Dialect-robust Evaluation of Generated Text

Jiao Sun[1,2*] **Thibault Sellam**[1] **Elizabeth Clark**[1] **Tu Vu**[1,3*] **Timothy Dozat**[1]
Dan Garrette[1] **Aditya Siddhant**[1] **Jacob Eisenstein**[1] **Sebastian Gehrmann**[1]
[1]Google Deepmind
[2]University of Southern California [3]University of Massachusetts Amherst
nano-dialect-eval@google.com

Abstract

Text generation metrics that are not robust to dialect variation make it impossible to tell how well systems perform for many groups of users, and can even penalize systems for producing text in lower-resource dialects. In this paper, we introduce a suite of methods to assess whether metrics are *dialect robust*. These methods show that state-of-the-art metrics are not dialect robust: they often prioritize dialect similarity over semantics, preferring outputs that are semantically incorrect over outputs that match the semantics of the reference but contain dialect differences. As a step towards dialect-robust metrics for text generation, we propose NANO, which introduces regional and language information to the metric's pretraining. NANO significantly improves dialect robustness while preserving the correlation between automated metrics and human ratings. It also enables a more ambitious approach to evaluation, *dialect awareness*, in which system outputs are scored by both semantic match to the reference and appropriateness in any specified dialect.

1 Introduction

Most natural language generation (NLG) evaluation metrics compare a system output against a human-written reference. References are usually drawn from a relatively narrow range of linguistic styles. They often exclude varieties like Indian English or Iberian Portuguese, which are *geographical dialects* with millions of speakers. As a result, outputs in dialects that are not represented in the reference may score poorly, discouraging the development of systems to meet the needs of these language communities. Although contemporary metrics such as COMET (Rei et al., 2020) can be reference-free, they still rely on training data and rater pools that do not cover all dialects of interest, leading to a high number of out-of-domain dialects. The performance of evaluation metrics on these out-of-domain dialects has not been quantified.

We define a *dialect-robust* evaluation metric as one that produces the same score for system outputs that share the same semantics, but are expressed in different dialects. To understand whether current evaluation metrics are dialect-robust, we propose to quantify the dialect robustness at the dialect feature-level and sentence-level. The analyses measure the dialect-sensitivity of evaluation metrics by comparing semantics-preserving dialect edits to perturbations that change the meaning of sentences.

Through our analyses, we demonstrate that multiple state-of-the-art NLG evaluation metrics are not robust to dialects of Mandarin, English, and Portuguese. In many cases, system outputs that are perturbed so as to differ semantically from the reference score higher than outputs in which the only change is to the dialect. With the goal of increasing the dialect robustness and without performance degradation on standard benchmarks, we propose a training schema NANO. NANO is an unsupervised pretraining step to a metric that distills dialect information of the multilingual pretraining dataset into a model, which we demonstrate leads to improved dialect robustness.

Based on our findings, we lay out research goals toward dialect-inclusive metrics. Moving beyond dialect robustness, we formalize the goal of *dialect awareness*, in which metrics can be applied to any user-specified language and dialect regardless of the language of the reference or source document.

2 Dialect Robustness and Awareness

Dialects can be regarded as linguistic subdivisions that align with communities of speakers, often grouped by geographical or demographic attributes (Chambers et al., 1998). A classic example is nation-level varieties, such as Brazilian and Iberian Portuguese. Dialects are distinguished from each other by a set of *dialect features*, which can op-

[*] Work done while Jiao and Tu interning at Google.

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 6010–6028
July 9-14, 2023 ©2023 Association for Computational Linguistics

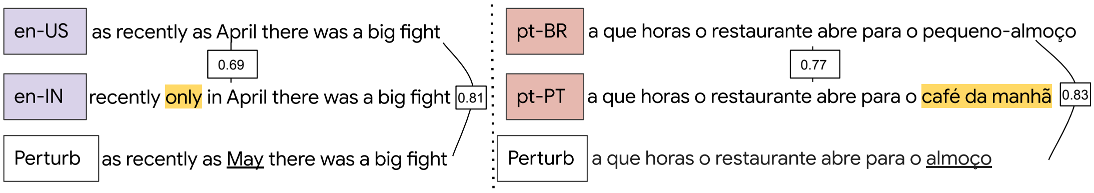

en-US
as recently as April there was a big fight
0.69
en-IN
recently only in April there was a big fight
0.81
Perturb
as recently as May there was a big fight
pt-BR
a que horas o restaurante abre para o pequeno-almoço
0.77
pt-PT
a que horas o restaurante abre para o café da manhã
0.83
Perturb
a que horas o restaurante abre para o almoço

