Exploring Robustness of Machine Translation Metrics: A Study of Twenty-Eight Automatic Metrics in the WMT22 Metric Task

Xiaoyu Chen, Daimeng Wei, Hengchao Shang, Zongyao Li, Zhanglin Wu, Zhengzhe Yu, Ting Zhu, Mengli Zhu, Ning Xie, Lizhi Lei, Shimin Tao, Hao Yang, Ying Qin

Huawei Translation Services Center, Beijing, China

{chenxiaoyu35,weidaimeng,shanghengchao,lizongyao,wuzhanglin2, yuzhengzhe,zhuting20,zhumengli,nicolas.xie,leilizhi,taoshimin, yanghao30,qinying}@huawei.com

Abstract

Contextual word embeddings extracted from pre-trained models have become the basis for many downstream NLP tasks, including machine translation automatic evaluations. Metrics that leverage embeddings claim better capture of synonyms and changes in word orders, and thus better correlation with human ratings than surface-form matching metrics (e.g. BLEU). However, few studies have been done to examine robustness of these metrics. This report uses a challenge set to uncover the brittleness of reference-based and reference-free metrics. Our challenge set[1] aims at examining metrics' capability to correlate synonyms in different areas and to discern catastrophic errors at both word- and sentence-levels. The results show that although embedding-based metrics perform relatively well on discerning sentence-level negation/affirmation errors, their performances on relating synonyms are poor. In addition, we find that some metrics are susceptible to text styles so their generalizability compromised.

1 Introduction

Automatic metrics compare machine-translated results with human-translated references or/and sources, and give scores accordingly. Such metrics offer a quick and inexpensive approach for researchers to evaluate model performances. Among these metrics, BLEU (Papineni et al., 2002) has dominated the area for twenty years since its birth in 2002. However, its limitations are obvious: (1) it weighs each word equally but in fact the entropy of each word varies; (2) it only counts n-grams that are exact in the reference and thus synonyms and elaborations are wrongly punished (Smith et al., 2016). Consequently, the correlation between BLEU and human evaluation is relatively low, which sometimes puzzles researchers.

In recent years, embedding-based approaches have been introduced to design new automatic metrics. These metrics, e.g. BERTScore (Zhang et al., 2019), COMET (Rei et al., 2020a), and BLEURT (Sellam et al., 2020a), claim better ability to capture synonyms and changes in word order, and thus better performance than BLEU. Apart from ref-based metrics, researches on quality estimation (QE) have been rising, as QE is an cheaper and more convenient approach considering no need of human-translated references.

In the WMT metric task, correlation with human annotators is the major indicator to evaluate metric performance (Freitag et al., 2021). However, in addition to that, a good metric should meet the following requirements (Banerjee and Lavie, 2005; Koehn, 2009): (1) sensitivity to nuances in quality among systems or outputs of the same system in different stages of its development so it can be used to direct system performance optimization; (2) consistency and reliability of scores; (3) usability in a great range of fields; (4) speed; (5) low cost. We believe the first three aforementioned requirements are crucial for judging metric performance as well. So we build a Zh→En challenge set to evaluate metrics' capability in these regards. Section 2 offers a brief description of metrics to be evaluated. Details of our challenge set are described in Section 3. Section 4 presents experiment results and Section 5 discusses our findings.

2 Metrics To Be Evaluated

2.1 Surface-Form Matching Metrics

Reference-based metrics measure the similarity between MT outputs and human translations, and believe that high similarity means high quality and vice versa. In the pre-neural era, metrics calculate the similarity based on surface forms and word stems. Two examples that fall into this category and used in this task as baselines are BLEU (Papineni

[1]We open-source our challenge set at: https://github.com/HwTsc/Challenge-Set-for-MT-Metrics

Proceedings of the Seventh Conference on Machine Translation (WMT), pages 530 - 540

et al., 2002) and chrF (Popović, 2015).

BLEU BLEU computes precision by comparing the n-gram of hypothesis with n-gram of the reference, coupled with a brevity penalty. In this task, sentence-level BLEU (SENT-BLEU) is used.

chrF chrF computes F1 score based on character-level n-grams instead of word-level n-grams.

