A Benchmark for Evaluating Machine Translation Metrics on Dialects Without Standard Orthography

Noëmi Aepli[1] **Chantal Amrhein**[1,2] **Florian Schottmann**[2,3] **Rico Sennrich**[1,4]

[1]University of Zurich, [2]Textshuttle, [3]ETH Zurich, [4]University of Edinburgh

{naepli,sennrich}@cl.uzh.ch, {amrhein,schottmann}@textshuttle.com

Abstract

For sensible progress in natural language processing, it is important that we are aware of the limitations of the evaluation metrics we use. In this work, we evaluate how robust metrics are to non-standardized dialects, i.e. spelling differences in language varieties that do not have a standard orthography. To investigate this, we collect a dataset of human translations and human judgments for automatic machine translations from English to two Swiss German dialects. We further create a challenge set for dialect variation and benchmark existing metrics' performances. Our results show that existing metrics cannot reliably evaluate Swiss German text generation outputs, especially on segment level. We propose initial design adaptations that increase robustness in the face of non-standardized dialects, although there remains much room for further improvement. The dataset, code, and models are available here: https://github.com/textshuttle/dialect_eval

1 Introduction

As multilingual NLP models include more and more languages, the community's focus on low-resource languages has also grown. This not only includes languages for which we have "little data" but also language varieties and dialects which often pose additional challenges, especially if they do not have a standardized orthography. Recent work has shown some progress in classification tasks (e.g. Wang et al., 2021; Touileb and Barnes, 2021; Aepli and Sennrich, 2022) as well as generation tasks where such language varieties appear on the input side only (e.g. Zbib et al., 2012; Honnet et al., 2018; Alam et al., 2023). For these scenarios, we can use established evaluation schemes. However, for research towards NLP models *generating* language varieties, Sun et al. (2023) have shown that current evaluation metrics are not robust to translations into different dialects.

GSW	... ufere	Webs ii te	aa glueg e t	w ä rd e .
GSW	... ufere	Webs i te	ah gluegt	w e rd ä .
de	... auf einer	Webseite	angeschaut	werden.
en		... *viewed on a website.*		

Figure 1: Example sentence that shows the extent of spelling variability in language varieties, here Swiss German dialect (GSW), with German (de) and English (en) translations.

What their evaluation does not consider is that language varieties often lack a standardized orthography and do not adhere to consistent spelling rules. This implies that even *within* a single dialect, notable orthographic variations can be observed, as illustrated in the Swiss German example in Figure 1. The same utterance with a similar but different spelling would result in a high word error rate of $\frac{3}{4}$.

Many languages have multiple regional variants, such as Spanish (Mexican, Argentinean, etc.), French (Canadian, Belgian, etc.), or English (British, American, Australian, Indian, etc.), among others. Such language varieties exhibit various lexical, grammatical, and orthographical distinctions. Importantly, these differences are *standardized*, meaning that they adhere to specific spelling rules and conventions, albeit with variations specific to each variant. This suggests that if a neural metric is exposed to a sufficient amount of data encompassing various language varieties, it should be able to develop similar representations and provide comparable scores for a given sentence in different varieties. Sun et al. (2023) show that pre-training a metric on data from multiple dialects indeed makes metrics more inter-dialect robust.

However, for a substantial number of languages and language varieties, there exists no established standard orthography. Many regions exhibit a dialect continuum where language varieties lack precise boundaries, and each dialect displays a significant range of diversity within itself. Furthermore, when speakers write in their dialect, they follow

Proceedings of the Eighth Conference on Machine Translation (WMT), pages 1045–1065
December 6–7, 2023.

their individual writing styles. Such kinds of variabilities, as can be observed in the example in Figure 1, are much less consistent and localized and will differ significantly between different writers. A metric designed to handle these kinds of varieties must be capable of addressing frequent spelling differences, which is considerably more challenging to learn solely from data compared to the standardized language variation differences mentioned in the previous paragraph.

In recent years, embedding-based metrics have gained increasing popularity (Sellam et al., 2020; Rei et al., 2020a) which – in theory – could be more appropriate for assessing non-standardized language varieties than string-based MT metrics like BLEU (Papineni et al., 2002) or chrF (Popović, 2015). However, these neural metrics are often not trained on the language varieties in question. Additionally, recent work showed that reference-based learned metrics still rely too much on subword overlap with the reference (Hanna and Bojar, 2021; Amrhein et al., 2022).

In this work, we follow Sun et al. (2023) and analyze the dialect robustness of machine translation metrics but specifically focus on non-standardized language varieties that were not seen during pretraining. Our contributions are:

- We collect a new dataset and design a challenge set for evaluating MT metrics on two Swiss German dialects.
- We benchmark existing string-based and neural metrics on our dataset and find that they are not reliable, especially on segment level.
- We propose initial adaptations to make metrics more robust for Swiss German but find that there is still a lot of room for improvement.

2 Related Work

There is a substantial amount of research on MT *into* language varieties (Scherrer, 2011b; Haddow et al., 2013; Fancellu et al., 2014; Hassani, 2017; Costa-jussà et al., 2018; Lakew et al., 2018; Myint Oo et al., 2019; Wan et al., 2020; Garcia and Firat, 2022). Most of these works exclusively evaluate with surface-level metrics like BLEU (Papineni et al., 2002) but some voice their concerns over a lack of reliable evaluation metrics (Kumar et al., 2021; Bapna et al., 2022).

Sun et al. (2023) confirm that existing machine translation evaluation metrics are not dialect-robust. They show that it is possible to train more robust metrics by including a language and dialect identification task in a second language model pre-training phase. While they focus on inter-dialect robustness between well-defined dialects, i.e. Brazilian and Iberian Portuguese, our study focuses on a setting where dialects lack standardized orthography. This absence of standardization introduces additional variability, resulting in distinct challenges and necessitating different solutions for MT systems, which need to generalize to often limited data; MT metrics, which need to be robust to spelling differences; and also meta-evaluation, which has its own challenges when collecting human assessments for dialects without standardized orthography as we outline in Section 3.1. To investigate how reliable MT metrics are for non-standardized varieties, we collect a new dataset with human translations and human judgments for MT outputs from English to two Swiss German dialects.

While other works also evaluate MT metrics on language varieties and dialects, Sun et al. (2023) is closest to our work: Alam et al. (2023) only look at language varieties on the source side and Riley et al. (2023) only evaluate language varieties for which a standard was included in the language model pretraining. Both studies also conclude that existing metrics are not robust to dialects. Riley et al. (2023) further propose a new automated lexical accuracy metric based on term dictionaries, similar to metrics used for automatic speech recognition (ASR) (Ali et al., 2017; Nigmatulina et al., 2020) which allow for more flexible string matching by using a look-up table of acceptable spellings. Riley et al.'s approach may work well if there is a limited set of term differences between dialects. However, such a metric is difficult to employ for language varieties without standardized spelling rules. Instead, we experiment with increasing dialect robustness by introducing character-level noise during metric training which has been shown to be useful for cross-lingual transfer to language varieties without standardized orthography (Aepli and Sennrich, 2022; Srivastava and Chiang, 2023; Blaschke et al., 2023).

3 Evaluation Data for Swiss German Dialects

While we focus on Swiss German because there are enough different MT systems that can be eval-

uated, Swiss German is by no means the only language where its varieties do not have standardized spelling. Many medium to high-resource languages like Arabic (Darwish et al., 2021) or Italian (Ramponi, 2022) include dialectal varieties that lack a standardized orthography. Additionally, this phenomenon extends to numerous low-resource settings (Bird, 2022), encompassing a wide array of language varieties across Africa (Adebara and Abdul-Mageed, 2022), Asia (Roark et al., 2020; Aji et al., 2022), Oceania (Solano et al., 2018) and the Americas (Littell et al., 2018; Mager et al., 2018). Historically, even many language varieties that now have a standardized orthography did not always have one, including English (Scragg, 1974). This makes our work on robust metrics for non-standardized dialects also relevant for NLP for historical texts.

To measure robustness against non-standardized dialects, we design two new datasets. With the first, we investigate how metrics behave in a realistic setup where we compare them against human judgments. The second is a challenge set that allows us to investigate score changes between different spellings and compare them to score changes when meaning is changed. This is inspired by similar experiments in Sun et al. (2023).