Figure 1: An illustration of dialect robustness in the context of generation evaluation. We define *dialect robustness* as evaluation metrics that are expected to have the same output across dialects that share the same semantics. Dialect edits (highlighted in yellow) should not lead to a greater degradation of score than edits that change the underlying semantics (highlighted in underline). BLEURT-20 in the figure assigns higher score to semantically-perturbed sentences than sentences with dialect features, exposing its vulnerability to dialects.

erate at the levels of pronunciation, lexicon, rhetorical devices, and grammar (Whiteman, 2013); one working definition of dialect is as a set of correlated features (Nerbonne, 2009).

Two examples of dialect features are shown in Figure 1. The left side shows the English dialect feature "focus *only*", which distinguishes Indian English from other varieties, such as US English (Lange, 2012). The feature changes the surface form but not the underlying semantics. The right panel of Figure 1 shows the Portuguese dialect feature of different lexical choice for the same semantics ("breakfast"), which distinguishes Iberian Portuguese from Brazilian Portuguese. Many dialect features are acceptable in multiple dialects: for example, zero definite article ("∅ main reason is . . . ")[1] is used in Indian English, Singapore English, and several other post-colonial dialects.

Dialect Robustness Consider a translation system that produces Iberian Portuguese outputs. If all the training data for the metric used to evaluate generation quality comes from Brazilian Portuguese, it will likely assign a lower score to Iberian Portuguese outputs, thereby misrepresenting system quality and disincentivizing further development of the more diverse system in favor of one that only produces Brazilian Portuguese. To formalize this intuition, we define dialect robustness in the context of NLG evaluation as:

Definition 2.1 (Dialect robustness)**.** Let $y^{(d)}$ and $y^{(d')}$ be two system outputs that are semantically equivalent but written in different dialects. An evaluation metric $m : \mathcal{Y} \to \mathbb{R}$ is **dialect robust** iff $m(y^{(d)}) = m(y^{(d')})$ for all such $(y^{(d)}, y^{(d')})$.[2]

This definition is strict: it would not apply to any system that produced even small differences in score between semantically equivalent, regionally distinct outputs. For that reason, we propose a relaxed criterion, which compares the change in the metric induced by dialect to changes induced by semantic perturbations:

Definition 2.2 (ϕ-Dialect robustness)**.** Let $y^{(d)}$ and $y^{(d')}$ be two semantically-equivalent system outputs that differ in dialect. Let $\phi : \mathcal{Y} \to \mathcal{Y}^*$ be a semantic perturbation function that maps an input to a set of outputs whose semantics are different from the input. An evaluation metric $m : \mathcal{Y} \to \mathbb{R}$ is ϕ-dialect robust if $m(y^{(d)}, y^{(d')}) > m(y^{(d)}, \tilde{y})$ for all semantically-equivalent $(y^{(d)}, y^{(d')})$ and all $\tilde{y} \in \phi(y^{(d)})$.

Dialect Awareness Consider a translation system that is supposed to translate into Brazilian Portuguese but instead produces Iberian Portuguese. In this case, a dialect-robust metric is undesirable because it is unable to detect this mistake. To account for these cases, we define dialect awareness:

Definition 2.3 (Dialect-awareness)**.** Let $\mathcal{T}$ be a set of dialect tags. A metric $m : \mathcal{Y} \times \mathcal{T} \to \mathbb{R}$ is **dialect aware** iff $m(y^{(d)}, d) \geq m(y^{(d')}, d)$ for all semantically-equivalent input pairs $(y^{(d)}, y^{(d')})$ where $y^{(d)}$ is in dialect $d \in \mathcal{T}$ and $y^{(d')}$ is in dialect $d' \neq d$.