2.2 Embedding-based Metrics

In the neural era, by leveraging pre-trained word embeddings, new metrics claim better understanding of sentence meanings and thus fare better in evaluation tasks. Some of the well-known metrics that fall into this category and used as baselines in this task include:

BERTScore BERTScore (Zhang et al., 2019) outputs F1 score by calculating token similarity based on contextual embeddings extracted from BERT.

BLEURT-20 BLEURT (Sellam et al., 2020a) is a BERT-based regression model trained on rating data. BLEURT-20 (Sellam et al., 2020b), which is fine-tuned based on Rebalanced mBERT is used in this task.

COMET-20 COMET (Rei et al., 2020a) employs the estimator-predictor architecture and leverages both source and reference information to assess translation quality. COMET-20 (Rei et al., 2020b), which utilizes XLM-RoBERTa, is used in this task.

Yisi-1 Yisi (Lo, 2019) measures semantic similarity between hypothesis and references. Yisi-1 (Lo, 2020) leverages contextual embeddings extracted from language models to compute the idf-weighted lexical semantic similarities.

2.3 QE as Metrics

Quality estimation approach evaluates machine translation quality totally without human intervention. It scores model outputs by leveraging information in source text. Among the seven baseline metrics, COMET-QE (Rei et al., 2021) is a reference-free version of COMET and thus falls into the QE category.

Table 1 is a summary of the seven baselines.

2.4 Participants in WMT22 Metric Task

The challenge set is also used to measure performances of metrics submitted to the WMT22 Metric Task, including twelve reference-based (ref-based) metrics: COMET-22, MATASE, three variants of MEE, four variants of Metricx, ME-COMET-22, two variants of UniTE; and nine QE metrics: COMET-Kiwi, Cross-QE, HWTSC-Teacher-Sim, HWTSC-TLM, KG-BERTScore, MATESE-QE, MS-COMET-QE, REUSE, and UniTE-src.

For details about their implementations, please refer to their system reports and summary report of WMT22 Metrics Task[2].

Metrics	Surface	CWE	Source	Ref	Rating
BLEU	Yes	No	No	Yes	No
chrF	Yes	No	No	Yes	No
BERTScore	No	Yes	No	Yes	No
BLEURT-20	No	Yes	No	Yes	Yes
COMET-20	No	Yes	Yes	Yes	Yes
COMET-QE	No	Yes	Yes	No	Yes
YISI-1	No	Yes	No	Yes	No

Table 1: A comparison of seven baseline metrics from aspects of whether they use surface form (Surface), contextual word embedding (CWE), Source text (Source), Target text (Ref), and human rating data (Rating).

3 Challenge Set & Method

3.1 Source of the Challenge Set

We build our Zh-En challenge set to evaluate metrics' ability to relate synonyms and identify crucial mistakes. The set is built based on two open-source test sets: Flores 101 (Goyal et al., 2022) En-Zh subset (but used as a Zh-En test set in this task) and WMT21 Zh-En news dev + test sets (Akhbardeh et al., 2021). We particularly pick up an En-Zh test set and a Zh-En test set because neural-based metrics may be style-sensitive (Hanna and Bojar, 2021): the English side of the En-Zh test set is natural language while that of the Zh-En test set is translation results, which may suffer from translationese. In addition, WMT sets focus on news domain while Flores is extracted from Wiki. We try to understand whether reference style and domain might influence metric performance so as to evaluate the generalizability of metrics.

3.2 Challenge Set Description

Our test set has 721 test cases and focuses on five categories of errors: (1) number; (2) date & time (D/T); (3) named-entity & terminology (NE&Term); (4) unit; and (5) affirmation/negation

[2]At the time of writing, we have not received descriptions from every participant.

Phenomenon	Flores	WMT	Overall
Number	183	172	355
D/T	50	90	140
NE&Term	68	42	110
Unit	23	35	58
AFF/NEG	58	0	58
Overall	382	339	721

Table 2: Challenge set composition

(AFF/NEG). Each case contains a source text, a reference, a good translation, a bad translation, a language phenomena label and a source of origin label indicating where the sentence comes from. Table 2 details the set composition. The first four categories focus on word-level crucial errors. If such information is translated wrong, human annotators will assign relatively low scores since the audience will be misled by such mistakes. In addition, the four categories feature rich types of expressions. For instance, a number can be presented in either numeral or number format; unit, named entity and terminology have widely-used abbreviations. We try to analyze whether metrics are able to relate synonyms and punish errors the way human annotators do. The last category – affirmation/negation – deals with phrase- to sentence-level errors and tests whether metrics are able to capture the overall meaning of a sentence.

