Robust MT evaluation with Sentence-level Multilingual Augmentation

Duarte M. Alves[*1], **Ricardo Rei**[1,3,4], **Ana C. Farinha**[3],
José G. C. de Souza[3], **André F. T. Martins**[1,2,3]
[1]Instituto Superior Técnico, University of Lisbon, Portugal
[2]Instituto de Telecomunicações, Lisbon, Portugal
[3]Unbabel, Lisbon, Portugal, [4]INESC-ID, Lisbon, Portugal

Abstract

Automatic translations with critical errors may lead to misinterpretations and pose several risks for the user. As such, it is important that Machine Translation Evaluation systems are robust to these errors in order to increase the reliability and safety of the translation process. Here we introduce SMAUG, a novel Sentence-level Multilingual AUGmentation approach for generating translations with critical errors and apply this approach to create a test set to evaluate the robustness of Machine Translation metrics to these errors. We show that current State-of-the-Art methods are improving their capability to distinguish translations with and without critical errors and to penalize the first accordingly. We also show that metrics tend to struggle with errors related to named entities and numbers and that there is a high variance in the robustness of current methods to translations with critical errors.

1 Introduction

In recent years, Machine Translation (MT) systems have been used in diverse real world environments. However, widespread adoption of these systems raises many concerns, namely in the quality of their outputs. Ideally, human translators would evaluate generated translations but this process is expensive and slow. As an alternative, automatic Machine Translation Evaluation relies on external systems to measure the quality of generated translations.

As a crucial aspect of Machine Translation Evaluation, it is vital to ensure that generated sentences do not contain critical errors. As detailed in Specia et al. (2021), translations with such errors deviate in meaning from their source sentence in ways that may lead to misinterpretations and pose health, safety, legal, reputation, religious or financial implications. Specia et al. (2021) group these translations into three categories, based on how their meaning deviates from the source sentence. Mistranslation errors have critical content in the source sentence translated into a different meaning, not translated (the content remains in the source language), or translated into gibberish. Hallucination errors introduce content in the translated sentence that is not present in the source sentence. Deletion errors exclude important content from the source sentence.

In this work, we propose SMAUG[1], a Sentence-level Multilingual AUGmentation framework to generate translations with critical errors, targeting all the aforementioned critical error categories.

We also introduce a novel test set to analyse the robustness of MT Evaluation systems to critical errors. This test set was created with the proposed augmentation framework and submitted to the WMT22 Challenge Sets Sub-task (Freitag et al., 2022).

Finally, we present the results obtained from evaluating metrics submitted to the WMT22 Metrics Shared Task with the developed test set. We show progress of submitted metrics with respect to baseline systems, particularly concerning Quality Estimation systems. Namely, we demonstrate that several metrics are able to correctly distinguish translations with and without critical errors and to penalize the former. Furthermore, we show that current metrics are less sensitive to translations containing errors in named entities and numbers and that there is a high variance in the performance of current SOTA evaluation metrics with respect to identifying and penalizing the occurrence of critical errors.

2 Related Work

Metrics for Machine Translation Evaluation produce a quality score for a given hypothesis, based on the source sentence and a possibly empty set

[*] Corresponding author: duartemalves@tecnico.ulisboa.pt

[1] Code available at: https://github.com/Unbabel/smaug

Proceedings of the Seventh Conference on Machine Translation (WMT), pages 469–478
Abu Dhabi, December 7–8, 2022. ©2022 Association for Computational Linguistics

of reference translations. These metrics can be divided into two main groups, given their reference set. Reference based metrics have a non-empty reference set, while reference free metrics have an empty reference set. Reference free evaluation is also denominated by Quality Estimation.

Within reference-based metrics, n-gram based metrics, such as BLEU (Papineni et al., 2002) and CHRF (Popović, 2015), measure lexical overlap from the hypothesis to the human references. Rei et al. (2020) advocate that these methods fail to capture semantic similarities beyond the lexical level. Their inability to capture meaning at a sentence level also makes them unfit for the detection of critical errors as they equally penalize the usage of synonyms or the mistranslation of a named entity.

As an alternative to n-gram matching, more recent methods leverage word representations to capture semantic similarities beyond the lexical level. As described in Rei et al. (2020), *embedding*-similarity methods, like YISI-1 (Lo, 2019) and BERTSCORE (Zhang et al., 2020), create an alignment between the vector representations of the words in the hypothesis and the reference and then compute a score that captures the semantic similarity between both sentences. As noted by Rei et al. (2020), the main issue with these approaches is that human judgements consider other information beyond semantic similarity, limiting the correlation of these methods with human evaluations.