Informally, a metric is dialect aware if, given a dialect identifier and a pair of semantically-equivalent texts that vary by dialect, it assigns the highest score to the text in the dialect specified by the identifier. Dialect awareness is undefined with respect to inputs that are not semantically equivalent. This means that the definition is agnostic as

[1] https://ewave-atlas.org/parameters/62#2/7.0/7.9

[2] For simplicity we do not include the reference in this definition. A corpus-level reference-based metric could be defined as $\frac{1}{N}\sum_i m_i(y_i)$ with $m_i(y_i) = \delta(y_i, r_i)$, with r_i indicating the reference for example i and $\delta : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$. Similarly, a corpus-level quality estimation metric could be defined with $m_i(y_i) = \delta(y_i, x_i)$ with x_i indicating the input, such as the source language or passage to be summarized. For the corpus-level metric to be dialect robust (or ϕ-robust), all m_i must be dialect robust (or ϕ-robust).

to whether the metric should prioritize matching the target semantics or the target dialect.

Figure 1 illustrates the concepts of dialect robustness and dialect awareness. The top two rows of each panel vary only by dialect; the bottom row shows semantic perturbations of the top row. ϕ-dialect robustness implies that the top row is scored as more similar to the middle row than to the bottom row. Dialect awareness implies that the quality of the surface form in each row should be highest when paired with the correct dialect label.

Is Semantic Equivalence Realistic? The above definitions presume that it is possible to characterize utterances in different dialects as semantically equivalent. Such characterizations have been criticized as lacking a strong foundation for semantic equivalence, outside the limited case in which the dialect differences are purely phonological (Lavandera, 1978; Romaine, 1981). One such criticism is that a pair of utterances might be semantically equivalent for some communicative purposes, but not for others. To avoid the gray area between dialect differences that change semantics and those that do not, we design perturbations that have a small surface-level impact on the original utterance but a strong effect on its meaning, e.g. by negating the main proposition or changing an important semantic argument. This establishes a necessary condition for dialect robustness: if a metric scores such perturbations more highly than dialect pairs, then it is certainly not dialect robust. Proving that a metric *is* dialect robust is more difficult, because it requires constructing more subtle semantic perturbations that are harder to distinguish (even conceptually) from dialect variables. Furthermore, from a practical standpoint we cannot evaluate $y^{(d)}$ with respect to *all* semantic perturbations $\tilde{y} \in \phi(y^{(d)})$, but the existence of perturbations for which $m(y^{(d)}, \tilde{y}) > m(y^{(d)}, y^{(d')})$ is enough to disprove dialect robustness.

3 Existing Metrics

To assess the quality of a generated text, most automatic evaluation approaches compare it to a "ground truth" reference, with higher similarity to the reference implying higher-quality output (Celikyilmaz et al., 2020). Similarity can be based on lexical features, as in BLEU (Papineni et al., 2002) and CHRF (Popović, 2015), or distributed representations, as in BLEURT (Sellam et al., 2020),[3] PRISM (Rei et al., 2020) and YISI (Lo, 2019). When distributed representations are used, they may be unsupervised (Zhang et al., 2020) or fine-tuned on a corpus of human ratings. In addition to these similarity-based metrics, there are also reference-free metrics for quality estimation (e.g., COMET-EQ; Rei et al., 2020), which we discuss in §5.2. Existing distributed metrics either use the multilingual representation from pretrained models, or create multilingual training data through various augmentation strategies. However, none of them explicitly accounts for dialectal variation.

4 Testing Dialect Robustness

In this section, we describe our methodology for assessing dialect robustness. We first introduce two ways to perturb sentences to get two comparable metrics' outputs and then describe the statistical tests we use to aggregate the outputs over a corpus.

4.1 Micro-level Dialect Features

Dialect features are local edits that distinguish dialects while avoiding changes to the meaning of the text, as described in §2. Our first robustness assessment uses such features. We start with a base sentence $y_i^{\text{(base)}}$ taken from a corpus of sentences D. We further assume access to a version of the same sentence in which a dialect feature was introduced, denoted $y_i^{\text{(dialect)}}$. Following Definition 2.2, we introduce a semantic perturbation that changes $y_i^{\text{(base)}}$ to $y_i^{\text{(perturb)}}$. Again using English as an example, from the U.S. English base sentence "*as recently as April…*", we may produce the Indian English version "*recently only in April…*" (using the feature *focus-only*), and the semantic perturbation "*as recently as May…*".