Since both sets provide only one translation result for each sentence, to generate an additional translation result, we employ a group of six in-house translators to post-edit MT results generated by our in-house model. We adopt List-based Attack (LIST) (Alzantot et al., 2018) to generate adversarial examples. LIST replaces word(s) in a candidate sentence with a list of similar words to construct adversarial examples. We use semi-auto and human-craft approaches to extract related sentences from the original data sets. We replace key words in those sentences to ensure that key information in references and good examples are semantically equal but in different formats, and that in references and adversarial examples are semantically different but in the same "surface" format.

Table 3 shows an example of our challenge set. The test case contains a source sentence, a reference, a good-translation that contains a correct translation for a language phenomenon, and an incorrect-translation with an error accordingly. Phenomenon to be evaluated and source of the sentence

SRC:	在已知的大约24,000 块坠落至地球的陨石中，经核实只有***34*** 块是来自火星。
REF:	Out of the approximately 24,000 known meteorites to have fallen to Earth, only about ***34*** have been verified to be martian in origin.
GOOD:	Of the roughly 24,000 meteorites known to have fallen to Earth, only ***thirty-four*** have been confirmed to have come from Mars.
BAD:	Of the roughly 24,000 meteorites known to have fallen to Earth, only ***30*** have been confirmed to have come from Mars.

Table 3: A case of number in different formats. GOOD refers to good translation and BAD refers to the adversarial example.

are also labelled in our challenge set. In this case, it is number and comes from Flores. For more examples, please see Appendix A.

3.3 Measurement

Kendall's tau-like correlation (Freitag et al., 2021) is used to evaluate metric performance. A good translation has higher quality than the corresponding bad one, so a good metric should assign a higher score to the good translation. If a metric does so, we label the metric "Concordant" on the case, and "Discordant" vice versa. The correlation is calculated based on the following formula:

$$\tau = \frac{Concordant - Discordant}{Concordant + Discordant}$$

4 Result

Table 4 presents the results of our challenge set. In general, 20 out of the 28 metrics struggle to discriminate between good and adversarial examples as they fail to achieve a medium correlation (above 0.4) with human annotators. The 8 metrics that manage to achieve medium-level correlation including: BLEURT-20 (baseline), four variants of Metricx (ref-based), HWTSC-Teacher-Sim (QE), KG-BERTScore (QE), and REUSE (QE).

4.1 Comparison across types of metrics

In general, embedding-based metrics perform much better than merely surface-form matching

Metric	Overall	Number	D/T	NE&Term	Unit	AFF/NEG
SENT-BLEU	-0.717	-0.735	-0.743	-0.691	-0.621	-0.690
chrF	-0.393	-0.301	-0.300	-0.745	-0.655	-0.241
BERTScore	-0.193	-0.149	-0.429	-0.291	-0.483	0.586
BLEURT-20	0.495	0.476	0.629	0.364	0.310	0.724
COMET-20	-0.132	-0.093	-0.400	-0.200	-0.414	0.690
COMET-QE	0.090	0.048	-0.343	0.400	0.069	0.828
Yisi-1	-0.140	-0.138	-0.271	-0.291	-0.379	0.690
Baseline Avg.	**-0.141**	**-0.128**	**-0.265**	**-0.208**	**-0.310**	**-0.369**
COMET-22	0.331	0.206	0.500	0.327	0.138	0.897
MATESE	-0.476	-0.673	-0.429	-0.109	-0.414	-0.138
MEE	-0.667	-0.662	-0.700	-0.564	-0.897	-0.586
MEE2	0.060	0.251	0.229	-0.600	-0.345	0.138
MEE4	0.171	0.307	0.443	-0.473	-0.138	0.207
metricx_xl_DA	0.778	0.746	0.900	0.727	0.586	0.966
metricx_xl_MQM	0.781	0.685	0.900	0.818	0.828	0.966
metricx_xxl_DA	0.822	0.820	0.800	0.800	0.828	0.931
metricx_xxl_MQM	0.870	0.865	0.829	0.873	0.897	0.966
MS-COMET-22	0.012	-0.054	-0.143	0.055	-0.103	0.828
UniTE	0.287	0.177	0.500	0.200	-0.069	0.966
UniTE-ref	0.343	0.234	0.529	0.327	0.000	0.931
Ref-based Avg.	**0.276**	**0.242**	**0.363**	**0.198**	**0.109**	**0.589**
COMET-Kiwi	0.337	0.177	0.243	0.582	0.483	0.931
Cross-QE	0.340	0.245	0.171	0.473	0.448	0.966
HWTSC-Teacher-Sim	0.445	0.504	0.314	0.309	0.345	0.759
HWTSC-TLM	0.393	0.425	0.271	0.364	0.310	0.621
KG-BERTScore	0.445	0.493	0.286	0.491	0.138	0.759
MATESE-QE	-0.675	-0.735	-0.771	-0.400	-0.690	-0.586
MS-COMET-QE	0.146	0.059	0.114	0.127	0.000	0.931
REUSE	0.528	0.577	0.657	0.291	0.241	0.655
UniTE-src	0.268	0.104	0.314	0.473	0.103	0.931
QE Avg.	**0.247**	**0.206**	**0.178**	**0.301**	**0.153**	**0.663**