More recently, learnt methods, such as BLEURT20 (Sellam et al., 2020) and COMET (Rei et al., 2020), address this issue by training to directly maximize correlation with human judgements. Results from the WMT21 Metrics (Freitag et al., 2021) and the WMT21 Quality Estimation (Specia et al., 2021) shared tasks suggest that these methods obtain higher correlations with human judgements, such as Direct Assessments (Graham et al., 2013), Human Translation Edit Rate (HTER) (Snover et al., 2006) or Multi-dimensional Quality Metrics (MQM) (Lommel et al., 2014).

However, as noted by Ribeiro et al. (2020), relying on accuracy on held-out sets can lead to an overestimation on the performance of NLP models. As such, Ribeiro et al. (2020) proposes `CheckList` that relies on data augmentation techniques to create examples that test specific behaviours of NLP systems in various situations. Within the field of Machine Translation Evaluation, as a case study for exploring the sensitivity of learnt metrics to specific phenomena, Amrhein and Sennrich (2022) employed Minimum Bayes Risk decoding with COMET as an utility function to identify good hypotheses. The authors show that hypotheses chosen with COMET are more likely to have errors in Named Entities and Numbers when compared to CHRF, indicating the metric is not sensitive enough to these errors.

Considering multiple metrics, Freitag et al. (2021) tested multiple systems on a challenge set with errors related to negation and sentiment polarity and found that most metrics struggle with these errors. Nonetheless, these examples were chosen from existing MT outputs, which can lead to a major human effort, as these errors are not common.

Regarding reference free evaluation, Kanojia et al. (2021) define multiple perturbations to test the robustness of QE systems in detecting specific errors. The authors show that overall the tested perturbations are well detected but some, such as polarity based perturbations, still pose a challenge to QE systems. However, the list of perturbations is not exhaustive and most rely on transformations that do not necessarily preserve the semantics of the phrases, such as random insertions, substitutions and deletions.

3 SMAUG Framework

In order to create an example of a critical error, the proposed framework receives an existing sentence and perturbs it, inducing one of the linguistic phenomena detailed in the following sections. For each linguistic phenomenon, the perturbation process is separated into two phases: transformation and validation. The first phase generates a candidate sentence by perturbing the original translation. This phase may not produce a candidate, as some perturbations are not applicable to all sentences. The second phase verifies whether the produced candidate meets a set of desirable criteria, discarding it otherwise.

3.1 Deviation in Named Entities

The first perturbation replaces a named entity in the original sentence for a different one that is consistent with the original context. The transformation phase of this perturbation, in Figure 1, starts by detecting all Named Entities in the original sentence with the Named Entity Recognition (NER) System in the Stanza library (Qi et al., 2020). If no entity is detected, the generation process stops. Otherwise, a single one is randomly chosen using an Uniform Distribution. This entity is replaced by employing the mT5 pretrained language model (Xue et al., 2021). For this, the span with the sampled entity is replaced by a single mask token and the model is used to generate the candidate sentence. The decoding strategy for the mT5 model is sampling considering the top 50 elements. When compared with other strategies, such as Beam-Search and Top-P sampling, this approach was empirically found to give realistic examples at a lower computational cost. The mT5 model was chosen for three main reasons: it is multilingual and trained on a massive set of different languages; it can generate multiple words from a single mask token, thus not requiring any special strategy for adding mask tokens in order to avoid only single word entities; and does not change the remainder of the sentence, avoiding unwanted side-effects. Nevertheless, the mT5 model was found to often generate punctuation symbols in the beginning of the sentence. In order to increase the credibility of the generated sentences, these symbols were removed.

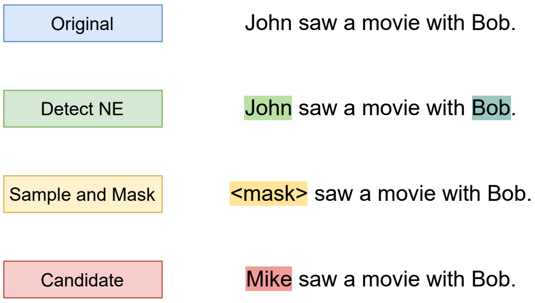
Original
John saw a movie with Bob.
Detect NE
John saw a movie with Bob.
Sample and Mask
<mask> saw a movie with Bob.
Candidate
Mike saw a movie with Bob.