3.1 Human Judgement Data

In order to realistically evaluate machine translation metrics on Swiss German dialects, it is essential to obtain human-translated reference segments and human judgments for machine-translated translation hypotheses. Since no such data exists for Swiss German, we compile our dataset based on the English NTREX-128 data[1] (Federmann et al., 2022). We selected this dataset because it originates from a standard test set[2], already contains human translations into 128 languages including some regional variants, has a permissive license[3] and offers document context which is important for collecting reliable human judgments (Läubli et al., 2018; Toral et al., 2018).

Human reference translations: For the reference translations, we provided two Swiss German translators with the English NTREX-128 source data (i.e. 1997 sentences from 123 documents). Translators saw sentences in document context and were asked to translate them into their respective native dialects (i.e. Bern and Zurich region). We provided translators with simple instructions where we stated that they must not post-edit machine translation outputs to translate the texts.

Human judgment scores: The hypotheses come from ten machine translation systems translating from English to Bern dialect and ten systems translating from English to Zurich dialect. For each dialect, we include nine neural MT systems in our rating setup and one rule-based system.

The neural models are provided by Textshuttle. They are based on a standard Transformer architecture (Vaswani et al., 2017) trained using different amounts of data, making use of data augmentation techniques like backtranslation (Sennrich et al., 2016). Some of the systems use German as a pivot language. In collaboration with Textshuttle, we decided to evaluate models for which they expect noticeable translation differences and not to compare the nine models that they think would perform the best. The rule-based system works by morphosyntactically analyzing the standard German NTREX-128 translation of the English source and then sequentially applying a set of dialect-specific rewriting rules to generate Swiss German output. The system is described in detail in Scherrer (2011a). The system version used for this task operates word by word without taking syntax into account. Notably, this means that past tense and genitive forms produce unpredictable output because they would require larger changes in the sentence structure.

We translated the English NTREX-128 source data with each neural system and the German NTREX-128 translation with the rule-based systems and let native dialect speakers rate the outputs via Appraise[4] (Federmann, 2018), a framework for the evaluation of machine translation outputs. Raters only had access to the source for context because providing the reference could incentivize raters to "quickly compare the surface forms of translation against reference without understanding" (Freitag et al., 2022). Note that in order to mitigate dialect preference biases as documented by Riley et al. (2023) and Abu Farha and Magdy (2022), the translators and raters were all native speakers of the dialect they were asked to rate or translate into. We collected continuous Direct Assessment (DA) scores (Graham et al., 2013) where the slider

[1] https://github.com/MicrosoftTranslator/NTREX

[2] newstest2019 from the 2019 news translation shared task at WMT (Barrault et al., 2019)

[3] Attribution-ShareAlike 4.0 International (CC BY-SA 4.0)

[4] https://github.com/AppraiseDev/Appraise

presented to the raters was annotated with Scalar Quality Metric (SQM) labels which increases the rating stability across annotators (Kocmi et al., 2022). Raters viewed segments in a document context and rated translations on the segment level as well as the document level. The document-level ratings are collected to enable future research on document-level metrics; in this study, we only focus on segment-level ratings.

Ideally, we would recruit professional translators for both the translation and the rating tasks. However, there exist no professional translators for Swiss German. Instead, we recruited translators and annotators from a pool of reliable candidates who already worked on similar Swiss German projects. To ensure the quality of the ratings we collect, we included control segments as implemented in Appraise. Based on this control, no raters needed to be excluded.

As Swiss German constitutes a dialect continuum, its various variations lack precise boundaries, and each dialect displays a significant range of diversity within itself. Consequently, during the recruitment process, we placed our trust in the annotators' self-identification of their native dialects. Furthermore, it is worth noting that all our contributors, comprising six women and five men, belong to younger generations, with raters ranging in age from 23 to 30, and translators aged 35 to 40, respectively. This age factor has an impact on their dialect. All translators and annotators were paid 30 CHF per hour for their work.

3.2 Challenge Set

As an additional evaluation, we compile a challenge set to directly pinpoint how robust metrics are to dialect variability. In the creation of this challenge set, we draw inspiration from the work of Sun et al. (2023), who propose measuring inter-dialect robustness by comparing metric scores between two language varieties and between one variety and a version with significant meaning changes. If segment pairs of the latter type are judged more or equally similar by a metric than those of the two varieties, Sun et al. (2023) argue the metric is not dialect-robust.

We build our challenge set from the collected data presented in the previous section. We filter for all MT hypotheses that humans rated as perfect (i.e. received a score of 100). If more than one unique hypothesis exists for a segment, we create all combinations of these hypotheses. For example, if four different machine translation outputs for the same source all receive a perfect human rating, this results in six pairs of semantically equivalent translation hypotheses that feature orthographic differences. For each pair, we then manually create a modified version of one of the hypotheses to change its meaning. Following Sun et al. (2023), we consider deletion, insertion, and substitution operations for introducing meaning changes which we randomly assign to each hypothesis pair. All changes are made either to a single word or if necessary a whole phrase. This process results in hypothesis triples as seen in this example:

A: S e chs Mitarbeiter s i wäg e Verletzige behandlet worde.
B: S ä chs Mitarbeiter s y wäg Verletzige behandlet worde.
Six members of staff have been treated for injuries.

C: Sechs Mitarbeiter si wäge Verletzige **beschtraft** worde.
*Six members of staff **were punished because of** injuries.*

Hypotheses A and B are semantically equivalent but exhibit spelling differences. Hypothesis C is very similar to hypothesis A on the surface level but differs significantly in meaning. During evaluation, metrics will have access to one of these hypotheses, as well as the reference and/or the source (depending on whether it is a reference-free or reference-based metric). We describe how we compare the different scores for these hypotheses in Section 4.3.

4 Experiment Setup

4.1 Benchmarking Existing Metrics

To document the performance of current MT metrics on dialects without a standard orthography, we evaluate the following metrics:

- **BLEU**[5] (Papineni et al., 2002), a string-based metric with a brevity penalty that calculates the word-level n-gram precision between a translation and one or multiple references.
- **chrF++**[6] (Popović, 2017), another string-based metric that provides a character n-gram, word unigram, and bigram F-score by computing overlaps between the hypothesis and reference translation.

[5]computed with SacreBLEU (Post, 2018), signature: nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.3.0.
[6]computed with SacreBLEU (Post, 2018), signature: nrefs:1|case:mixed|eff:yes|nc:6|nw:0|space:no|version:2.3.0.

We expect surface-level, string-based metrics to perform badly on dialects without standard spelling rules as they are entirely based on overlap with a reference translation. These are also the metrics used by most works that explored text generation for language varieties without standardized orthography (e.g. Jeblee et al., 2014; Meftouh et al., 2015; Kumar et al., 2021). We further benchmark the following neural metrics:

- **COMET-20**[7] (Rei et al., 2020b) and **COMET-22**[8] (Rei et al., 2022), two reference-based neural metrics built on the COMET framework (Rei et al., 2020a). These are trained neural metrics that are built on top of a large, pre-trained language model and are fine-tuned on human judgment data from previous metric evaluation campaigns. COMET-20 is fine-tuned to predict DA scores. COMET-22 is an ensemble between a COMET-20-like model and a multi-task model that predicts segment-level Multidimensional Quality Metric (MQM) scores (Uszkoreit and Lommel, 2013) as well as word-level error tags.
- **COMET-20-QE**[9] (Rei et al., 2020b) and **COMET-Kiwi**[10] (Rei et al., 2022), two reference-free neural metrics for quality estimation. COMET-20-QE is trained similarly to COMET-20 and COMET-KIWI to COMET-22, but both versions do not have access to the reference during training on human judgments.

While these metrics go beyond surface-level comparisons to the reference due to their hidden representations and embedding-based nature, we expect that they still struggle to reliably evaluate translations into Swiss German for several reasons: First, no Swiss German data was included for pre-training the language model (XLM-R; Conneau et al., 2019) that is used as the basis for training COMET. Second, neural metrics are often fine-tuned on Standard German data which shares many similar words with Swiss German and could falsely bias metrics towards Standard German spelling. Third, reference-based metrics have been shown to still be influenced by surface overlap with the reference (Hanna and Bojar, 2021; Amrhein et al., 2022) which is a disadvantage in situations where numerous spelling variations exist.

