluation. We also investigate peculiar properties of the metrics such as generalisation to new languages, alternatives to references for reference-based metrics in an online setting, and conduct task-specific ablation studies.

Our findings are summarised as follows:

1. We devise a new classification task that measures whether the segment level scores are indicative of the downstream performance of the extrinsic task (See Section 3).
2. We find that segment level scores provided by the nine metrics have a negligible correlation with the performance of the end task (See Section 4.1). Most metrics produce scores that are uninformative (Section 4.3). Also, varying tasks have varying sensitivity to different MT errors (Section 4.5).
3. We make recommendations to develop MT metrics that predict labels instead of scores and suggest reusing existing post-editing datasets/MQM labels (See Section 5).

2 Related Work

Evaluation of machine translation has been of great research interest across different research communities (Nakazawa et al., 2021; 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. A common protocol in evaluating these metrics is to compare the scores produced by these metrics with human judgements collected for the output translations. Designing protocols for human evaluation of machine translated outputs and meta evaluation is not straightforward (Mathur et al., 2020a), leading to the development of several different methodologies and analyses over the years.

Human evaluation of MT systems has been carried out based on guidelines for fluency, adequacy and/or comprehensibility (White et al., 1994) evaluating every generated translation often on a fixed scale of 1 to 5 (Koehn and Monz, 2006) or 1 to 100 (Graham et al., 2013) (direct assessments). For some years, the ranking of MT systems was based on a binary comparison of outputs from two different MT systems (Vilar et al., 2007). More recently, expert-based evaluation is carried out based on Multidimensional Quality Metrics (Lommel et al., 2014) where translation outputs are scored on the severity of errors using a fine-grained error ontology (Freitag et al., 2021a,b). Over the years, different methods to compute the correlation between the scores produced by the metrics and this human evaluation have been suggested based on the drawbacks of the previous ones (Callison-Burch et al., 2006; Bojar et al., 2014, 2017). Most metrics claim their effectiveness by comparing their performance with competitive metrics on the recent method for computing correlation with human judgements on the system-level.

The meta evaluation progress is generally documented in the metrics shared task overview papers (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 the overview papers, Mathur et al. (2020a) show that then meta evaluation regimes were sensitive to outliers and small changes in evaluation metrics were not sufficient to claim the effectiveness of any metric. Kocmi et al. (2021) conduct a comprehensive evaluation of metrics 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.

All the above works draw their conclusions based on some comparison with human judgement. Our work focuses on the usability of the metrics which is solely judged on their ability to work with downstream tasks where MT is used as an intermediate step. The emphasis of the meta evaluation is also on their segment level performance. Task based MT evaluation has been well studied in the literature (Jones and Galliers (1996); Laoudi et al. (2006); Zhang et al. (2022), *inter alia*) However, these works focus on evaluating individual MT systems than investigating MT metrics. Scarton et al. (2019) propose a task based evaluation of metrics where the task is to rank translations based on the time to post-edit them. Our work is the first to address the evaluation of MT metrics through extrinsic tasks when MT is used as an intermediate step.

3 Methodology

We describe the construction of the classification setup and the metrics evaluated on the classification task.

3.1 Setup

For all the tasks described below, we first train a model for that task in the monolingual setup. We then evaluate the source language on that task and store the predictions of the model. As we follow the *Translate-test* setup, the target language inputs for the task are the first machine translated into the source language and then the translations are given as input to the task model. We use either (i) OPUS translation models or (ii) M2M100 translation or (iii) translations provided by the authors of the respective datasets. Note that the data across the target languages are parallel. We obtain the predictions for the translated data to construct a binary classification benchmark for the metrics.

We only consider all the examples in the target language that were predicted as correct in the source language to avoid any errors that arise from the complexity of the task. Thus, all the incorrect predictions for the target language in the end task should arise from erroneous translations. We use these predictions to build a binary classification benchmark - all the examples from the target language that are correctly predicted receive a positive label while the incorrect predictions receive a negative label.

We consider the input from the target language as "source", the corresponding machine translation as "hypothesis" and the input from the source language as "reference". These triples are then scored by the respective metrics. After obtaining the segment-level scores for these triples, we find a threshold for the scores, thus turning the metrics into classifiers. The metrics are then evaluated on how well their predictions for a good/bad translation correlate with the success/failure of the end task for the target language.

3.2 Tasks

We evaluate the metrics on the following tasks.