Let $m(y_i, y_j)$ be a metric function that takes a candidate sentence y_i and a reference y_j as input, and produces a score σ. Given the above defined variations of y_i, we define the dialect and perturbed scores as

$$\sigma_{m,i}^{\text{(dialect)}} = m(y_i^{\text{(dialect)}}, y_i^{\text{(base)}}) \tag{1}$$

$$\sigma_{m,i}^{\text{(perturb)}} = m(y_i^{\text{(perturb)}}, y_i^{\text{(base)}}). \tag{2}$$

To satisfy Definition 2.2, $\sigma_{m,i}^{\text{(dialect)}}$ should score higher than $\sigma_{m,i}^{\text{(perturbation)}}$ across the sentences in

[3] In practice, we use the latest BLEURT-20 (Pu et al., 2021), following the authors' recommendation in https://github.com/google-research/bleurt.

the corpus. This implies as a necessary condition that $\mathbb{E}_{i\sim D}[\sigma^{(\text{dialect})}_{m,i}] > \mathbb{E}_{i\sim D}[\sigma^{(\text{perturb})}_{m,i}]$.

We consider three perturbation types: deletion, replacement and insertion. Each perturbation aims to change the sentence by only a single word or phrase, so as to induce a strong semantic change with a minimal impact to the surface form. Such perturbations are expected to yield challenging but clear comparisons against dialect variation. There are no standard techniques for introducing semantic perturbations, so we apply few-shot learning by prompting LaMDA (Cohen et al., 2022). For each perturbation type, we provide five exemplars (see Appendix A) and then prompt LaMDA for automatic semantic perturbation given a sentence $y_i^{(\text{en-base})}$. Some sentences are not amenable to all perturbations — for example, some are too short to support deletion — so we choose one perturbation per sentence, with the preference order of replacement, insertion and then deletion.

4.2 Sentence-level Dialect Rewrites

Micro-level dialect features require significant linguistic expertise to identify and have been defined for only a few languages. We thus introduce a less granular method that is based on parallel human translations. Given an English base sentence EN_i, we obtain human translations $y_i^{(j)}$ and $y_i^{(k)}$ in dialects j and k of the target language, e.g., Brazilian and Iberian Portuguese. We can again use the metric m to score the pair, $\sigma^{(\text{dialect})}_{m,i} = m(y_i^{(j)}, y_i^{(k)})$.

Because we have access to the English base sentence, we can use machine translation to generate a sentence in the target language $\text{EN}_i \underset{\text{MT}}{\Longrightarrow} \hat{y}_i^{(j^*)}$ which we can compare to, yielding $\sigma^{(\text{MT})}_{m,i} = m(y_i^{(j)}, \hat{y}_i^{(j^*)})$. Here, j^* indicates the locale that we believe is most strongly targeted by the machine translation system ("pt-BR" for Portuguese, "zh-CN" for Mandarin).

Finally, we construct target language perturbations by first perturbing the English source $\text{EN}_i \Rightarrow \tilde{\text{EN}}_i$ and then automatically translating the perturbed sentence $\tilde{\text{EN}}_i \Rightarrow \tilde{y}^{(j^*)}$, yielding $\sigma^{(\text{perturb})}_{m,i} = m(y_i^{(j)}, \tilde{y}_i^{(j^*)})$. The perturbations are produced by prompting LaMDA with the same exemplars as in §4.1.[4]

We expect $\mathbb{E}[\sigma_m^{(\text{MT})}] > \mathbb{E}[\sigma_m^{(\text{perturb})}]$, because both involve machine translation while the latter also involves perturbation to the source. If we have $\mathbb{E}[\sigma_m^{(\text{perturb})}] > \mathbb{E}[\sigma_m^{(\text{dialect})}]$ then metric m strongly disprefers dialect variants, even in favor of inputs that are different in meaning due to the perturbation of the English source.

4.3 Statistical Methods

As a necessary condition for dialect robustness, we test whether the expected scores for dialect rewrites exceed the expected scores for semantic perturbations. A challenge in correctly characterizing the uncertainty of these comparisons is that there is a substantial amount of variance over the original examples. We handle this with two styles of analysis:

Mixed-effect Regression For metric m, example i, and condition $j \in \{\text{perturb}, \text{dialect}, \text{MT}\}$, we model the metric $\sigma^{(j)}_{m,i}$ via a mixed-effects regression (Baayen, 2012; Speelman et al., 2018),

$$\sigma_i^{(j)} = \theta_i + \phi_j + \epsilon_{i,j}, \tag{3}$$

with the subscript m implicit in each term. The first term θ_i is a random intercept associated with example i, which helps to address the variance across examples; ϕ_j, the parameter of interest, is a fixed effect associated with the condition j; $\epsilon_{i,j}$ is a Gaussian error. Because all methods and conditions are applied to all examples, the predictors are uncorrelated. This makes it possible to interpret $\phi_{m,j}$ as an estimate of the expected change in the metric value corresponding to the application of metric m in condition j. By including the θ_i term, the regression is conceptually equivalent to a pairwise comparison, in the sense that the regression also benefits from the additional power obtained by controlling for per-example variation.