[7] wmt20-comet-da
[8] wmt22-comet-da
[9] wmt20-comet-qe-da
[10] wmt22-cometkiwi-da

4.2 Developing Dialect-Robust Metrics

Similar to Sun et al. (2023), we also experiment with training more robust metrics but we focus on robustness against non-standardized dialects rather than inter-dialect robustness. The following list summarizes our metrics:

- **COMET-REF** and **COMET-QE**, a baseline trained as a reference to compare our modifications to because our COMET models differ slightly from COMET-20 and COMET-22 (see details below).
- **+gsw**, same as the baseline but the pre-trained model is fine-tuned on Swiss German data before the COMET models are fine-tuned on human judgment data. This is similar to the second pre-training phase for the inter-dialect-robust metric proposed in Sun et al. (2023). However, we do not include the additional language and dialect identification task during continued pre-training as we do not have dialect labels for the Swiss German pre-training data.
- **+noise**, same as the baseline but during the fine-tuning process on human judgment data we introduce character-level noise. This is inspired by previous work that showed that this method allows for better cross-lingual transfer to closely related languages (Aepli and Sennrich, 2022; Srivastava and Chiang, 2023). Blaschke et al. (2023) hypothesize that injecting noise into standard language data results in a similar tokenization rate as for unseen dialects. We apply noise injection to all languages within the COMET fine-tuning dataset that have an alphabetic writing system, therefore excluding languages like Chinese which were not considered in the original work introducing character-level noise. Following Aepli and Sennrich (2022), we inject character-level noise (essentially typos) into a random selection of 15% of the tokens within each sentence. Specifically, we alter, delete, or add one character per chosen token. We execute this process using the characters specific to the relevant language, taking into account all characters that occur more than 1,000 times in the respective dataset. We apply this noise

injection to all segments, including the source, translation, and reference segments.

We provide details of how we trained those models here:

Continued pre-training of XLM-R To expose our models to Swiss German data, we modify the encoder model upon which COMET models are usually based: XLM-RoBERTa[11] (Conneau et al., 2019). We continue the training of the XLM-R model on SwissCrawl[12] (Linder et al., 2020), a corpus containing 500K dialect sentences crawled from the web in late 2019. For the continued pre-training, we work with the Huggingface Transformers library[13] (Wolf et al., 2020), following the default configurations for language model fine-tuning which involves a training duration of three epochs.

Training COMET models We train COMET models using the official code base[14] with the default settings from version 2.0.2. We use the "regression model" configuration for the reference-based models and the "referenceless model" configuration for the reference-free models. Our models are trained on the direct assessment data collected by the organizers of the WMT news translation task spanning the years 2017 to 2021 (2021 as dev set)[15] (Bojar et al., 2017, 2018; Barrault et al., 2019, 2020; Akhbardeh et al., 2021). It is important to highlight that our models are not directly comparable to the original WMT shared task COMET models, for which the 2020 models were exclusively trained on data from 2017-2019 and the 2022 models used a different configuration.

4.3 Evaluation

We evaluate our metrics in five different ways. For the human judgment data, we compute two scores on system (sys) and two on segment (seg) level using the reference implementation from the WMT metrics shared task[16] (Freitag et al., 2022), except for *success rate* where we use our own implementation.

[11] xlm-roberta-base
[12] swisscrawl
[13] https://github.com/huggingface/transformers
[14] https://github.com/Unbabel/COMET
[15] https://github.com/Unbabel/COMET/tree/master/data
[16] https://github.com/google-research/mt-metrics-eval

System level The *pairwise accuracy* as defined by Kocmi et al. (2021), measures the accuracy with which a metric agrees with human preference between pairs of systems where the human ratings are significantly different according to a two-sided Wilcoxon test. Note that the score difference between the two systems is not important in this analysis. Furthermore, we provide results for the *sys-level Pearson correlation*, quantifying the strength of the linear relationship between metrics and human judgment scores for systems.

Segment level At the segment level, our evaluation includes the *seg-level accuracy* with an optimized tie threshold, which resembles a global accuracy but also acknowledges metrics for correctly predicting tied human judgment scores (Deutsch et al., 2023). Further, we present the *seg-level Kendall correlation*, akin to pairwise accuracy but employing a distinct normalization technique.

Challenge set For the challenge set, we compute the *success rate* (seg level) following Sun et al. (2023). This measures the accuracy with which a metric assigns more similar scores (s) to two equivalent translations A and B compared to a version with a semantic change C. Consequently, a metric is considered robust to non-standardized dialects for a segment if the score difference between s_A and s_B is smaller than the score difference between s_C and either s_A or s_B (depending on which score is smaller):

$$|s_A - s_B| < min(s_A, s_B) - s_C \quad (1)$$

5 Results

Table 1 provides a comprehensive summary of our results with scores for existing metrics (top), COMET models trained for this work (bottom), system-level evaluations (left), and segment-level evaluations (right). Additional results can be found in the appendices. Appendix A.1 contains results related to the incorporation of additional languages in the pre-training process, Appendix B presents an evaluation of performance on an official WMT benchmark, and Appendix C presents pairwise accuracy plots for our metrics.

Existing vs. GSW metrics As expected, the surface-level metrics perform worse than trained metrics in almost all evaluations. Our baseline metrics often perform a bit worse than the existing COMET metrics, this is particularly true for

	system-level			segment-level					
	pairwise accuracy	**Pearson correlation**		**tie-optim. accuracy**		**Kendall correlation**		**success rate**	
		BE	ZH	BE	ZH	BE	ZH	BE	ZH
BLEU	0.740	0.728	0.587	0.544	0.560	0.142	0.163	0.135	0.194
chrF	0.753	0.806	0.665	0.486	0.478	0.076	0.079	0.121	0.145
COMET-20	0.766	0.849	0.816	0.565	0.583	0.205	0.227	0.250	0.298
COMET-22	0.766	0.897	0.901	0.570	0.587	0.184	0.212	0.243	0.306
COMET-20-QE	0.675	0.875	0.872	0.508	0.516	0.134	0.134	0.131	0.161
COMET-KIWI	0.636	0.952	0.876	0.536	0.533	0.146	0.142	0.240	0.290
COMET-REF	0.740	0.864	0.793	0.567	0.570	0.180	0.194	0.221	0.234
+ gsw	0.792	**0.906**	0.862	0.611	0.627	**0.286**	**0.317**	0.320	0.347
+ noise	0.727	**0.940**	**0.903**	0.561	0.567	**0.223**	**0.233**	0.237	0.290
+ gsw + noise	0.792	**0.917**	**0.868**	0.597	0.621	**0.271**	**0.304**	0.287	0.323
COMET-QE-KIWI	0.636	0.781	0.689	0.486	0.507	0.104	0.099	0.127	0.145
+ gsw	0.844	**0.978**	**0.987**	0.595	0.587	**0.257**	**0.283**	0.292	0.298
+ noise	0.675	**0.915**	**0.817**	0.524	0.528	**0.154**	**0.158**	0.149	0.177
+ gsw + noise	**0.896**	**0.968**	**0.981**	0.582	0.596	**0.246**	**0.269**	0.273	0.274

Table 1: Results for the baselines metrics (above) and our trained metrics (below) on system level (left) and segment level (right). Darker shades indicate lower scores. Bold denotes statistically significant improvement compared to their respective baselines COMET-REF or COMET-QE-KIWI. There is no information about significance for tie-optim. accuracy (columns 4-5) and success rate (columns 8-9). Note that BE and ZH represent the abbreviations for the two Swiss German (GSW) dialect regions under consideration.

our reference-free model. However, continued pre-training on Swiss German data improves their performance considerably and they strongly outperform existing metrics. This highlights the importance of the model to have seen the target language (variety) during the language model pre-training. It also shows that metrics can be extended to include new languages and language varieties with limited effort although this impacts their performance on other language pairs as we show in Appendix B. Continued pre-training on multiple languages and language varieties can mitigate this effect (see Appendix A.1).