3.2.1 Dialogue State Tracking

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 example is marked correct only if the predicted dialogue state is exactly equal to the ground truth to provide labels for the binary classification task. The metric scores are produced for the current utterance said by the user.

3.2.2 Semantic Parsing

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 similarly follows Sherborne and Lapata (2022), as an encoder-decoder Transformer model based on MBART50 (Tang et al., 2021). The model 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 outputs.

3.2.3 Extractive Question Answering

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 [2]. We use the "Exact-Match" metric, i,e:

[2] https://huggingface.co/csarron/roberta-base-squad-v1

Metric / Lang Good / Bad	zh 1465 / 1796		de 2162 / 1099		ar 1744 / 1517		ru 1517 / 1744	
Method	F1	MCC	F1	MCC	F1	MCC	F1	MCC
random	0.489	0.02	0.499	-0.002	0.497	0.008	0.486	0.006
BLEU	0.401	-0.166	0.601	0.211	0.421	-0.153	0.469	-0.062
chrF	0.384	-0.227	0.601	0.212	0.421	-0.153	0.486	0.197
BERTScore	0.355	0	0.374	-0.106	0.425	-0.146	0.392	-0.203
COMET-DA	0.357	-0.115	0.554	0.141	0.458	-0.085	0.43	-0.052
COMET-MQM	0.462	0.127	0.51	0.097	0.462	0.073	0.474	0.14
UniTE	0.46	-0.077	0.566	0.158	0.478	0.094	0.434	-0.031
COMET-QE-DA	0.526	0.196	0.535	0.083	0.55	0.173	0.429	-0.053
COMET-QE-MQM	0.525	0.193	0.511	0.023	0.55	0.172	0.62	0.243
UniTE-QE	0.426	-0.112	0.571	0.164	0.483	0.095	0.613	0.227

Table 1: Performance of different metrics when the extrinsic task is dialogue state tracking on the Multi^{2}WoZ dataset where the state tracker is trained in English. The good/bad are the number of examples in the respective labels for the classification task. Macro F1 scores and MCC scores are reported to quantify if the metric can actually detect a breakdown for the extrinsic task. Metrics have negligible correlation with the outcomes of the end task.

the model's predicted answer span exactly matches the gold standard answer span; for the binary classification 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 threshold.

3.3 Metrics

We describe three types of metrics based on their design

3.3.1 Surface level overlap

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

chrF (Popović, 2017) evaluates translation outputs based on a character n-gram F-score by computing overlaps 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 organizes 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 between 0 to 100 based according to 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. We now describe these metrics:

COMET-DA (Rei et al., 2020) uses a cross-lingual encoder (XLM-R (Conneau et al., 2020)) and pooling operations to obtain sentence-level representations of the source, hypothesis, and reference. These sentence embeddings are combined and then passed through a feedforward network to produce a score. COMET is trained on human evaluation scores of machine translation systems submitted to WMT until 2020.

COMET-QE-DA was trained similarly to COMET-20 but only the source and the hypothesis are combined to produce a final score as this is a reference-free metric.

COMET-MQM was trained similarly as COMET-DA. The training data consists of direct assessments till 2021 which is then fine-tuned with MQM scores. **COMET-QE-MQM** is the reference free version of COMET-MQM i,e: references are excluded during training and evaluation.

In all the variants of the COMET, the MQM scores and DA scores are converted to z-scores to reduce the effect of outlier annotations.

UniTE (Wan et al., 2022), Unified Translation Evaluation, is a metric approach where the model-

Metric / Lang Good / Bad	en 220 / 119		fr 293 / 46		pt 210 / 129		es 193 / 146		zh 174 / 165	
	F1	MCC	F1	MCC	F1	MCC	F1	MCC	F1	MCC
random	0.49	-0.003	0.409	-0.007	0.462	-0.07	0.479	-0.035	0.468	-0.063
BLEU	0.611	0.226	0.523	0.046	0.592	0.184	0.605	0.23	0.614	0.241
chrF	0.598	0.21	0.539	0.078	0.641	0.3	0.621	0.242	0.67	0.34
BERTScore	0.623	0.251	0.515	0.033	0.638	0.312	0.569	0.2	0.571	0.15
COMET-DA	0.624	0.263	0.595	0.196	0.684	0.394	0.666	0.332	0.614	0.242
COMET-MQM	0.637	0.328	0.613	0.226	0.683	0.406	0.631	0.264	0.553	0.124
UniTE	0.629	0.303	0.608	0.243	0.619	0.302	0.684	0.37	0.458	-0.076
COMET-QE-DA	0.556	0.161	0.592	0.185	0.645	0.331	0.596	0.206	0.524	0.054
COMET-QE-MQM	0.62	0.25	0.522	0.044	0.619	0.295	0.576	0.181	0.532	0.088
UniTE-QE	0.503	0.024	0.464	0	0.556	0.132	0.631	0.279	0.327	0