Win/loss Analysis and Binomial Test For a coarse-grained evaluation that is more easily comparable across metrics, we count how often each condition j receives a higher score than condition k in a pairwise comparison. When j represents dialect rewrites and k represents semantic perturbations, a high win rate indicates that the metric is more likely to be dialect robust. To measure statistical significance, we apply a one-tailed binomial test, which computes the likelihood of achieving at least n wins on T trials given a null hypothesis win probability $\frac{1}{2}$. In words, we test against the null hypothesis that for each example, a dialect rewrite

[4] While it is possible directly perturb the sentences in the target language, using the same English validated few-shot setup scales to more languages at the cost of a more English-centric perturbation style.

and a semantic perturbation are equally likely to get the higher score.

As discussed in the next section, we perform multiple comparisons per metric, across different conditions and different languages. To adjust the p-values for multiple comparisons, we apply the Bonferroni correction (Dror et al., 2017).

5 NANO

We hypothesize that explicitly encoding dialect information while pretraining a model will lead to improved downstream robustness. To test this hypothesis on learned metrics for text generation, we introduce NANO,[5] a model-agnostic pretraining schema with the goal of improving dialect robustness without performance degradation on downstream metric benchmarks.

5.1 Acceptability Pretraining

Given a pretrained model, we add a second pretraining phase to distill dialect information into the model. Specifically, we define the NANO-task as: given an expression y, determine whether it is from a text that has been identified as written in language ℓ and/or dialect region d (e.g., en-IN).

Data To construct a training corpus for NANO, we process mC4 (Xue et al., 2021). We split the corpus into sentences and use a Language Identification (LangID) model (Zhang et al., 2018) by Botha et al. (2017) to identify the language and locale information for the sentences.[6] Besides LangID output, mC4 provides the URL where a sentence originated from which we extract the region information as an indication of geographic dialect. For Portuguese and Mandarin, we filter an instance if the predicted locale does not agree with the region information from the URL. For other languages, we combine the LangID and region information as a noisy approximation for a dialect of the language in the specific region. For example, if the LangID model predicts that the language is English and the region in the URL indicates India (.in), we treat the instance as en-IN.[7] In total, we include ten languages with metric finetuning data evaluated during the WMT benchmark with ninety-five language variants following the classification by van Esch et al. (2022).[8]

Given a sentence, we balance the ratio of sampling a dialect or language tag using a parameter λ. For instance, a sentence with gold dialect tag “pt-BR” can be a positive instance for the dialect itself or the general language “pt-any”. At the same time, it can also be a negative instance for other dialect (e.g., “en-IN”) or language (“en-any”). The ratio of positive instances versus negatives instances is always 0.5. For more discussion, see Appendix E.

Modeling We use mT5 (Xue et al., 2021) as our base model because the model is pretrained on the mC4 dataset, matching with our corpus choice and ensuring tokenizer compatibility. During pretraining, we transform each sentence into the string `candidate:` *{sentence}* `language:` *{language_tag}*, where the *language_tag* can be the dialect or the general language tag. The target label is zero or one, indicating whether the sentence belongs to the language tag. We adapt the encoder-decoder architecture of mT5 for regression by taking the logits of the first decoded token and applying the RMSE loss function between the logits and the label during model training. For more details about training, please see Appendix D.

5.2 Finetuning

Following Pu et al. (2021), we use the test data from the WMT 2015-2019 shared tasks as training data and use the WMT shared task 2020 as test data. There are three possible model specifications: (1) quantifying the semantic similarity between candidate and reference within the same reference, as in BLEURT (Pu et al., 2021) and YiSi (Lo, 2019); (2) measuring the similarity between candidate and a cross-lingual reference, as in COMET (Rei et al., 2020); (3) reference-free quality estimation, also performed by COMET. To compare to all models, we finetune on all three settings, using the input formats in Appendix subsection F.1.

6 Experiments

In this section, we demonstrate that existing metrics are not dialect robust by applying our proposed methods and statistical tests to existing corpora in English, Portuguese, and Mandarin (§6.1). We

[5] The name is motivated by the dialect feature “invariant tag (‘isn’t it’, ‘no’, ‘na’)” (Lange, 2012).