Noise injection While continued LM pre-training on Swiss German data generally outperforms noise injection during task fine-tuning, we still see gains over the baselines. This suggests that metrics that were trained on noised data are more robust to unseen language (varieties) and may be a good strategy for language (varieties) without sufficient data for continued pre-training. Combining both continued pre-training and noise injection generally does not lead to further improvements.

Reference-based vs reference-free While both types of metrics perform similarly with continued pre-training on Swiss German, both existing reference-free metrics perform worse than the existing reference-based metrics in the segment-level evaluations. Since these metrics did not see any Swiss German during the pre-training phase, having access to the reference as an anchor might help the reference-based metrics for unseen languages. Amrhein et al. (2022) reported a similar finding where the reference acted as an anchor when metrics were used to identify copied source sentences.

Challenge set The success rate for all metrics is extremely low. Metrics assign more similar scores to a hypothesis with a semantic change than to a different translation hypothesis in the majority of cases. Again, continued pre-training on Swiss German results in the best metric performance. However, even these scores are lower than a random success rate of 50% by far. Our findings highlight that even though system-level correlations may seem convincing, none of the metrics studied in this work are robust to non-standardized dialect variations.

Since our results show that there is still significant room for improvement toward metric robustness to non-standardized language varieties, we provide suggestions for future work.

6 Open Questions

We hope that our benchmark inspires more work on robust evaluation metrics for language varieties in the future. In this section, we list several directions we think are worthwhile exploring:

Expanding the benchmark: We were not able to include additional language varieties in our benchmark at the time because we could not find enough *different* machine translation systems that translate into these varieties. While we recognize that without reliable metrics this is a "chicken-and-egg" problem, we still advocate for more MT research that focuses on translating *into* language varieties. Expanding our benchmark would not only allow us to draw more general conclusions but would also help with sample size for the *pairwise accuracy* analysis (Kocmi et al., 2021) since we find that a large number of systems are required for confident results.

More focus on segment level: Segment-level metric scores tend to be much less correlated with human judgments when contrasted with system-level correlations (Freitag et al., 2022) and have also been shown to be unreliable in downstream tasks (Moghe et al., 2023). We hope that future work aimed at enhancing metric performance on our challenge set will also contribute to greater metric reliability on segment level in general, as over-reliance on reference overlap is also a problem for languages with standardized spelling (Hanna and Bojar, 2021; Amrhein et al., 2022).

Training neural metrics that model character-level similarities: A segment in a dialect often resembles a reference in certain characters only rather than in full words (see Figure 1 as an example). As the underlying language models of neural evaluation metrics use a fixed tokenization scheme that was learned on text that likely does not include many examples of language varieties, these similarities might be hard to account for by the neural metric. Thus, we believe that character-based language models, such as Canine (Clark et al., 2022), could provide a better basis for neural evaluation metrics to model character-level similarities.

7 Conclusion

We evaluated the reliability of machine translation metrics when evaluating dialects without standard orthographies. As part of this work, we collected a new dataset consisting of human translations, human judgments, and a challenge set from English to two Swiss German dialects. We benchmark several existing metrics and find that they are not robust to variation featured by non-standardized dialects. Based on this finding, we explore several modifications that allow us to train metrics that are more robust towards spelling variation. Our results show that there is still a lot of room for improvement and we offer a set of recommendations for future work on dialect robust metrics.

Limitations

The goal of this work is to evaluate and develop machine translation metrics that take into account the spelling variability of dialects and languages without established writing norms. We recognize that evaluating metrics on varieties from different languages would help generalize our results. However, we were not able to find enough differing machine translation systems that translate *into* the same language variety for other languages. Therefore, we had to limit this study to two Swiss German dialects. We hope to include further language varieties in our benchmark in the future (when such machine translation systems become available) to encourage research toward metrics that are reliable for many non-standardized language varieties.

We did our best to avoid dialectal preference bias within our annotators by selecting only annotators who consider themselves native speakers of the respective dialect. However, as Swiss German is a dialect continuum, this can only be controlled to a certain degree.

Ethics Statement

This work includes the compilation of a new dataset as a test set for evaluating various machine translation metrics. All translators and annotators were compensated at a rate of 30 CHF per hour. Our dataset is based on a publicly available dataset and will be released under the same license for future use. **Intended use:** The dataset and the models resulting from this work are intended to be used by the research community to evaluate machine translation metrics.

Acknowledgements

We thank Yves Scherrer for providing the rule based systems and helpful comments. Furthermore, we thank Annette Rios, Tom Kocmi, Mathias Müller, and the anonymous reviewers for their valuable inputs. We are also grateful to the Swiss German raters and translators for their important contribution. This work was supported by the Swiss National Science Foundation (project nos. 191934 & 176727), Textshuttle, and the Department of Computational Linguistics at the University of Zurich.

References

Ibrahim Abu Farha and Walid Magdy. 2022. The effect of Arabic dialect familiarity on data annotation. In *Proceedings of the The Seventh Arabic Natural Language Processing Workshop (WANLP)*, pages 399–408, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Ife Adebara and Muhammad Abdul-Mageed. 2022. Towards afrocentric NLP for African languages: Where we are and where we can go. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3814–3841, Dublin, Ireland. Association for Computational Linguistics.

Noëmi Aepli and Rico Sennrich. 2022. Improving zero-shot cross-lingual transfer between closely related languages by injecting character-level noise. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 4074–4083, Dublin, Ireland. Association for Computational Linguistics.

Alham Fikri Aji, Genta Indra Winata, Fajri Koto, Samuel Cahyawijaya, Ade Romadhony, Rahmad Mahendra, Kemal Kurniawan, David Moeljadi, Radityo Eko Prasojo, Timothy Baldwin, Jey Han Lau, and Sebastian Ruder. 2022. One country, 700+ languages: NLP challenges for underrepresented languages and dialects in Indonesia. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7226–7249, Dublin, Ireland. Association for Computational Linguistics.

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.

Md Mahfuz Ibn Alam, Sina Ahmadi, and Antonios Anastasopoulos. 2023. Codet: A benchmark for contrastive dialectal evaluation of machine translation. *arXiv preprint arXiv:2305.17267*.

Ahmed Ali, Preslav Nakov, Peter Bell, and Steve Renals. 2017. Werd: Using social text spelling variants for evaluating dialectal speech recognition. In *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 141–148.

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

Ankur Bapna, Isaac Caswell, Julia Kreutzer, Orhan Firat, Daan van Esch, Aditya Siddhant, Mengmeng Niu, Pallavi Nikhil Baljekar, Xavier Garcia, Wolfgang Macherey, Theresa Breiner, Vera Saldinger Axelrod, Jason Riesa, Yuan Cao, Mia Chen, Klaus Macherey, Maxim Krikun, Pidong Wang, Alexander Gutkin, Apu Shah, Yanping Huang, Zhifeng Chen, Yonghui Wu, and Macduff Richard Hughes. 2022. Building machine translation systems for the next thousand languages. Technical report, Google Research.

Loïc Barrault, Magdalena Biesialska, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Matthias Huck, Eric Joanis, Tom Kocmi, Philipp Koehn, Chi-kiu Lo, Nikola Ljubešić, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Santanu Pal, Matt Post, and Marcos Zampieri. 2020. Findings of the 2020 conference on machine translation (WMT20). In *Proceedings of the Fifth Conference on Machine Translation*, pages 1–55, Online. Association for Computational Linguistics.

Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019. Findings of the 2019 conference on machine translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy. Association for Computational Linguistics.

Steven Bird. 2022. Local languages, third spaces, and other high-resource scenarios. In *Proceedings of the*

60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 7817–7829, Dublin, Ireland. Association for Computational Linguistics.

Verena Blaschke, Hinrich Schütze, and Barbara Plank. 2023. Does manipulating tokenization aid cross-lingual transfer? a study on POS tagging for non-standardized languages. In *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023)*, pages 40–54, Dubrovnik, Croatia. Association for Computational Linguistics.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Raphael Rubino, Lucia Specia, and Marco Turchi. 2017. Findings of the 2017 conference on machine translation (WMT17). In *Proceedings of the Second Conference on Machine Translation*, pages 169–214, Copenhagen, Denmark. Association for Computational Linguistics.