Table 4: Kendall's tau-like correlation results of each metric on our challenge set. The horizontal lines delimit baseline metrics (top), participating ref-based metrics (middle), and participating QE metrics (bottom).

metrics. Ref-based QE metrics perform slightly better than QE metrics. Regarding the two surface-form matching metrics, character-level chrF performs much better than SENT-BLEU on AFF/NEG, Number and D/T test cases, although slightly worse on the other two categories. The performances of embedding-based metrics vary greatly across both ref-based and QE metrics.

4.2 Comparison across error categories

Embedding-based metrics perform well on AFF/NEG cases as we assumed, as most embedding-based metrics (both ref-based and QE) achieve medium to strong correlations with human ranking. However, regarding the other four categories on word-level crucial errors, performances of some embedding-based metrics deteriorate significantly and only few metrics manage to reach medium-level correlation.

5 Discussion

5.1 Number as A Tough Issue

One of the focuses of our challenge set is number. Numbers are dispersed, rich in format, and semantically similar, making metrics hard to grasp the exact meaning. To analyze how metrics perceive and score numbers, we further divide it into four sub-categories:

- Same Format (SAME): Good and bad exam-

Metric	SAME	DIFF	SWAP	SEP
STEN-BLEU	-0.908	-0.807	-0.333	-0.630
chrF	-0.333	-0.572	-0.286	0.210
BERTScore	0.632	-0.393	-0.476	-0.383
BLEURT-20	0.678	0.490	0.000	0.481
COMET-20	0.011	-0.559	-0.095	0.630
COMET-QE	-0.034	-0.159	0.000	0.531
Yisi-1	0.586	-0.379	-0.476	-0.309
Baseline Avg.	**0.090**	**-0.340**	**-0.238**	**0.076**
COMET-22	0.747	-0.103	-0.143	0.358
MATESE	-0.839	-0.710	-0.857	-0.333
MEE	-0.701	-0.876	-0.810	-0.160
MEE2	0.747	0.283	-0.571	0.086
MEE4	0.816	0.421	-0.571	0.012
metricx_xl_DA	0.954	0.862	0.190	0.605
metricx_xl_MQM	0.770	0.724	0.571	0.580
metricx_xxl_DA	0.977	0.903	0.762	0.531
metricx_xxl_MQM	0.931	0.890	0.905	0.728
MS-COMET-22	0.057	-0.103	-0.190	-0.012
UniTE	0.655	-0.103	-0.381	0.457
UniTE-ref	0.655	-0.076	-0.190	0.556
Ref-based Avg.	**0.481**	**0.176**	**-0.107**	**0.284**
COMETKiwi	0.425	-0.090	0.048	0.457
Cross-QE	0.218	0.145	-0.095	0.630
HWTSC-Teacher-Sim	0.632	0.503	0.000	0.630
HWTSC-TLM	0.448	0.393	0.238	0.556
KG-BERTScore	0.655	0.503	-0.143	0.630
MATESE-QE	-0.839	-0.821	-0.762	-0.457
MS-COMET-QE-22	-0.011	0.034	-0.095	0.259
REUSE	0.747	0.641	-0.048	0.605
UniTE-src	0.264	-0.241	-0.143	0.679
QE Avg.	**0.282**	**0.119**	**-0.111**	**0.443**

Table 5: Kendall's tau-like correlation results on our challenge set. The horizontal lines delimit baseline metrics (top), participating reference-based metrics (middle), and participating QE metrics (bottom).

ples use different numbers in the same format (e.g. 1 & 2; three & four).