[6] We use a more current model that is capable of identifying the locale for Mandarin and Portuguese.

[7] This is an approximation because many dialects do not align with national borders. The development of a data-gathering approach for subnational and transnational dialects is an important topic for future work.

[8] Appendix B provides the full list of WMT language variants, which does not cover Portuguese. Our reported results on PT shows NANO’s capability in a zero-shot setting.

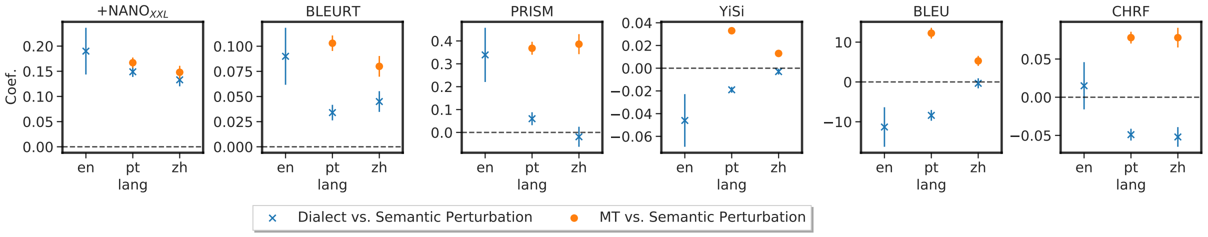


Figure 2: Differences in metric values for *Dialect vs. Semantic Perturbation* and *MT vs. Semantic Perturbation*, as estimated by the regression model. Higher values for *Dialect vs Semantic Perturbation* indicate more dialect robustness; negative values indicate that semantic perturbations are preferred over dialect differences. Error bars show 99% confidence intervals; they are larger for the English evaluations because there is less data. The *MT vs Semantic Perturbation* comparison is a stress test to measure whether the evaluation metrics can distinguish semantic differences from paraphrases.

show that language-aware pretraining improves the dialect robustness and leads to promising preliminary steps toward dialect-aware metrics (§6.4).

Datasets As described in §4, we consider micro-level and sentence-level dialect rewrites. The micro-level rewrites are based on pairwise data from Demszky et al. (2021), in which each example includes a form containing at least one dialect feature from Indian English and a corresponding "base sentence" in U.S. English. We then apply the semantic perturbation to the base sentence as described in §4.1. For each perturbation type, one of the coauthors manually examined whether the perturbation successfully changes the meaning of the sentence. If all of the three perturbations fail, we exclude the instance from analysis.[9]

For sentence-level dialect analysis, we use the test set of the FRMT benchmark (Riley et al., 2022). Each instance contains an English sentence and its translations into dialects of the target languages Portuguese and Mandarin. For Portuguese, the two dialects are Brazilian Portuguese (pt-BR) and European Portuguese (pt-PT); for Mandarin, we consider mainland Mandarin and Taiwanese Mandarin, both in simplified script. As described in §4.2, semantic perturbations are obtained by perturbing the English sentences and then translating, using the Google Translate API. Table 8 (Appendix B) shows the number of evaluation examples.

6.1 Dialect robustness

We use the statistical methods reported in §4.3 to test metrics' sensitivity to dialects.

Regression Following Equation 3, we use $\sigma^{(\text{perturb})}_{m,i}$, $\sigma^{(\text{dialect})}_{m,i}$, $\sigma^{(\text{MT})}_{m,i}$ as conditions and model each metric as a mixed-effects regression. For a dialect-robust metric, we expect $\phi_{\text{dialect}} > \phi_{\text{perturb}}$, indicating that dialect rewrites score more highly than semantic perturbations, as required by definition 2.2. The difference $\phi_{\text{dialect}} - \phi_{\text{perturb}}$ is shown in the Y-axis of Figure 2. We also evaluate $\phi_{\text{MT}} - \phi_{\text{perturb}}$ as a stress test to measure metrics' abilities to recognize semantic changes, and to ensure that the semantic perturbations are effective. For all metrics except BLEURT and NANO, $\phi_{\text{dialect}} - \phi_{\text{perturb}}$ is negative for at least one language, indicating that these metrics are not dialect robust even in the average case. At the same time, all evaluation metrics can distinguish the MT and PERTURB conditions, showing that the issue is specific to dialect and not generally applicable to other paraphrases. Table 2 shows the coefficients before and after using NANO, which improves dialect robustness across all model sizes and languages.