Table 2: Performance of different metrics when the extrinsic task is parsing on the MultiATIS++ dataset with the parser trained in German. The good/bad are the number of examples in the respective labels for the classification task. Typically, metrics have negligible correlation with the outcomes of the end task.

based metrics can possess the ability to evaluate translation outputs following all three evaluation scenarios, i.e. source-only, reference-only, and source-reference-combined.

3.4 Evaluation

We evaluate the performance of the metric on the respective binary classification tasks using macro-F1 and Matthew's Correlation Coefficient (MCC) (Matthews, 1975). As the class distribution will change depending on the task and the language pair, we selected metrics that are robust to class imbalance. We included MCC to interpret the MT metric's standalone performance for the given extrinsic task. The range of macro-F1 is between 0 to 1 and the range of MCC is between -1 to 1. An MCC value around 0 indicates no correlation. Any MCC value between 0 and 0.3 indicates negligible correlation, 0.3 to 0.5 indicates low correlation.

4 Results

We report the results for dialogue state tracking (Table 1), semantic parsing (Table 2) and question answering (Table 3). We only report the results when the source language is De in Table 2 and report the results for all the other source languages in Appendix. We also report the macro-F1 scores for question answering in Appendix. 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 the metrics perform better than the random baseline on the macro-F1 metric (except BERTScore for dialogue state tracking). We use MCC to identify if this increase in macro-F1 makes the metric usable in the end task. Evaluating on MCC, we find that all the metrics show negligible correlation under almost all settings, across all three tasks. Contrary to the trend where neural metrics are better than metrics based on surface overlap (Freitag et al., 2021b), we find this binary classification to be difficult across all the metrics.

While comparing the reference-based versions of trained metrics (COMET-DA, COMET-MQM, UniTE) with their reference-free equivalents (COMET-QE-DA, COMET-QE-MQM, UniTE-QE respectively), we observe that generally reference-based perform better than their reference-free versions for semantic parsing and question answering. However, this trend flips for the dialogue state tracking where reference-free performs the same or better than reference-based metrics. We also note that references are unavailable at test time, hence reference-based metrics are not suitable in this setting. We discuss alternative ways of obtaining references in Section 4.4.

Between the use of MQM-scores and DA-scores during fine-tuning the different COMET variants, we find that both COMET-QE-DA and COMET-DA are better than COMET-QE-MQM and COMET-MQM respectively for question answering. There is no clear winner for both dialogue state tracking and semantic parsing. The metrics exhibit similar performance for dialogue state tracking in the reference-based scenario. For the remaining cases, the MQM-based metrics and DA-based metrics have varying performance.

We now conduct some further analyses on the results:

4.2 Performance on zero-shot language pairs

The language pairs in the trained metrics are only based on the language pairs in WMT data, around

Metric / Lang	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	0.023	-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	0.16	0.083	0.172	0.092	0.202	0.106	0.162	0	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	0.12
COMET-MQM	0.096	0.011	0.025	0.017	0.062	-0.023	-0.001	-0.05	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	0.11
COMET-QE-MQM	0.099	0.05	-0.013	0.025	0.09	-0.025	0.041	-0.077	0.068	0.07
UniTE-QE	0.065	-0.031	0.012	-0.008	0.035	0.069	0.073	0.056	-0.009	-0.069

Table 3: MCC for different metrics when the extrinsic task is extractive question answering where the model is trained for English question. The good/bad are the number of examples in the respective labels for the classification task. Metrics have poor performance on the classification task as most of them report MCC < 0.3