Figure 1: Example of the transformation phase for the Deviation in Named Entities phenomenon.

The validation phase for this perturbation encompasses several sub-validations. On the one hand, in order to ensure the mT5 model generates a named entity, the candidate is only accepted if the above NER model detects the same number of entities in both the candidate and the original sentence. On the other hand, the mT5 model can "guess" the correct named entity from the remaining context. As such, the generated sentence can not be equal to the original. Furthermore, to prevent cases where mT5 produces a small variation of the original entity (for example by adding an hyphen between two words or changing the accentuation), candidate sentences may only be accepted if they have a character-level minimum edit distance to the original above a threshold. This procedure can discard many valid candidates and thus, depending on the desired quality and quantity of generated sentences, may be applied or not. Through manual experimentation, a distance greater or equal to 5 was found to produce a good balance between ensuring the generated entities are different without discarding too many valid candidates. Finally, to increase the overall quality of the generated sentences, several sub-validations can be employed. Candidates with words matching the regular expression of the mT5 masking token (`<extra_id_\d{1,2}>`) are discarded, as they represent cases where the model was unable to generate content. This can be extended by considering more generic expressions such as `extra_*`. Furthermore, since named entities do not usually have characters such as `()[]\{\}_`, candidates that have more of these characters than the original can also be removed. As before, these validations can remove valid candidates and they should be adapted to the use case in question.

3.2 Deviation in Numbers

Another perturbation, similar to the deviation in named entities, replaces a number in the original sentence by a different one. The transformation phase for this phenomenon follows the same procedure as the deviation in Named Entities. However, it employs the regular expression `[-+]?\.?(\d+[.,])*\d+` to detect numbers in the original sentence. From the detected numbers, the process to sample a single number and replace it with another one using the mT5 model is the one described above, from masking the span with the chosen number to generating the candidate sentence.

Regarding the validation phase, it also employs a set of sub-validations. As before, the candidate is accepted only if the regular expression to detect numbers is matched the same number of times in both the original and candidate sentences, ensuring a number was generated. Furthermore, the original and candidate sentences must be different to ensure the mT5 model did not "guess" the number by the context. In this perturbation, the minimum edit distance sub-validation was not applied as small variations in numbers mostly lead to critical errors (for example changing the place of a comma within the number). Finally, candidates matching the mT5 masking token (`<extra_id_\d{1,2}>`) or that introduce one of the following characters

( ) [ ] \ { \ } _ are also removed to increase the overall quality of the generated sentences.

3.3 Deviation in Meaning

Concerning deviations in meaning, a phenomenon that either introduces or removes a negation in the original sentence was developed, thus generating a sentence with the opposite meaning.

In order to negate the original sentences, this perturbation relies on the POLYJUICE (Wu et al., 2021) model conditioned for negation. POLYJUICE can either negate an entire sentence or a span, by masking the sentence or only the desired text span, respectively. Initial experiments showed that, when trying to negate the entire sentence, the model often forgot some content, specially in longer phrases. Thus, the developed approach, shown in Figure 2, masks a verb in the original sentence, as well as any adjacent auxiliary verbs before it, in order to produce a small perturbation that changes the meaning of the sentence. Specifically, the transform used a Part-of-Speech tagger from the Stanza library (Qi et al., 2020) in the original sentence and recovered all spans with 0 or more AUX tags immediately followed by a VERB tag. If no spans are detected, the generation process stops. Otherwise, one span is sampled using an uniform distribution. Finally, the conditioned POLYJUICE model produces the candidate sentence by negating the original sentence with a mask over the chosen span.

The validation phase for this phenomenon first verifies whether the candidate sentence is equal to the original or if the POLYJUICE model produced its empty token, meaning it was unable to generate a sentence. Furthermore, a RoBERTa (Liu et al., 2019) model trained for Multi-Genre Natural Language Inference (MNLI) corpus was used to verify whether the candidate contradicts the original sentence. This procedure is employed as a proxy for validating whether the generated sentence is a negation over the original.

3.4 Insertion of Content

Regarding the generation of Hallucinated content, a phenomenon to insert new content in the original sentence was devised.

The transformation phase of this perturbation employs a similar strategy to the Named Entities phenomenon. In this case, the masking pattern randomly inserts mask tokens between adjacent words in the original sentence. In order to avoid inserting too much content, a maximum of three mask tokens are introduced. After this step, the masked sentence is fed to the mT5 model, which generates the candidate sentence.