Success Rates In Table 1 we report the success rates of a metric in assigning higher scores to dialect rewrites than to semantic perturbations. BLEURT performs better than other existing evaluation metrics which consistently fail to rank the dialect change above the perturbations. However, no metric correctly ranks the English examples at better than a random chance win rate (0.5), and even BLEURT as the most robust metric only has a 0.59 win rate for PT and ZH. In comparison with BLEURT, NANO achieves a higher win rate when scaled to XL and XXL, marked with ♦ in Table 1. The same trend can be observed in the regression analysis, where NANO's coefficients are positive for all metrics and languages. However, the marginal benefit of NANO over standard fine-

[9] For the sentences that have multiple dialect rewritings, we treat each one as an individual data point. When multiple semantic perturbations can be applied, we choose a single one, preferring replacements, then insertions, and then deletions.

	Learned			*Lexical*		**mT5$_{\text{base}}$**		**mT5$_{\text{XL}}$**		**mT5$_{\text{XXL}}$**	
	BLEURT	**PRISM**	**YISI**	**BLEU**	**CHRF**	**-NANO**	**+NANO**	**-NANO**	**♦+NANO**	**-NANO**	**🏆+NANO**
EN	0.53	0.51	0.53	0.49	0.46	0.50	0.50	0.55	0.54	0.57	0.57
PT	**0.59**	0.53	0.36	0.35	0.35	0.39	0.44	**0.57**	**0.65**	**0.82**	**0.81**
ZH	**0.59**	0.47	0.46	0.35	0.36	0.46	0.45	0.51	**0.59**	**0.74**	**0.74**

Table 1: Success Rates of $\sigma^{(\text{dialect})} > \sigma^{(\text{perturb})}$. Training with NANO starts to improve upon the strongest baseline BLEURT with mT5$_{\text{XL}}$ (♦) and achieves the best performance with mT5$_{\text{XXL}}$ (🏆). We **boldface** the success rates that are better than random chance (0.5) and significant after applying Bonferroni correction for multiple comparisons. Training with NANO improves dialect robustness for the XL- and base-scale model.

		EN	PT	ZH
mT5$_{\text{base}}$	-NANO	0.01$_{0.01}$	-0.02$_{0.00}$	-0.02$_{0.00}$
	+NANO	**0.04**$_{0.01}$	**-0.01** $_{0.00}$	0.00$_{0.00}$
mT5$_{\text{XL}}$	-NANO	0.01$_{0.01}$	0.02$_{0.00}$	0.02$_{0.00}$
	+NANO	**0.06**$_{0.01}$	**0.05**$_{0.00}$	**0.05**$_{0.00}$
mT5$_{\text{XXL}}$	-NANO	0.15$_{0.02}$	0.12$_{0.00}$	0.11$_{0.00}$
	+NANO	**0.19**$_{0.02}$	**0.15**$_{0.00}$	**0.13**$_{0.00}$

Table 2: Differences in metric values for *Dialect vs. Semantic Perturbation* before and after using NANO, as estimated by the regression model. We **boldface** significant coefficients where NANO helps.

tuning diminishes at scale—while NANO leads to significant improvements at XL scale, it has only a minor effect on the XXL model.

6.2 Align with Human Judgments

Does dialect robustness come at the cost of sacrificing the metrics' performance on standard benchmark of evaluation metrics? To study this, we evaluate on the test set of WMT 2020. We calculate the segment-level agreement with human ratings and report DaRR (Mathur et al., 2020), a robust variant of Kendall Tau. We follow Pu et al. (2021) and omit $*$-en results because of inconsistencies between benchmark implementations.