Ondřej Bojar, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, and Christof Monz. 2018. Findings of the 2018 conference on machine translation (WMT18). In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 272–303, Belgium, Brussels. Association for Computational Linguistics.

Jonathan H. Clark, Dan Garrette, Iulia Turc, and John Wieting. 2022. Canine: Pre-training an efficient tokenization-free encoder for language representation. *Transactions of the Association for Computational Linguistics*, 10:73–91.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Marta R. Costa-jussà, Marcos Zampieri, and Santanu Pal. 2018. A neural approach to language variety translation. In *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2018)*, pages 275–282, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Kareem Darwish, Nizar Habash, Mourad Abbas, Hend Al-Khalifa, Huseein T. Al-Natsheh, Houda Bouamor, Karim Bouzoubaa, Violetta Cavalli-Sforza, Samhaa R. El-Beltagy, Wassim El-Hajj, Mustafa Jarrar, and Hamdy Mubarak. 2021. A panoramic survey of natural language processing in the arab world. *Commun. ACM*, 64(4):72–81.

Daniel Deutsch, George Foster, and Markus Freitag. 2023. Ties matter: Modifying kendall's tau for modern metric meta-evaluation. *arXiv preprint arXiv:2305.14324*.

Federico Fancellu, Andy Way, and Morgan O'Brien. 2014. Standard language variety conversion for content localisation via SMT. In *Proceedings of the 17th Annual Conference of the European Association for Machine Translation*, pages 143–149, Dubrovnik, Croatia. European Association for Machine Translation.

Christian Federmann. 2018. Appraise evaluation framework for machine translation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 86–88, Santa Fe, New Mexico. Association for Computational Linguistics.

Christian Federmann, Tom Kocmi, and Ying Xin. 2022. NTREX-128 – news test references for MT evaluation of 128 languages. In *Proceedings of the First Workshop on Scaling Up Multilingual Evaluation*, pages 21–24, Online. Association for Computational Linguistics.

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

Xavier Garcia and Orhan Firat. 2022. Using natural language prompts for machine translation. *arXiv preprint arXiv:2202.11822*.

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.

Barry Haddow, Adolfo Hernández, Friedrich Neubarth, and Harald Trost. 2013. Corpus development for machine translation between standard and dialectal varieties. In *Proceedings of the Workshop on Adaptation of Language Resources and Tools for Closely Related Languages and Language Variants*, pages 7–14, Hissar, Bulgaria. INCOMA Ltd. Shoumen, BULGARIA.

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.

Hossein Hassani. 2017. Kurdish interdialect machine translation. In *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, pages 63–72, Valencia, Spain. Association for Computational Linguistics.

Pierre-Edouard Honnet, Andrei Popescu-Belis, Claudiu Musat, and Michael Baeriswyl. 2018. Machine translation of low-resource spoken dialects: Strategies for normalizing Swiss German. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Serena Jeblee, Weston Feely, Houda Bouamor, Alon Lavie, Nizar Habash, and Kemal Oflazer. 2014. Domain and dialect adaptation for machine translation into Egyptian Arabic. In *Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP)*, pages 196–206, Doha, Qatar. Association for Computational Linguistics.

Tom Kocmi, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Thamme Gowda, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Rebecca Knowles, Philipp Koehn, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Michal Novák, Martin Popel, and Maja Popović. 2022. Findings of the 2022 conference on machine translation (WMT22). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 1–45, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

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.

Sachin Kumar, Antonios Anastasopoulos, Shuly Wintner, and Yulia Tsvetkov. 2021. Machine translation into low-resource language varieties. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 110–121, Online. Association for Computational Linguistics.

Surafel Melaku Lakew, Aliia Erofeeva, and Marcello Federico. 2018. Neural machine translation into language varieties. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 156–164, Brussels, Belgium. Association for Computational Linguistics.

Samuel Läubli, Rico Sennrich, and Martin Volk. 2018. Has machine translation achieved human parity? a case for document-level evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4791–4796, Brussels, Belgium. Association for Computational Linguistics.

Lucy Linder, Michael Jungo, Jean Hennebert, Claudiu Cristian Musat, and Andreas Fischer. 2020. Automatic creation of text corpora for low-resource languages from the internet: The case of swiss german. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 2706–2711, Marseille, France. European Language Resources Association.

Patrick Littell, Anna Kazantseva, Roland Kuhn, Aidan Pine, Antti Arppe, Christopher Cox, and Marie-Odile Junker. 2018. Indigenous language technologies in Canada: Assessment, challenges, and successes. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2620–2632, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Manuel Mager, Ximena Gutierrez-Vasques, Gerardo Sierra, and Ivan Meza-Ruiz. 2018. Challenges of language technologies for the indigenous languages of the Americas. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 55–69, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Karima Meftouh, Salima Harrat, Salma Jamoussi, Mourad Abbas, and Kamel Smaili. 2015. Machine translation experiments on PADIC: A parallel Arabic DIalect corpus. In *Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation*, pages 26–34, Shanghai, China.

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

Thazin Myint Oo, Ye Kyaw Thu, and Khin Mar Soe. 2019. Neural machine translation between Myanmar (Burmese) and Rakhine (Arakanese). In *Proceedings of the Sixth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 80–88, Ann Arbor, Michigan. Association for Computational Linguistics.

Iuliia Nigmatulina, Tannon Kew, and Tanja Samardzic. 2020. ASR for non-standardised languages with dialectal variation: the case of Swiss German. In *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 15–24, Barcelona, Spain (Online). International Committee on Computational Linguistics (ICCL).

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.

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.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Alan Ramponi. 2022. Nlp for language varieties of italy: Challenges and the path forward. *arXiv preprint arXiv:2209.09757*.

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

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020a. 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.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020b. Unbabel's participation in the WMT20 metrics shared task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 911–920, Online. Association for Computational Linguistics.

Parker Riley, Timothy Dozat, Jan A. Botha, Xavier Garcia, Dan Garrette, Jason Riesa, Orhan Firat, and Noah Constant. 2023. FRMT: A benchmark for few-shot region-aware machine translation. *Transactions of the Association for Computational Linguistics*, 11:671–685.

Brian Roark, Lawrence Wolf-Sonkin, Christo Kirov, Sabrina J. Mielke, Cibu Johny, Isin Demirsahin, and Keith Hall. 2020. Processing South Asian languages written in the Latin script: the Dakshina dataset. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2413–2423, Marseille, France. European Language Resources Association.

Yves Scherrer. 2011a. Morphology generation for swiss german dialects. In *Systems and Frameworks for Computational Morphology*, pages 130–140, Berlin, Heidelberg. Springer Berlin Heidelberg.

Yves Scherrer. 2011b. Syntactic transformations for Swiss German dialects. In *Proceedings of the First Workshop on Algorithms and Resources for Modelling of Dialects and Language Varieties*, pages 30–38, Edinburgh, Scotland. Association for Computational Linguistics.

D. Scragg. 1974. *A History of English Spelling*. Manchester University Press, United Kingdom.

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.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Rolando Coto Solano, Sally Akevai Nicholas, and Samantha Wray. 2018. Development of natural language processing tools for Cook Islands Māori. In *Proceedings of the Australasian Language Technology Association Workshop 2018*, pages 26–33, Dunedin, New Zealand.

Aarohi Srivastava and David Chiang. 2023. Fine-tuning BERT with character-level noise for zero-shot transfer to dialects and closely-related languages. In *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023)*, pages 152–162, Dubrovnik, Croatia. Association for Computational Linguistics.

Jiao Sun, Thibault Sellam, Elizabeth Clark, Tu Vu, Timothy Dozat, Dan Garrette, Aditya Siddhant, Jacob Eisenstein, and Sebastian Gehrmann. 2023. Dialect-robust evaluation of generated text. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6010–6028, Toronto, Canada. Association for Computational Linguistics.

Antonio Toral, Sheila Castilho, Ke Hu, and Andy Way. 2018. Attaining the unattainable? reassessing claims of human parity in neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 113–123, Brussels, Belgium. Association for Computational Linguistics.

Samia Touileb and Jeremy Barnes. 2021. The interplay between language similarity and script on a novel multi-layer Algerian dialect corpus. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3700–3712, Online. Association for Computational Linguistics.