half of which contains English as one of the languages in the translation pair. We also note that our dialogue state tracking and question answering tasks only have the task specific language as English. Similar to Kocmi et al. (2021), we consider a subset of results from the semantic parsing task to investigate if the use of multilingual embeddings in these trained metrics allows generalization to unseen language pairs. We consider five different types of language pairs in the zero-shot setting. (i) *en-es* is unseen language pair when the source language is en (ii) *pt-en* is unseen language pair when the target language is en (iii) *es-de* is the pair where the two languages are from different language families but share the same script, (iv) *zh-fr* contains source language with logogram script and (v) *de-zh* contains target language with logogram script. We report the difference between the chosen MT metric and the random baseline for different language pairs in Table 4. We find that trained metrics have similar performance on unseen languages except UniTE metrics for *en-es* and UniTE reference based for *pt-en*. Metrics based on string matches have similar (and sometimes better) performance than the trained metrics. Reference-based metrics have higher gains than their respective reference-free counterparts. As the metrics already have a poor performance on the seen languages, we cannot verify if this similar poor performance of unseen language pairs is an indicator of generalization. We also find that the metrics have higher gains in the case of *zh-fr* where the translation is of poor quality for the end task. This is expected as the errors in these translations will be easier to detect.

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

src-tgt	en-es	pt-en	es-de	zh-fr	de-zh
examples	193/92	171 / 172	193 / 196	53/275	144 / 183
random	0.455	0.466	0.479	0.384	0.503
COMET-DA	0.15	0.089	0.187	0.283	0.145
COMET-MQM	0.146	0.143	0.152	0.274	0.159
UniTE	-0.031	-0.039	0.205	0.121	0.094
COMET-QE-DA	0.089	0.061	0.117	0.137	0.036
COMET-QE-MQM	0.109	0.118	0.097	0.118	0.053
UniTE-QE	-0.022	0.065	0.152	0.194	0.014

Table 4: Difference of macro F1 scores of different metrics and random baseline with semantic parsing as extrinsic task to check generalisation of metrics . The top row has absolute macro F1 scores for the random baseline and every other row is $\Delta = metric_{F1} - random_{F1}$. The first language in the pair is the source language of the task-specific model. The second language is the target language used in the translate test setting. The metrics show similar yet poor performance to Table 2.

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 observe that finding the right threshold to identify if a translation needs correction is not straightforward.

We report the mean and standard deviation of best thresholds for every language pair for every metric in Table 5. Surprisingly, the thresholds are not consistent and off from the midpoint in the case of bounded metrics - BLEU (0-100), chrF (0-100), and BERTScore (0 to 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. The COMET metrics except COMET-DA have smaller 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

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

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

metric	SP	QA	DST
chrF	44.0 ± 13.7	53.9 ± 07.8	58.0 ± 10.6
BLEU	15.5 ± 08.8	16.1 ± 04.9	27.5 ± 08.3
BERTScore	0.50 ± 0.21	0.54 ± 0.08	0.79 ± 0.15
COMET-DA	0.21 ± 0.35	0.30 ± 0.23	0.75 ± 0.13
COMET-MQM	0.03 ± 0.01	0.06 ± 0.01	0.04 ± 0.01
UniTE	0.01 ± 0.37	-0.4 ± 0.38	0.43 ± 0.10
COMET-QE-DA	0.02 ± 0.07	0.02 ± 0.01	0.11 ± 0.02
COMET-QE-MQM	0.11 ± 0.01	0.00 ± 0.04	0.12 ± 0.01
UniTE-QE	-0.07 ± 0.48	-0.24 ± 0.13	0.36 ± 0.16

Table 5: 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 not consistent across language pairs and across tasks for both bounded and unbounded metrics. QA is question answering. DST=Dialogue state tracking. SP is semantic parsing

Method	zh	de	ar	ru
BLEU	0.078	0.025	-0.065	-0.103
chrf	0.314	0.211	0.128	-0.018
BERTScore	0.313	0.229	0.15	-0.017
COMET-DA	0.158	0.14	0.075	-0.105
COMET-MQM	0.18	0.167	0	-0.115

Table 6: MCC scores of reference based metrics for the extrinsic task of dialogue state tracking. The setup simulates an online setting where gold standard references are not available. Instead, the translation is translated back into the target language. MCC scores improve in this setup over Table 1 as long as the quality of machine translation is high (zh, de, ar).

range for COMET-MQM is found to lie between -0.2 to 0.2 which questions whether this small standard deviation is an indicator of the consistency of the threshold. 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 notion of negative scores for good translation 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 in this work do not provide segment level scores whose range can be interpreted meaningfully across both tasks and different language pairs.

4.4 Reference-based metrics in online setting

In an online setting, we don't have access to references at test time. To test the effectiveness of reference-based methods in this setting, we consider translating the translation back into the source language. While evaluating the reference-based methods on the new setup, the language pair flips the direction; the src_{new} is mt_{old}, mt_{new} is the translation of mt_{old}, and ref_{new} is src_{old}. We generate these new translations using the mBART translation model (Tang et al., 2020). We report these results for the dialogue state tracking task in Table 6. The table calculates the difference between the MCC values of the round trip translation task and the MCC values reported in Table 1.