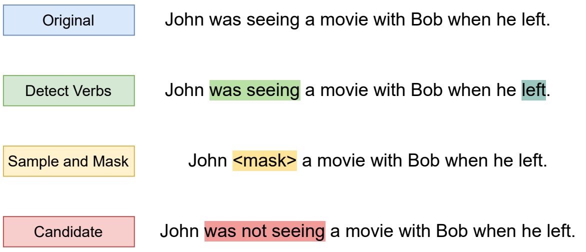
Original
John was seeing a movie with Bob when he left.
Detect Verbs
John was seeing a movie with Bob when he left.
Sample and Mask
John <mask> a movie with Bob when he left.
Candidate
John was not seeing a movie with Bob when he left.

Figure 2: Example of the transformation phase for the Deviation in Meaning phenomenon. Although not shown in this example, the POLYJUICE model receives additional information besides the masked sentence to know the text that was replaced by the mask.

In the validation phase, as in the Named Entities phenomenon, candidate sentences that are equal to the original or that match the regular expression for the mT5 masking pattern (<extra_id_\d{1,2}>) are discarded. Moreover, another sub-validation that ensures the minimum edit distance at a word-level between the candidate and original sentences is above a threshold was applied. As there are only insertions, this sub-validation ensures at least a minimum number of words are introduced in the candidate sentence. Furthermore, higher thresholds increase the likelihood of the candidate sentences having hallucinated content as, with a fixed number of masks (defined in the masking strategy), the model has to generate spans of text with multiple words and it is unlikely that only function words are introduced. Through manual experimentation, a threshold of eight words was found to produce a good balance between ensuring content was added without discarding too many valid candidates.

3.5 Removal of Content

Finally, translations with deletion errors were tackled by a phenomenon that removes a span of text between two punctuation symbols. By considering text spans between adjacent punctuation symbols, this method aims to remove a sub-phrase of the original sentence that likely contains some information, thus generating a sentence which is missing content.

As shown in Figure 3, the transformation phase of this perturbation starts by detecting all instances of the symbols . , ? ! in the original sentence. Then, a span between two adjacent symbols is randomly sampled with an Uniform Distribution. The chosen span, as well as the punctuation symbol after it, are deleted in order to generate the candidate sentence. In order to increase the likelihood of removing content, the deleted span has a minimum number of words. Furthermore, to increase the credibility of the generated sentence, three additional constraints were enforced. First, the first text span was not considered, as translation models are less likely to forget content in the beginning of the sentence. Second, the deleted span has a maximum size, as it is unlikely the translation model drops a large portion of the sentence. Third, if the generated candidate does not end in . ! ? , the final symbol is replaced by a punctuation mark. If no span exists in the previous conditions, the transform does not generate a candidate sentence.

This transform does not require any extra validation, as all verifications are enforced when choosing the text span to delete.

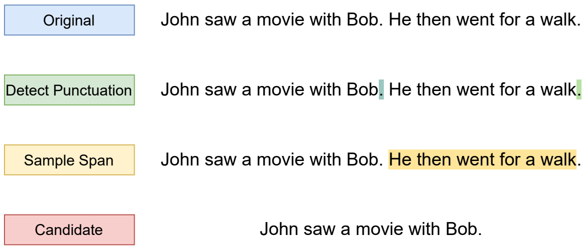
Original
John saw a movie with Bob. He then went for a walk.
Detect Punctuation
John saw a movie with Bob. He then went for a walk.
Sample Span
John saw a movie with Bob. He then went for a walk.
Candidate
John saw a movie with Bob.

Figure 3: Example of the transformation phase for the Removal of Content phenomenon.

4 Challenge Set

The created test set comprises of records in the format $(s, h_{good}, h_{bad}, r, p)$, where s is a source sentence, h_{good} and h_{bad} are "good" and "bad" hypothesis, r is a reference and p is an identifier for the linguistic phenomenon present in h_{bad}. Three language pairs were considered: English-Portuguese, Spanish-English, Portuguese-English. For each language pair, a data augmentation approach was applied to an existing parallel corpus to generate a the final set of records.

4.1 Parallel Corpora

To create our challenge set we extracted sentences from OPUS (Tiedemann, 2012) ranging several domains such as News and Euro Parliament. To guarantee high-quality references we used Bicleaner tool (Ramírez-Sánchez et al., 2020) with a threshold of 0.85.