- Different Format (DIFF): The good examples contain correct numbers in a different format as reference while the bad examples contains an incorrect number in the same format as reference (e.g. 1 & two; 1,000,000 & 1 million).
- Swapped Number (SWAP): When a sentence contains two or more numbers, we swap the numbers to generate the bad translation.
- Thousand Separator (SEP): Thousand separators are not required but help improve readability. Test sets under this category compare numbers without thousand separators with those in wrong formats (e.g. 1 000; 10,00; 1.000).

Table 5 presents results on the number subcategories. According to the table, surface-form matching metrics perform worse under all the four subcategories. Although embedding-based metrics in general perform much better, those metrics still perform worse under the SWAP subcategory.

5.1.1 Numeral vs. Number

In daily usage, there is no strict rule about when to use numerals or numbers. In some cases, numeral and number are just two different symbols to express the same meaning and as a result can be regarded as synonyms. According to table 5, the majority of embedding-based metrics perform relatively well on discerning differences among numerals or among numbers (SAME). To be more specific, if the good and adversarial examples contain different numbers in the same format, even if the format is different from that used in the reference, the possibility for metrics to discern between the correct and incorrect numbers is relatively high.

However, performances of these metrics under the DIFF category deteriorate to varying extents. In other words, if the good example contains a correct number in different format and the adversarial example contains an incorrect number but in the same format as that in the reference, metrics are likely to assign a higher score to the adversarial example.

The result demonstrates that contextual embeddings fail to relate semantically similar numbers and numerals. Instead, they seem to rely more on the "surface similarity". Another example to buttress this assumption is the metrics' performances in the thousand separator category, where there is about 50% of chance that metrics score numbers with wrong separator formats higher than those without separators.

Although neural machine translation models seldom translate numbers wrong, outputs do use different number formats. When these metrics are used to measure model performances, they incline to wrongly penalize sentences using a different number format, thus leading to unfair evaluations.

5.1.2 Does Number Difference Count?

We further conducted two experiments to examine if metrics' capability of distinguishing numbers improves when the difference between the correct and incorrect numbers turns greater. The sentence shown in Table 3 is used for the two experiments.

In the first experiment, we replace the number in the reference (REF in table 3) to its numeral format "thirty-four" and denote the sentence as good-

translation x1. Then we replace the number in the reference to other Arabic numbers ranging from 1 to 100 to generate a set of comparative candidates denoted as bad-translations Y{y_1,y_2,...y_{100}}.

In the second experiment, we denote another correct post-edit result as good-translation x2 (GOOD in table 3), and alter the numeral in x2 to Arabic numbers ranging from 1 to 100 (denoted as bad-translations Z{z_1,z_2,..., z_{100}}).

We calculated BERTScore of x1, x2, Y and Z against the reference and the result is presented in figure 1. When there is no other difference between reference and candidates except the number, it seems easier for BERTScore and BLEURT to discern number differences even in different formats. In addition, as the difference between numbers becomes greater, the gap of scores expands. However, when there are other differences between the reference and candidates, it becomes harder for BERTScore to quantify the error, as BERTScore gives the majority of candidates in Z higher scores than x2. And greater difference between numbers seems not help. However, BLEURT remains a good performance in the second experiment, which is consistent with our challenge test results.

5.1.3 Do Metrics Understand Number?

Another interesting finding regarding number is that all metrics perform badly under the SWAP category (only three ref-based metrics managed to achieve medium-level correlation). Swapping two numbers in a sentence causes drastic changes in meaning but metrics lack the capability to identify such changes.