Hans Uszkoreit and Arle Lommel. 2013. Multidimensional quality metrics: A new unified paradigm for human and machine translation quality assessment. *Localization World, London*, pages 12–14.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Yu Wan, Baosong Yang, Derek F. Wong, Lidia S. Chao, Haihua Du, and Ben C.H. Ao. 2020. Unsupervised neural dialect translation with commonality and diversity modeling. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):9130–9137.

Xinyi Wang, Yulia Tsvetkov, Sebastian Ruder, and Graham Neubig. 2021. Efficient test time adapter ensembling for low-resource language varieties. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 730–737, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2020. CCNet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4003–4012, Marseille, France. European Language Resources Association.

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.

Rabih Zbib, Erika Malchiodi, Jacob Devlin, David Stallard, Spyros Matsoukas, Richard Schwartz, John Makhoul, Omar F. Zaidan, and Chris Callison-Burch. 2012. Machine translation of Arabic dialects. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 49–59, Montréal, Canada. Association for Computational Linguistics.

A Appendix

A.1 Mixed Continued Pre-training

In our main experiments in Section 5, we evaluate continued language model pre-training only on Swiss German data. While this increases the performance on our benchmark, it remains unclear whether this leads to a "specialized" metric that does not perform well on other language pairs. We will evaluate this in the next section, but first, we introduce a set of contrastive models that are less specialized to Swiss German. Continued pre-training for contrastive models involves incorporating mixed data from five languages apart from Swiss German, namely: German (de), English (en), French (fr), Hindi (hi), and Chinese (zh). We train one metric based on XLM-R with continued pre-training only on these five languages ("5 langs"), and another one where we also add GSW to the training data ("6 langs"). For both settings, we also test character-level noise in the COMET fine-tuning step, as described in Section 4.2. The data for the five additional languages is sourced from the CC-100 corpus[17] (Wenzek et al., 2020), which is a reconstructed version of XLM-R's training dataset. Specifically, we utilize the first 100,000 sentences from the training data of each language.

Table 2 shows the results we obtained from incorporating mixed data into the continued LM pre-training. We see a similar effect as when continuing the pre-training only on GSW in the main results in Section 5. The performance of the metrics increases in all evaluations. Comparing these results to the metric where we only continued pre-training on Swiss German (+6 langs vs. +gsw), the results are comparable and often not significantly different. In the next section, we investigate how these metrics behave on other language pairs.

B Correlations on WMT Benchmarks

As discussed in the previous section, we evaluate the performance of our metrics on an official WMT benchmark to monitor their performance on language pairs that do not involve Swiss German. To do this, we reproduce the evaluations from the WMT 2022 metrics task (Freitag et al., 2022) for a subset of language pairs. We evaluate on the following five language pairs:

[17] https://data.statmt.org/cc-100/

	system-level			segment-level					
	pairwise accuracy	**Pearson correlation**		**tie-optim. accuracy**		**Kendall correlation**		**success rate**	
		BE	ZH	BE	ZH	BE	ZH	BE	ZH
COMET-REF	0.740	0.864	0.793	0.567	0.570	0.180	0.194	0.221	0.234
+ noise	0.727	**0.940**	**0.903**	0.561	0.567	**0.223**	**0.233**	0.237	0.290
+ gsw	0.792	**0.906**	0.862	0.611	0.627	**0.286**	**0.317**	0.320	0.347
+ gsw + noise	0.792	**0.917**	**0.868**	0.597	0.621	**0.271**	**0.304**	0.287	0.323
+ 5 langs	0.766	0.877	0.774	0.561	0.583	**0.212**	**0.230**	0.235	0.274
+ 5 langs + noise	0.766	**0.938**	**0.890**	0.570	0.593	**0.241**	**0.256**	0.265	0.290
+ 6 langs	0.805	**0.932**	**0.887**	0.592	0.616	**0.286**	**0.316**	0.357	0.452
+ 6 langs + noise	0.779	**0.956**	**0.917**	0.599	0.622	**0.282**	**0.311**	0.323	0.379
COMET-QE-KIWI	0.636	0.781	0.689	0.486	0.507	0.104	0.099	0.127	0.145
+ noise	0.675	**0.915**	**0.817**	0.524	0.528	**0.154**	**0.158**	0.149	0.177
+ gsw	0.844	**0.978**	**0.987**	0.595	0.587	**0.257**	**0.283**	0.292	0.298
+ gsw + noise	**0.896**	**0.968**	**0.981**	0.582	0.596	**0.246**	**0.269**	0.273	0.274
+ 5 langs	0.610	0.758	0.773	0.514	0.505	**0.134**	**0.135**	0.164	0.202
+ 5 langs + noise	0.701	0.898	**0.831**	0.513	0.521	**0.178**	**0.184**	0.166	0.266
+ 6 langs	0.831	**0.985**	**0.984**	0.583	0.605	**0.261**	**0.284**	0.304	0.331
+ 6 langs + noise	0.870	**0.983**	**0.983**	0.579	0.591	**0.251**	**0.269**	0.284	0.323

Table 2: Results for systems with continued pre-training only on Swiss German (+ gsw), on 5 other languages (+ 5 langs) and the same languages including Swiss German (+ 6 langs). Darker shades indicate lower scores. Bold denotes statistically significant improvement compared to their respective baselines COMET-REF or COMET-QE-KIWI. There is no information about significance for tie-optim. accuracy (columns 4-5) and success rate (columns 8-9). Note that BE and ZH represent the abbreviations for the two Swiss German (GSW) dialect regions under consideration.

- **en-de:** evaluation against MQM ratings collected specifically for the metrics shared task.
- **en-zh:** evaluation against MQM ratings collected specifically for the metrics shared task.
- **de-en:** evaluation against reference-based DA scores collected for the translation shared task.
- **cs-uk:** evaluation against DA + SQM scores collected for the translation shared task.
- **en-liv:** evaluation against DA + SQM scores collected for the translation shared task.

Note that all these languages except for Livonian (liv) are part of the CC-100 corpus[18] (Wenzek et al., 2020). Consequently, they form a part of the training dataset for XLM-R and are thus included in the COMET models. Moreover, English (en), German (de), and Chinese (zh) were incorporated into the mixed continued pre-training, as explained in Section A.1. Lastly, all the languages mentioned above, with the exception of Ukrainian (uk) and Livonian (liv; a language of Latvia), are included in the COMET training data.

This evaluation allows us to assess the effects of our modifications both on language pairs that were included during COMET training, during continued LM pre-training, and those that were not.

The results are shown in the following Tables: 3 (system-level Pearson correlation), 4 (segment-level accuracy), and 5 (segment-level Kendall). We do not report pairwise accuracy here because they cannot be directly compared with the WMT22 results, given that we have only included a subset of the language pairs. Versions of COMET-ref that were continued pretrained on Swiss German data demonstrate comparable or improved performance compared to the baseline metrics. In contrast, continued pretrained COMET-qe performs worse. When examining individual languages, we observe that fine-tuning is advantageous for translations into Livonian (liv), which is the only language in our selection not included in XLMR. Conversely, for translations into English, continued pretrained systems, particularly COMET-qe, tend to perform slightly worse.

C Pairwise Comparison Plots

In the subsequent plots displayed in Figures 2 (existing metrics), 3 (our trained COMET-ref metrics), and 4 (our trained COMET-qe metrics), every point represents a difference in average human judgment (y-axis) and a difference in automatic metric (x-axis) over a pair of systems. Metrics disagree with human ranking for system pairs in pink quadrants. These plots follow the example of Figure 1 in (Kocmi et al., 2021).