4.2 Augmentation Approach

For each language pair, the source side of the respective corpus was considered as source sentences and the target as references. First, the source sentences were translated using an OPUS-MT bilingual model (Tiedemann and Thottingal, 2020)[2]. Second, all the perturbations were applied to the references, generating sentences with at most one critical error. This information was aggregated to create records in the format $(s, h_{good}, h_{bad}, r, p)$, where h_{good} is the translation of the source sentence, h_{bad} is a perturbation of the reference and p is the linguistic phenomenon that was induced. With this approach, multiple records can be created from an original source and reference pair, one for each perturbation applied to the reference. In this case, all the records have the same good hypothesis.

The generated records were then manually filtered and validated to ensure its quality. In this process, we ensured that both the references and the good translations were high quality and that the bad translation contained a critical error. Furthermore, we chose records where h_{good} was different from r to force the metrics to attend to the meaning of the sentence instead of analysing lexical overlap. In the end, around 50 records for each phenomenon and language pair were obtained, as shown in Table 1. The Deviation in Named Entities and Meaning phenomena for the English-Portuguese language pair have 0 records since the Portuguese language is not supported by the NER model in the Stanza library or the POLYJUICE model.

5 Experiments

The developed test set was submitted to the WMT22 Challenge Set Sub-task and the scores for several State-of-the-Art metrics were gathered. The following sections detail the evaluation method for the tested metrics and the obtained results.

5.1 Evaluation Method

We rely on two evaluation methods to assess the robustness of metrics to the developed critical errors.

The first is the official evaluation method for the Shared Task in order to compare the performance of the several metrics. This method used a Kendall-Tau like formulation, defined as:

$$\tau = \frac{Concordant - Discordant}{Concordant + Discordant}, \qquad (1)$$

[2] Available at Hugging Face Transformers (Wolf et al., 2020)

en-pt		pt-en		es-en	
Phenomenon	**Size**	**Phenomenon**	**Size**	**Phenomenon**	**Size**
NE	0	NE	50	NE	48
NUM	49	NUM	48	NUM	50
MEAN	0	MEAN	50	MEAN	48
INS	44	INS	48	INS	50
DEL	48	DEL	50	DEL	49

Table 1: Number of selected records for each phenomenon and language pair. The Deviation in Named Entities and Meaning phenomenon have 0 records for the English-Portuguese language pair as the phenomenon do not support to-Portuguese language pairs.

where $Concordant$ is the number of times the metric assigned a higher score to the good hypothesis and $Discordant$ is the number of times the metric assigned a higher score to the bad hypothesis.

The second method measures the average difference between the scores assigned to h_{good} and h_{bad}, when the score assigned to the h_{good} is higher. For a given set S with pairs of scores, this method is defined as

$$d = \frac{\sum\limits_{(s_{good}, s_{bad}) \in S} \mathbb{I}[s_{good} > s_{bad}](s_{good} - s_{bad})}{\sum\limits_{(s_{good}, s_{bad}) \in S} \mathbb{I}[s_{good} > s_{bad}]} \tag{2}$$

where s_{good} and s_{bad} are respectively the scores for multiple good and bad hypothesis pairs. This formulation is used as a proxy for the confidence of the evaluated metric when it assigns a higher score to the good hypothesis. In order to compare multiple metrics with different scoring intervals, the metric scores are normalized before this evaluation method is applied.

5.2 Baseline Metrics

All the baseline metrics from the Sub-task were considered. These comprise of several State-of-the-Art methods: BLEU and CHRF are *n-gram* based metrics; BERTSCORE and YiSi-1 are *embedding*-similarity methods, and BLEURT20, COMET-20 and COMET-QE are learnt methods.

Figure 4 shows the obtained results for these metrics. For each phenomenon, results show the average Kendall-Tau considering all language pairs and the black bars represent the standard deviation. We observe that the metrics obtain mostly negative correlations, indicating they are assigning higher scores to the bad hypothesis. *n-gram* based metrics show the worst correlations. This result is to be expected as the perturbations create localized changes, such as changing a number, which do not significantly modify the alignments with the reference. Embedding-similarity based metrics exhibit a better performance as contextual embeddings can capture divergence in meaning of the bad hypothesis, but still the obtained correlations are mostly negative. Pretrained models obtain the best results, having positive correlations for the phenomenon Deviation in Meaning, Insertion and Removal of Content. Nevertheless, they still show poor correlations and struggle with Deviation in Named Entities and Numbers.