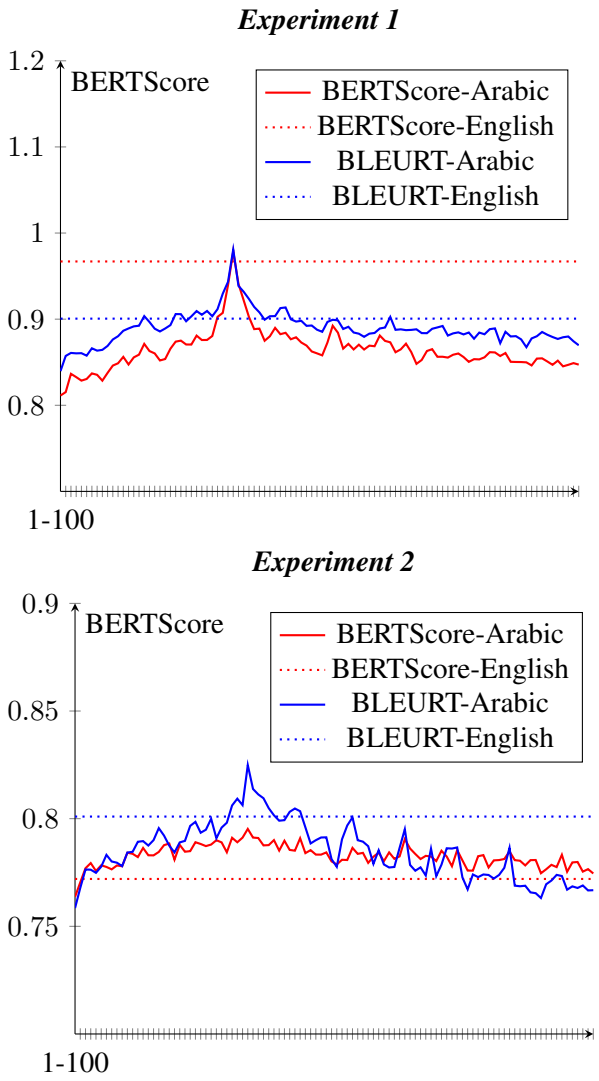

Figure 1: Results for experiment 1 and 2. The dotted lines indicate the scores for x1 and x2, while the solid lines represent the results of Y and Z.

5.2 Is Source/Ref Information Helpful?

In general, QE metrics perform relatively worse than ref-based metrics, but the gap is smaller than we assumed. By just leveraging source-side information, the average of QE metrics almost reaches medium-level correlation. This gives rise to a question: if a metric leverages both source-side and target-side information, will the accuracy improve?

The implementations of COMET-22 and COMET-Kiwi are almost the same but COMET-22 leverages both source-side and target-side text while COMET-Kiwi uses only source-side text. When we compare the performances of the two metrics, we find that COMET-22 outperforms COMET-Kiwi under the Number and D/T categories. However, COMET-Kiwi outperforms COMET-22 under the NE&Term, Unit and AFF/NEG categories. The result indicates that while in some cases, reference-side information helps improve accuracy; in other cases, reference-side information surprisingly causes performance deterioration.

Among all the participating ref-based metrics, although some leverage source-side information while the others do not, their implementations vary. So we are unable to draw a conclusion that whether adding source-side information to a ref-based metric helps improve accuracy. More ablation experiments are required.

5.3 Synonym is Still A Tough Issue

Although embedding-based metrics claim better capture of synonyms, the result shows that there is still a long way to go. Not only numbers, test cases under NE&Term, D/T, and Unit categories

all aim at examining metrics' ability to relate different formats of words that express the same meaning. The results show that metric performances vary greatly under these categories. The variations demonstrate that there should be a solution to this problem. However, at the time of writing, we have no detailed information about the implementations of those well-performed metrics. For more details, please refer to WMT Metric summary report and their system reports.

Metric	FLORES	WMT
BLEU	-0.728	-0.705
chrF	-0.398	-0.386
BERTScore	-0.073	-0.327
BLEURT-20	0.529	0.457
COMET-20	-0.042	-0.233
Yisi-1	-0.016	-0.280
COMET-QE	0.225	-0.062
Baseline Avg.	**-0.072**	**-0.220**
COMET-22	0.450	0.198
MATESE	-0.492	-0.457
MEE	-0.644	-0.693
MEE2	0.047	0.074
MEE4	0.178	0.162
metricx_xl_DA$_2$019	0.801	0.752
metricx_xl_MQM$_2$019	0.796	0.764
metricx_xxl_DA	0.848	0.794
metricx_xxl_MQM	0.911	0.823
MS-COMET-22	0.084	-0.068
UniTE	0.398	0.162
UniTE-ref	0.435	0.239
Ref-based Avg.	**0.318**	**0.229**
COMETKiwi	0.450	0.209
Cross-QE	0.445	0.221
HWTSC-Teacher-Sim	0.450	0.440
HWTSC-TLM	0.393	0.392
KG-BERTScore	0.487	0.398
MATESE-QE	-0.649	-0.705
MS-COMET-QE-22	0.215	0.068
REUSE	0.571	0.481
UniTE-src	0.335	0.192
QE Avg.	**0.300**	**0.188**

Table 6: A comparison of metric performances on Flores and WMT test cases. The horizontal line delimit baseline metrics (top) and participating reference-based metrics (bottom).