We find that most metrics improve their performance by using the new reference except when the target language is ru. This is reassuring that as reference-based metrics improve, their deployment in a reference-less setting can still be useful when the quality of the translation outputs is high. However, their correlation coefficients are within the range of negligible correlation. Using reference-based metrics in an online setting comes with the overhead of producing another translation. Using ru as input language has the lowest performance on the downstream task indicating the machine translation quality of ru-en translation is inferior. The back translation from en to ru is likely to add additional errors to the existing erroneous translation. This cascading of errors confuses the metric and it can mark a perfectly useful translation from ru-en as "bad" due to the error present in the en-ru translation. For example, in the ru-en case, COMET-MQM has a false negative rate of 0.796 in the round-trip translation setup compared to 0.097 when the human reference is used instead. Thus, this setting is likely to fail when the machine translation models generate poor-quality outputs.

4.5 Qualitative evaluation

The development of machine translation metrics largely caters towards only the intrinsic task of evaluating the quality of a translated text in the target language. The severity to penalize 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. However, different downstream tasks will demonstrate varying levels of sensitivity to the same machine translation errors (Zhang et al., 2022). For example, the fluency of a translation is likely to be more crucial when translating an utterance from the system to the user than translating from the user to the system.

`html`

To quantify which translation errors are most crucial to the respective extrinsic tasks, we conduct a qualitative evaluation of the outputs of the respective classification tasks. We consider the behaviour COMET-DA for this case study. We annotate at least 100 examples containing the false positives and the false negatives for semantic parsing when the parser is trained in English and the target language input is Chinese translated into English. We annotate the MT errors (if present) in these examples based on the MQM ontology.

For semantic parsing, we find that 54% of the errors belong to mistranslation. The other MT errors belong to addition(2%), omission (6%), and fluency (8%). The rest 30% of the errors did not have any MT errors (none). These translations were paraphrases of the references which were undetected as correct translations by the metric; except in six instances where the parsing model could not handle the diverse input. Within mistranslation, 93% of the errors are sensitive to the downstream task; like named entity errors. We observe that fluency error in the translation largely does not have an impact on the parsing task.

We also find that approx 20% of the errors made by the classification model arise from the task-specific model. For example, the MT model uses an alternative term of *shuttle* instead of *round-trip* while generating the translation for the reference "show me round trip flights from montreal to orlando". The semantic parser fails to generalise despite being trained with mBART.

4.6 Ablation

We look at some ablations studies dependent on the nature of the end tasks.

4.6.1 Cascading errors in dialogue

The results reported in Table 1 illustrate a scenario where the automatic translation is applied on every utterance. The dialogue state tracking model includes the history of the conversation while predicting the current state. Thus, if an entity is incorrectly translated at the start of the conversation, it is likely to produce cascaded errors on the dialogue state tracking task. To eliminate this effect of cascading errors, we also consider a setting where every utterance except the current utterance uses the gold standard translation. We then perform the classification task and report these results in Table 7.

The number of Good examples has increased over Table 1 confirming that using machine translation directly without correction causes cascading errors in state tracking. However, there is a drop in the performance for most metrics as compared to Table 1.

Metric \Lang	zh	de	ar	ru
Good / Bad	2542/ 719	2876/385	2699/562	2563/698
random	0	0.044	-0.019	0.005
BLEU	0.067	0	0.144	0.03
chrF	-0.058	0	0	-0.009
BERTScore	0	0.012	-0.02	0.048
COMET-DA	0.076	-0.011	-0.023	0.026
COMET-MQM	-0.038	-0.023	-0.024	-0.051
UniTE	-0.058	-0.036	-0.032	-0.067
COMET-QE	-0.058	-0.024	-0.04	-0.048
COMET-QE-MQM	-0.057	0.033	-0.021	-0.069
UniTE-QE	-0.069	-0.012	-0.04	-0.072

Table 7: MCC scores for metrics when extrinsic task is dialogue state tracking when the dialogue history is gold translations and only the context is automatically translated. Number of bad examples decrease compared to Table 1 indicating cascading of errors due to incorrect translations in the dialogue history. Metrics also perform poor than Table 1.