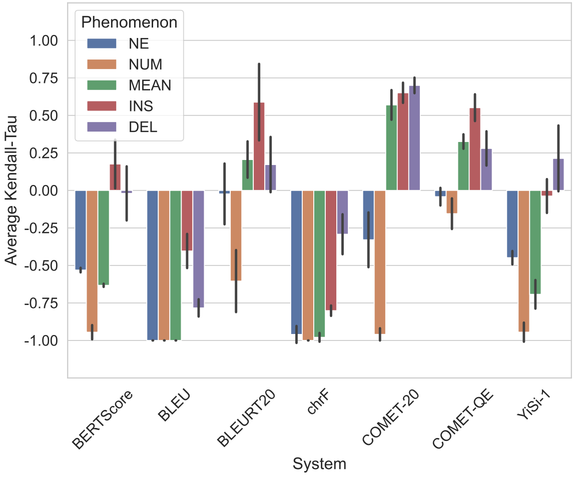

Figure 4: Average Kendall-Tau for baseline metrics discriminated by phenomenon. The coloured bars indicate the average score for all language pairs and the black bars represent the standard deviation.

5.3 Submitted Metrics

The submissions that rely on the reference to predict a score encompass COMET-22 (Rei et al., 2022), metricx_xl_DA_2019[3], MS-COMET-22 (Kocmi et al., 2022) and UniTE (Wan et al., 2022).

[3] Citation was not available.

As depicted in Figure 5, these metrics obtain much higher correlations, when compared to the baselines. The metric metricx_xl_DA_2019 obtains the overall best results, achieving high correlations for all phenomena. Across all metrics, the Deviation in Numbers phenomenon is the one with lowest scores. Furthermore, it is also the one with the highest standard deviation over the several language pairs, showing the uncertainty of these metrics when faced with this perturbation.

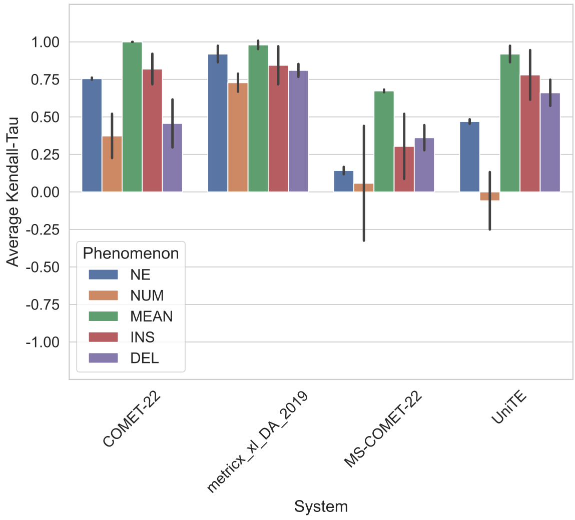

Figure 5: Average Kendall-Tau for submitted reference based metrics discriminated by phenomenon. The coloured bars indicate the average score for all language pairs and the black bars represent the standard deviation.

Regarding reference free metrics, submissions comprise of COMET-Kiwi (Rei et al., 2022), HWTSC-Teacher-Sim (Liu et al., 2022), HWTSC-TLM (Liu et al., 2022), KG-BERTScore (Liu et al., 2022) and MS-COMET-QE-22 (Kocmi et al., 2022). Here, it is important to note that HWTSC-TLM is a system that only receives the hypothesis as input and, as such, it is likely in disadvantage in this task, as the developed bad hypothesis are only critical errors in the context of the source sentence.

As shown in Figure 6, several reference free metrics obtain very high correlations for all linguistic phenomena. The main exception is HWTSC-TLM, which can be attributed to the reasons explained above. KG-BERTScore obtains the best overall results, with almost perfect correlations. Furthermore, we observe that reference free metrics outperform reference based metrics. This result is further discussed in the following section.