5.4 Do Metrics Suffer from Domain Issue?

We build our challenge set based on two open-source test sets: Flores 101 and WMT21 Zh-En. Hanna and Bojar (2021) claim that when the reference is a post-edit, BERTScore performs poorly as the post-edit may have high lexical overlap with machine translations. In our experiment setting, the candidate sentences are post-edits, which are stylistically similar to references in the WMT21 Zh-En news test sets, as the references are translations provided by professional translators. On the contrary, the Flores 101 En-Zh test set is translated from English to Chinese, so the English side is original and less stylistically similar to post-edits.

We calculate each metric's performance on Flores and WMT test cases (see table 6). Surface-form matching metrics are least influenced by the difference. For both ref-based and ref-free metrics, while some metrics (e.g. Metricx, HWTSC-Teacher-Sim) remain almost same performance on cases from the two sources, some metrics (e.g. COMET-22, UniTE) perform far worse on WMT cases.

The result shows that the generalizability of metrics varies. While good metrics can remain the same performance on test sets in different domains and of different styles, some metrics suffer greatly from domain issues. We assume the reasons for such performance gaps including: 1) WMT test cases are longer than Flores cases in average, making the cases harder to score; 2) big data for training pre-trained models are mostly native monolinguals so these models are better at encoding native languages than "translationese". However, more ablation experiments are required and generalizability should be concerned when developing metrics.

6 Conclusion

This paper presents our submitted challenge set to the WMT22 Metrics Challenge Sets Subtask and various metrics' performances on our set. Our set focuses on five categories of errors and the result shows that while most metrics are able to identify catastrophic sentence-level affirmation/negation errors, some metrics fail at discerning word-level keyword errors and capturing synonyms of such words. The results show that references are not always useful for a metric to identify errors. In addition, generalizability of metrics should be considered as some metrics are susceptible to test sets styles. The majority of metrics fail to meet the requirements (Banerjee and Lavie, 2005; Koehn,

2009) we discuss in the introduction section. They fail to identify nuances in quality and provide reliable scores, and suffer from domain issues as well.

The limitation of this research is that all of the perturbations are human-crafted, and these errors may seldom occur in neural machine translations. To further analyze metric performance in real settings, we will try to annotate and categorize real machine translation errors and evaluate metric performance accordingly.

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.

Moustafa Alzantot, Yash Sharma, Ahmed Elgohary, Bo-Jhang Ho, Mani Srivastava, and Kai-Wei Chang. 2018. Generating natural language adversarial examples. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2890–2896, Brussels, Belgium. Association for Computational Linguistics.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

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

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

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.

Philipp Koehn. 2009. *Statistical Machine Translation*. Cambridge University Press.

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

Chi-kiu Lo. 2020. Extended study on using pretrained language models and YiSi-1 for machine translation evaluation. In *Proceedings of the Fifth Conference on Machine Translation*, pages 895–902, 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.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Ricardo Rei, Ana C Farinha, Chrysoula Zerva, Daan van Stigt, Craig Stewart, Pedro Ramos, Taisiya Glushkova, André F. T. Martins, and Alon Lavie. 2021. Are references really needed? unbabel-IST 2021 submission for the metrics shared task. In *Proceedings of the Sixth Conference on Machine Translation*, pages 1030–1040, Online. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020a. Comet: A neural framework for mt evaluation. *arXiv preprint arXiv:2009.09025*.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020b. Unbabel's participation in the wmt20 metrics shared task. *arXiv preprint arXiv:2010.15535*.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020a. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

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

Aaron Smith, Christian Hardmeier, and Joerg Tiedemann. 2016. Climbing mont BLEU: The strange world of reachable high-BLEU translations. In *Proceedings of the 19th Annual Conference of the European Association for Machine Translation*, pages 269–281.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