[18] https://data.statmt.org/cc-100/

sys-level Pearson correlation	de-en	en-de	en-zh	en-liv	cs-uk
BLEU	0.353	0.178	0.065	-0.575	0.890
chrF++	0.356	0.304	0.203	-0.517	0.925
COMET-20	0.424	0.876	0.744	0.893	0.985
COMET-22	0.450	0.873	0.756	-0.517	0.989
COMET-20-QE	0.443	0.577	0.752	0.564	0.953
COMET-KIWI	0.421	0.748	0.767	-0.563	0.987
COMET-ref	0.423	0.888	0.626	0.909	0.992
+ noise	0.420	**0.931**	0.618	0.912	0.991
+ gsw	0.410	0.904	0.450	0.693	0.983
+ gsw + noise	0.407	0.930	0.375	0.610	0.964
+ 5 langs	0.412	0.897	0.656	0.826	0.993
+ 5 langs + noise	0.415	**0.933**	0.658	0.689	0.991
+ 6 langs	0.417	0.908	0.636	0.892	0.992
+ 6 langs + noise	0.413	**0.951**	0.626	0.627	0.989
COMET-qe	0.384	0.453	0.639	0.598	0.954
+ noise	0.398	0.464	0.659	0.589	0.961
+ gsw	0.365	0.300	0.444	0.806	0.874
+ gsw + noise	0.387	0.354	0.446	0.859	0.893
+ 5 langs	0.371	0.434	0.650	0.621	0.923
+ 5 langs + noise	0.377	0.429	0.667	0.639	0.939
+ 6 langs	0.372	0.424	0.657	0.694	0.921
+ 6 langs + noise	0.380	0.440	0.640	0.725	0.939

Table 3: System-level Pearson correlation scores for baseline metrics (above) and our trained metrics (below) on a subset of language pairs from the WMT 2022 metrics task. Bold denotes statistically significant improvement compared to their respective baselines COMET-REF or COMET-QE-KIWI, underlined denotes statistically significant decline.

seg-level tie-optim. **accuracy**	de-en	en-de	en-zh	en-liv	cs-uk
BLEU	0.394	0.539	0.096	0.319	0.490
chrF++	0.391	0.545	0.352	0.237	0.466
COMET-20	0.439	0.580	0.466	0.589	0.563
COMET-22	0.437	0.584	0.468	0.368	0.567
COMET-20-QE	0.442	0.566	0.460	0.513	0.556
COMET-KIWI	0.412	0.580	0.470	0.338	0.567
COMET-ref	0.439	0.565	0.462	0.540	0.556
+ noise	0.434	0.556	0.458	0.615	0.564
+ gsw	0.434	0.551	0.470	0.507	0.542
+ gsw + noise	0.432	0.543	0.470	0.453	0.531
+ 5 langs	0.444	0.570	0.471	0.593	0.543
+ 5 langs + noise	0.428	0.553	0.478	0.500	0.552
+ 6 langs	0.445	0.567	0.475	0.461	0.551
+ 6 langs + noise	0.430	0.560	0.483	0.523	0.545
COMET-qe	0.433	0.550	0.470	0.545	0.555
+ noise	0.436	0.561	0.470	0.520	0.544
+ gsw	0.445	0.546	0.472	0.583	0.518
+ gsw + noise	0.441	0.552	0.463	0.505	0.500
+ 5 langs	0.445	0.561	0.467	0.522	0.530
+ 5 langs + noise	0.439	0.550	0.462	0.526	0.520
+ 6 langs	0.443	0.555	0.480	0.520	0.528
+ 6 langs + noise	0.453	0.552	0.470	0.517	0.526

Table 4: Segment-level accuracy scores (the darker the lower) for baseline metrics (above) and our trained metrics (below) on a subset of language pairs from the WMT 2022 metrics task. There is no information about significance.

seg-level Kendall correlation	de-en	en-de	en-zh	en-liv	cs-uk
BLEU	0.009	0.169	0.032	-0.158	0.133
chrF++	0.007	0.146	0.056	-0.158	0.086
COMET-20	0.018	0.319	0.141	0.208	0.280
COMET-22	0.019	0.343	0.137	-0.111	0.295
COMET-20-QE	0.022	0.234	0.123	0.126	0.254
COMET-KIWI	0.016	0.231	0.123	-0.147	0.281
COMET-ref	0.015	0.320	0.139	0.213	0.267
+ noise	0.019	0.310	0.125	0.165	0.251
+ gsw	0.016	0.293	0.120	0.096	0.225
+ gsw + noise	0.020	0.298	0.101	0.059	0.213
+ 5 langs	0.017	0.316	0.131	0.127	0.252
+ 5 langs + noise	0.018	0.321	0.128	0.095	0.238
+ 6 langs	0.017	0.309	0.133	0.140	0.246
+ 6 langs + noise	**0.018**	0.309	0.126	0.070	0.234
COMET-qe	0.017	0.225	0.121	0.152	0.235
+ noise	0.014	**0.228**	0.114	0.137	0.217
+ gsw	0.013	0.178	0.093	0.146	0.161
+ gsw + noise	0.013	0.182	0.094	0.102	0.162
+ 5 langs	0.020	0.214	0.115	0.145	0.214
+ 5 langs + noise	0.019	0.217	0.117	0.142	0.203
+ 6 langs	0.015	0.216	0.117	0.167	0.198
+ 6 langs + noise	0.016	0.212	0.114	0.147	0.186

Table 5: Segment-level Kendall correlation scores for baseline metrics (above) and our trained metrics (below) on a subset of language pairs from the WMT 2022 metrics task. Bold denotes statistically significant improvement compared to their respective baselines COMET-REF or COMET-QE-KIWI, underlined denotes statistically significant decline.

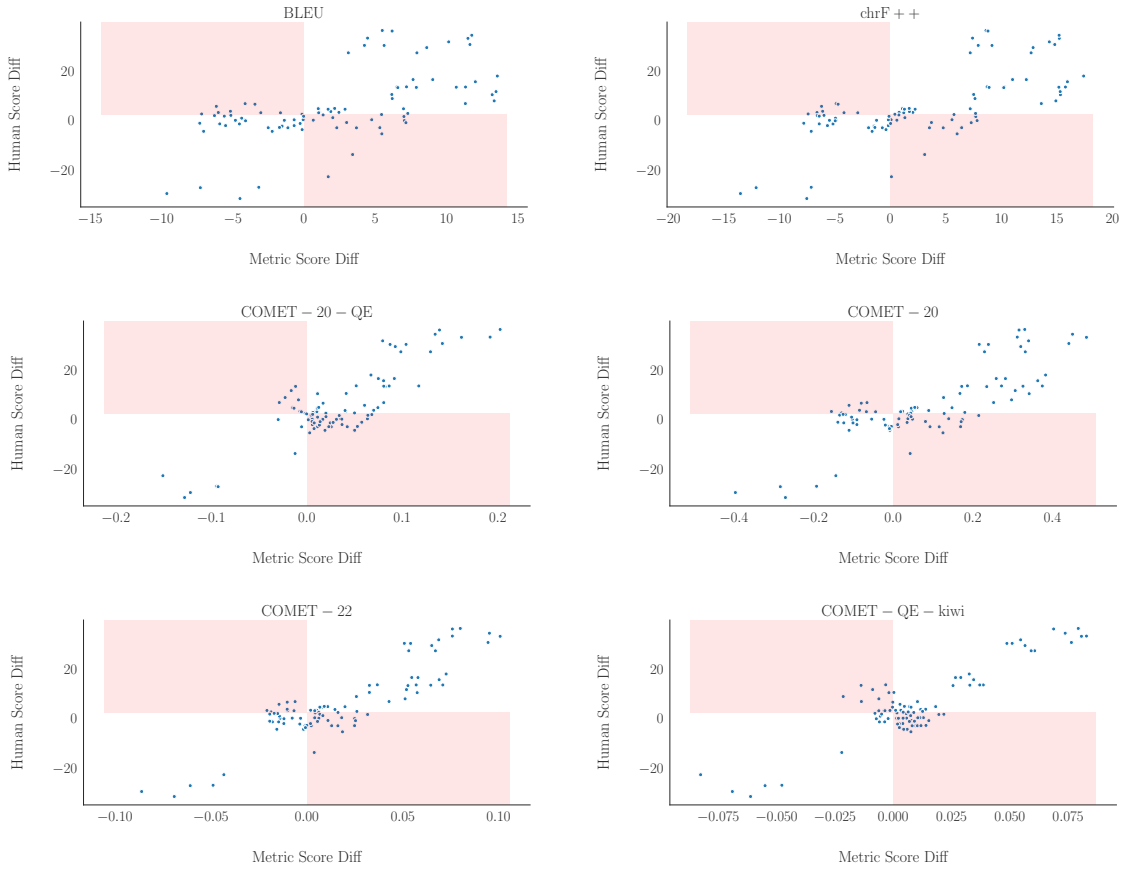

Figure 2: Pairwise comparison plots for existing metrics.