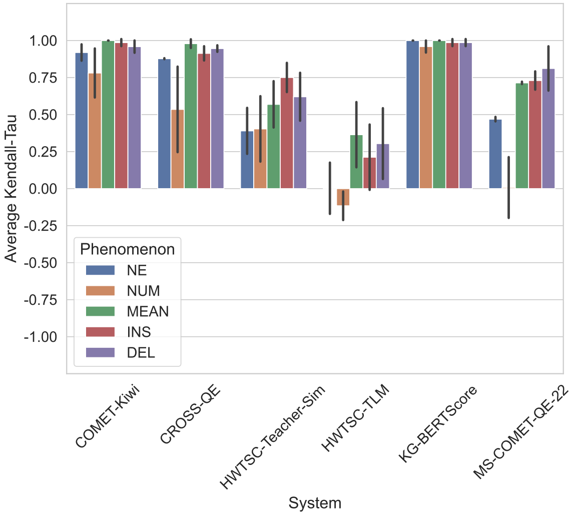

Figure 6: Average Kendall-Tau for submitted reference free metrics discriminated by phenomenon. The coloured bars indicate the average score for all language pairs and the black bars represent the standard deviation.

5.4 Reference based vs Reference free

Figure 7 compares the performance of reference based and reference free metrics across all phenomena. We observe that reference free metrics obtain higher correlations on all perturbations, which can be attributed to the adversarial nature of the bad hypothesis that is specifically generated with a localized perturbation of the reference.

This result reveals the dependency of reference based metrics on the reference and, in particular, on the word overlap of the reference with the hypothesis. Reference-free metrics are forced to attend to the source and compare its meaning with the hypothesis, as there is little word overlap between the two sentences. This issue is particularly visible in the Deviation in Named Entities and Numbers phenomena, where the reference and bad hypothesis differ on a single named entity or number, respectively.

Comparing the performance of metrics for each phenomenon, we verify that both groups of metrics obtain lower correlations for Deviation in Named Entities and Numbers, indicating these phenomena are not well detected by current methods. Moreover, the results show large standard deviations, suggesting an inherent unpredictability on the performance of current methods for all phenomena.

5.5 Penalisation of critical errors

In order to measure whether the metrics penalize the critical errors when they score the bad hypoth-

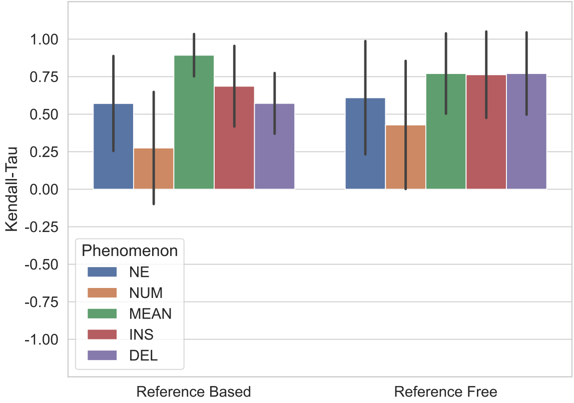

Figure 7: Average Kendall-Tau for submitted reference based and reference free metrics discriminated by phenomenon. The coloured bars indicate the average score for all language pairs and the black bars represent the standard deviation.

esis lower, we compare their Kendall-Tau values with their average difference between the scores for good and bad hypothesis, as described in Section 5.1.

In Figure 8, we observe that submitted metrics not only obtain higher correlations but also have a greater difference between the scores attributed to the good and bad hypothesis. Moreover, the two variables follow a linear relationship, obtaining a Pearson Correlation Coefficient of 0.8924. This shows the metrics that correctly distinguish the good from the bad hypothesis also penalize the bad hypothesis accordingly.

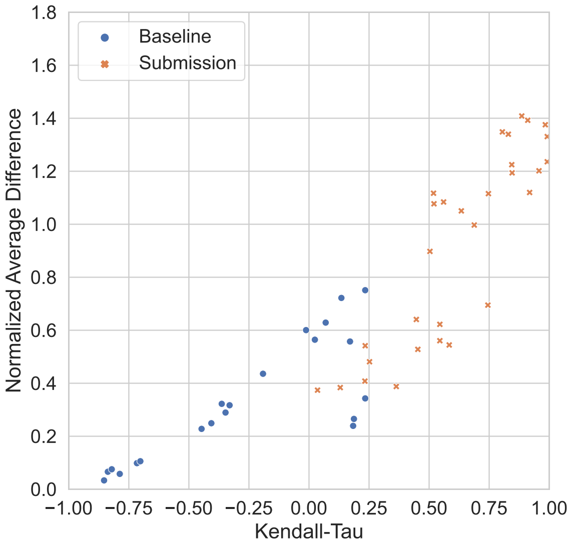

Figure 8: Average Kendall-Tau and Difference for all metrics. Each data point represents a single metric and language pair.