A Appendix

SRC:	死亡人数至少为***15*** 人，预计还会增加。
REF:	The death toll is at least ***15***, a figure which is expected to rise.
GOOD:	The death toll is at least ***fifteen*** and is expected to rise.
BAD:	The death toll is at least ***fourteen*** and is expected to rise.
Phenom:	Number (Same Format)
Source:	Flores

SRC:	湖南红色旅游文化节已成功举办***16***届，是全国红色旅游的知名品牌。
REF:	The Hunan Red Tourism and Culture Festival has been successfully held for ***16*** years, making it a famous red tourism brand in China.
GOOD:	Hunan Red Tourism Culture Festival has been successfully held for ***sixteen*** times and is a well-known brand of red tourism in China.
BAD:	Hunan Red Tourism Culture Festival has been successfully held for ***14*** times and is a well-known brand of red tourism in China.
Phenom:	Number (Different Format)
Source:	WMT

SRC:	投票存在两极分化的情况，***29%*** 的受访者认为澳大利亚应该尽快成立共和国，***31%*** 的人则认为澳大利亚永远不应该成立共和国。
REF:	At the extremes of the poll, ***29 per cent*** of those surveyed believe Australia should become a republic as soon as possible, while ***31 per cent*** believe Australia should never become a republic.
GOOD:	The vote was polarised, with ***29%*** of respondents saying Australia should become a republic as soon as possible and ***31%*** saying it should never become a republic.
BAD:	The vote was polarised, with ***31%*** of respondents saying Australia should become a republic as soon as possible and ***29%*** saying it should never become a republic.
Phenom:	Number (Swapped Number)
Source:	Flores

SRC:	“我是***7***月***7***日来北京的，当时其实有点担心疫情，还提前三天做了核酸检测，是带着酒精棉和检测报告来布展的。”
REF:	“I arrived in Beijing on ***July 7***, and at the time I was a little worried about the pandemic, so took the nucleic acid test three days in advance, and I came here with alcohol pads and my test report. ”
GOOD:	"I arrived in Beijing on ***the 7th of July***. At that time, I was a little worried about the pandemic so I did a nucleic acid test three days in advance, and I took alcohol pads and the test report to set up the exhibition."
BAD:	"I arrived in Beijing on ***June 7***. At that time, I was a little worried about the pandemic so I did a nucleic acid test three days in advance, and I took alcohol pads and a test report to set up the exhibition."
Phenom:	Date & Time
Source:	WMT

SRC:	除了大件，让傅昆宝两口子头疼的还有家里***1000***多斤粮食和新买的一些家具。
REF:	Apart from the large items, the over ***1,000*** jin (500 kg) of grain and newly bought furniture was also a headache Fu Kunbao and his wife.
GOOD:	In addition to big items, Fu Baokun and his wife don’t know how to deal with more than ***1000*** jin of grain and some newly bought furniture in the home.
BAD:	In addition to big items, Fu Baokun and his wife don’t know how to deal with more than ***1.000*** Jin of grain and some newly bought furniture in the home.
Phenom:	Number (Thousand Separator)
Source:	WMT

SRC:	美国地质调查局国际地震地图显示，冰岛在前一周并未发生地震。
REF:	The ***United States*** Geological Survey international earthquake map showed no earthquakes in Iceland in the week prior.
GOOD:	The ***U.S.*** Geological Survey International Earthquake Map shows no earthquakes in Iceland in the previous week.
BAD:	The ***United Kingdom*** Geological Survey International Earthquake Map shows no earthquakes in Iceland in the previous week.
Phenom:	Named Entity & Terminology
Source:	Flores

SRC:	到今天早些时候，风速为每小时***83*** 公里左右，预计会不断减弱。
REF:	By early today, winds were around 83 ***km/h***, and it was expect to keep weakening.
GOOD:	By early today, the wind speed was about 83 ***kilometers per hour***, and it is expected to continue to weaken.
BAD:	By early today, the wind speed was about 83 ***m/h***, and it is expected to continue to weaken.
Phenom:	Unit Format
Source:	Flores

SRC:	不久前，他在布里斯班公开赛上败于拉奥尼奇。
REF:	He ***recently*** lost against Raonic in the Brisbane Open.
GOOD:	***Not long ago***, he lost against Raonic at the Brisbane International tournament.
BAD:	***Long ago***, he lost against Raonic at the Brisbane International tournament.
Phenom:	Unit Format
Source:	Flores