4.6.2 Effect of true casing

The original ATIS dataset (Hemphill et al., 1990) in English is available only in lowercase. However, contemporary machine translation models are trained to produce true case outputs irrespective of the input. We investigate whether this mismatch of casing in the hypothesis and reference causes a change in the performance of the reference-based metrics on the downstream task. We truecase the references using an automated tool. Note, some entities are still in lowercase after running the tool. We report these results in Table 8 which contains the F1 difference between the scores computed using the truecase references and the lowercase references.

We find that both surface form overlap based and neural metrics have a difference in the performance but there is no clear trend if using true case reference is beneficial for the end task. The effects of using correct casing are most beneficial for the de-en and fr-en semantic parsing tasks.

5 Recommendations

Our experiments suggest that evaluating MT metrics on the segment level for MT metrics has considerable room for improvement. We make some recommendations based on our observations-

Explicit error Analysis: We reinforce the proposal of using the MQM scoring scheme for eval-

Metrics	de	fr	pt	es	zh
BLEU	0.003	0.055	-0.05	0.002	-0.033
chrF	0.014	0.034	-0.039	-0.014	-0.034
BERTScore	0.013	0.039	-0.009	-0.005	-0.037
COMET-DA	0.018	0.029	0.075	-0.051	-0.021
COMET-MQM	0.078	0.082	-0.008	-0.015	-0.027
UniTE	-0.004	0.305	0.032	-0.093	0.082

Table 8: We report the difference between F1 scores of different metrics for the classification task when the reference is in the (synthetic) truecase v/s lowercase for reference based metrics. The extrinsic task is semantic parsing on the MultiATIS++ dataset where the semantic parser is trained in English. The coloumns report the F1 difference of the respective target language. Casing does not have a conclusive effect on the reference based metrics. $\Delta = truecase - lowercase$

uating MT outputs as introduced in Freitag et al. (2021a). As seen in Section 4.5, different tasks have varying tolerance to different MT errors. With explicit errors marked per MT output, future classifiers can only be trained on a subset of human evaluation data containing errors most relevant to the downstream application.

Combination of metrics: We do not find a "winning metric" across our tasks (See Section 4). We also find that neural metrics have performance similar to surface overlap metrics. Similar to Amrhein et al. (2022), we recommend using a combination of different families of metrics to judge the usability of the MT output for the downstream task.

Adding diverse references: From Section 4.5, we find that both the neural metric and the task-specific model are not robust to paraphrases. Bawden et al. (2020) proposed automatic paraphrasing of references to improve the coverage for BLEU. We also recommend the inclusion of diverse references through automatic paraphrasing or data augmentation during the training of neural metrics.

Produce labels instead of scores: From Section 4.3, we find that a single score from the metric is difficult to interpret the quality of the produced MT translation. 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. The MQM annotations released by Freitag et al. (2021a) contain spans for erroneous words and the corresponding types of errors which can help set up the classification task. 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 is the work by Perrella et al. (2022) that treats MT evaluation as sequence-tagging problem i,e: words in the candidate translation are tagged with 'minor" or "major" error and then converted into a weighted score. We hope to incorporate these techniques in developing future evaluation regimes when using MT as an intermediate step in extrinsic tasks.

6 Conclusion

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 - dialogue state tracking, question answering, and semantic parsing. We found 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 attributed this result to segment scores produced by these metrics being uninformative and that different extrinsic tasks demonstrate different levels of sensitivity for different MT errors. We made recommendations to predict error types instead of error scores to facilitate the use of MT metrics in downstream tasks.

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.

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.

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.

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, 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.

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.

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*.

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. *CoRR*, 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.

Benjamin Marie. 2022. An automatic evaluation of the wmt22 general machine translation task.

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.

B.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, Hideki Nakayama, Chenchen Ding, Raj Dabre, Shohei Higashiyama, Hideya Mino, Isao Goto, Win Pa Pa, Anoop Kunchukuttan, Shantipriya Parida, Ondřej Bojar, Chenhui Chu, Akiko Eriguchi, Kaori Abe, Yusuke Oda, and Sadao Kurohashi. 2021. Overview of the 8th workshop on Asian translation. In *Proceedings of the 8th Workshop on Asian Translation (WAT2021)*, pages 1–45, Online. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting 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. Machine Translation Evaluation as a Sequence Tagging Problem. In *Proceedings of the Seventh Conference on Machine Translation*, 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.

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, 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.

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.

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. 2020. Multilingual translation with extensible multilingual pretraining and finetuning.

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.

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.