6 Conclusions

Ensuring generated translations do not have critical errors is a crucial aspect of Machine Translation Evaluation, as they can pose various risks. In this work, we propose SMAUG, a multilingual augmentation framework to create translations with critical errors by inducing several linguistic phenomena in existing translations. We also apply these perturbations to create a manually verified test set to assess the robustness of Machine Translation Evaluation systems to critical errors.

With the created test set, we evaluate multiple metrics and show promising progress in current State-of-the-Art methods in both distinguishing translations with and without critical errors and significantly penalizing the occurrence of critical errors in translations. Nevertheless, errors related to named entities and numbers were found to pose a challenge for several tested metrics. Additionally, we observe a high variance in the measured correlations across all the developed phenomena, suggesting an unpredictability on the performance of current methods with respect to detecting critical errors.

One of the challenges in the automatic generation of translations with critical errors is the validation of the output. In this work, we relied on a preliminary automatic validation but also required a manual verification of the outputs. Future work will explore high-precision validation techniques, such as the work of Raunak et al. (2022) that uses very specific detectors to find examples of critical errors in translations.

Furthermore, support for multiple languages is a crucial aspect of this framework. However, several of the devised perturbations support a limited number of languages pairs. For example, the Deviation in Meaning phenomenon only supports to-English language pairs, as the POLYJUICE model is an English only model. A future avenue of research will investigate methods to expanding the number of languages supported by the linguistic phenomena.

Acknowledgements

This work was supported by the P2020 programs (MAIA, contract 045909), by EU's Horizon Europe Research and Innovation Actions (UTTER, contract 101070631), by the European Research Council (ERC StG DeepSPIN 758969), and by the Fundação para a Ciência e Tecnologia through contract UIDB/50008/2020.

References

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*, Online. Association for Computational Linguistics.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André F. T. Martins. 2022. Results of WMT22 Metrics Shared Task: Stop Using BLEU – Neural Metrics Are Better and More Robust. In *Proceedings of the Seventh Conference on Machine Translation*, Abu Dhabi. 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.

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.

Diptesh Kanojia, Marina Fomicheva, Tharindu Ranasinghe, Frédéric Blain, Constantin Orăsan, and Lucia Specia. 2021. Pushing the right buttons: Adversarial evaluation of quality estimation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 625–638, Online. Association for Computational Linguistics.

Tom Kocmi, Hitokazu Matsushita, and Christian Federmann. 2022. MS-COMET: Larger Filtered Human Annotations Help Metric Performance. In *Proceedings of the Seventh Conference on Machine Translation*, Abu Dhabi. Association for Computational Linguistics.

Yilun Liu, Xiaosong Qiao, Zhanglin Wu, Su Chang, Min Zhang, Yanqing Zhao, Song Peng, shimin tao, Hao Yang, Ying Qin, Jiaxin Guo, Minghan Wang, Yinglu Li, Peng Li, and Xiaofeng Zhao. 2022. Partial could be better than whole. HW-TSC 2022 Submission for the Metrics Shared Task. In *Proceedings of the Seventh Conference on Machine Translation*, Abu Dhabi. 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.

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

Arle Lommel, Aljoscha Burchardt, and Hans Uszkoreit. 2014. Multidimensional Quality Metrics (MQM): A Framework for Declaring and Describing Translation Quality Metrics. *Tradumàtica: tecnologies de la traducció*, 0:455–463.

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.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.

Gema Ramírez-Sánchez, Jaume Zaragoza-Bernabeu, Marta Bañón, and Sergio Ortiz Rojas. 2020. Bifixer and bicleaner: two open-source tools to clean your parallel data. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 291–298, Lisboa, Portugal. European Association for Machine Translation.

Vikas Raunak, Matt Post, and Arul Menezes. 2022. SALTED: A Framework for SAlient Long-Tail Translation Error Detection.

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*, 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.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. In

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 4902–4912, Online. Association for Computational Linguistics.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. 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.

Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas.

Lucia Specia, Frédéric Blain, Marina Fomicheva, Chrysoula Zerva, Zhenhao Li, Vishrav Chaudhary, and André F. T. Martins. 2021. Findings of the WMT 2021 shared task on quality estimation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 684–725, Online. Association for Computational Linguistics.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

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.

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.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Tongshuang Wu, Marco Tulio Ribeiro, Jeffrey Heer, and Daniel Weld. 2021. Polyjuice: Generating counterfactuals for explaining, evaluating, and improving models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6707–6723, Online. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. In *8th International Conference on Learning Representations (ICLR)*, Addis Ababa, Ethiopia. OpenReview.net.