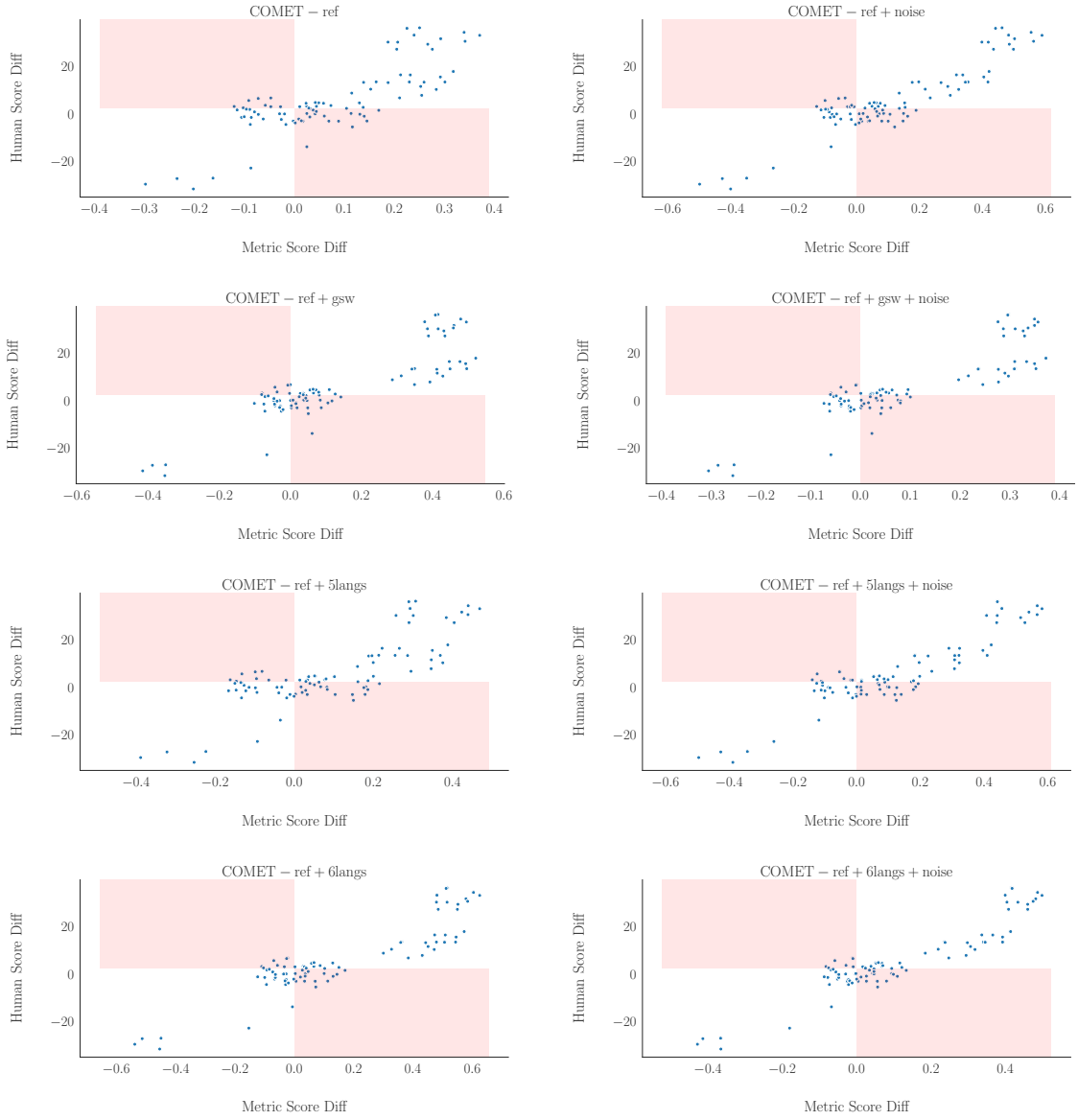

Figure 3: Pairwise comparison plots for the COMET-ref metrics trained for this work.

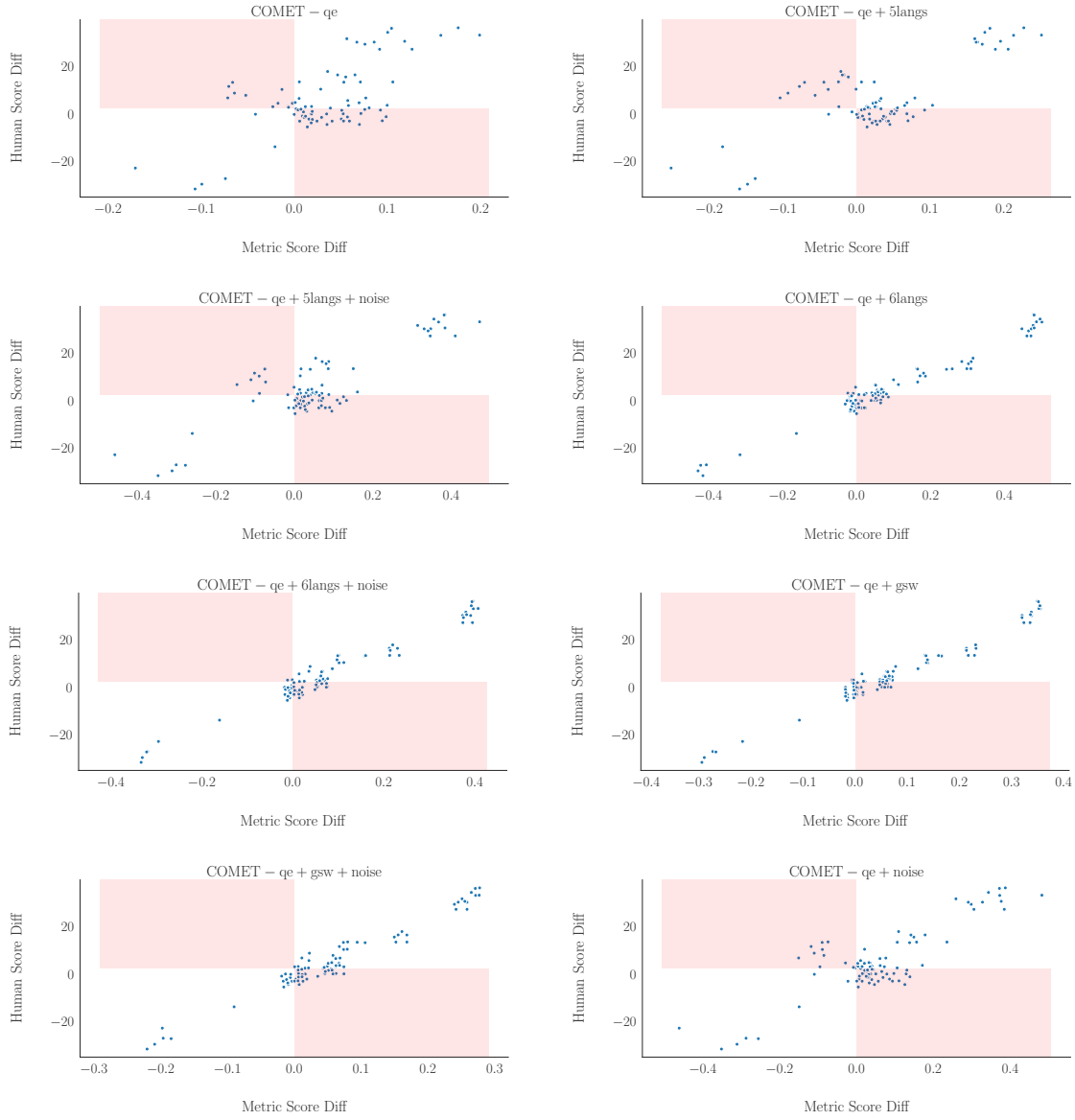

Figure 4: Pairwise comparison plots for the COMET-qe metrics trained for this work.