Results Table 3 and Table 12 (Appendix F.4) show the performance of existing methods and NANO on WMT 2020 test sets for within the same language and quality estimation settings respectively. In both settings, adding NANO improves the WMT benchmark performance of the mT5$_{\text{XL}}$ model compared to the finetuning-only setup. As in the dialect robustness tests, NANO does not help much for the model size XXL and achieves comparable results to finetuning-only settings. Our results are on par with or exceed those of prior metrics, demonstrating that dialect robustness is not in tension with other measures of metric quality.

	en-*	**en-cs**	**en-de**	**en-ja**	**en-pl**	**en-ru**	**en-ta**	**en-zh**
BLEURT	55.2	70.8	45.3	63.0	51.0	36.8	67.9	51.6
Prism	-	63.8	39.8	60.2	46.0	33.9	-	41.6
YiSi	35.6	50.1	32.7	44.8	21.7	24.0	35.7	40.0
-NANO $_{\text{XL}}$	49.2	68.2	41.0	63.0	48.6	30.8	68.5	51.0
+NANO $_{\text{XL}}$	54.2	69.8	41.9	63.7	49.9	33.2	70.2	50.9
-NANO $_{\text{XXL}}$	58.6	73.0	47.9	66.3	54.1	38.7	72.0	58.1
+NANO $_{\text{XXL}}$	58.3	72.4	47.1	66.3	53.6	39.4	72.2	56.9

Table 3: Segment-level agreement with human ratings on the WMT 2020 test set. The metric is WMT Metrics DaRR (Mathur et al., 2020), a robust variant of Kendall Tau and higher is better.

6.3 Transfer to Quality Estimation

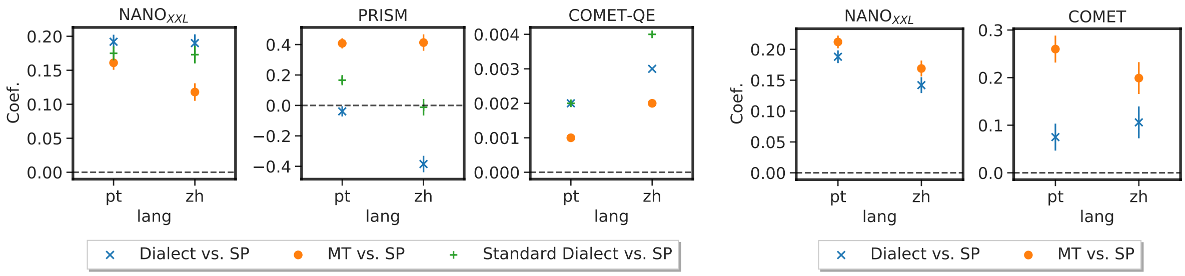
Quality Estimation While we have been focusing the cross-dialect setting within the same language, all the statistical methods can be applied to the cross-language cross-dialect setting, and training with NANO can serve as a quality estimation of the translation quality. Similar to §4.2, given an English base sentence EN_i and its translation to two locales (j and k) of a target language. We have

$$\sigma^j_{m,i} = m(\text{EN}_i, y^{(j)}_i) \tag{4}$$

$$\sigma^k_{m,i} = m(\text{EN}_i, y^{(k)}_i). \tag{5}$$

For a system that produces equally-good quality translations that are in different dialects j and k, we expect $\mathbb{E}[\sigma^j_m] \approx \mathbb{E}[\sigma^k_m] > \mathbb{E}[\sigma^{\text{perturb}}_m]$ for a metric that is robust to dialect variations. For the quality estimation, we can also use one dialect (k) as reference and evaluate other conditions (e.g., perturb, MT, dialect j) against dialect k as candidates for evaluation. We can use all statistical methods in §4.3 to understand the difference of outputs from evaluation metrics.

Experiment Setup We use the datasets for sentence-level dialect features for the quality estimation with and without references experiments. For quality estimation, we take the English sentences as the source and candidate from each of four conditions: two human-written dialects of target language (e.g., pt-BR), translated outputs to



(a) QE

(b) QE with Reference

Figure 3: Coefficients from the regression model for metric as quality estimation without and with references. NANO consistently rates dialects higher than semantic perturbations for both setups, and assigns similar ratings across dialects (● and ×) for quality estimation with references.

		PRISM	COMET	-NANO $_{XL}$	+NANO $_{XL}$	-NANO $_{XXL}$	+NANO $_{XXL}$
QE	PT	0.44	0.54	0.67	0.76	0.84	0.85
	ZH	0.30	0.53*	0.67	0.75	0.84	0.84
QE ref	PT	-	0.53	0.63	0.64	0.86	0.85
	ZH	-	0.53*	0.55	0.55	0.79	0.75

Table 4: The success rates of $\sigma^{(\text{dialect})} > \sigma^{(\text{perturb})}$ for Quality Estimation without and with references. NANO improves the dialect robustness upon existing metrics on both quality estimation settings.

target language from English and semantic perturbation as the input for the quality estimation. The translated outputs are from the Google Translate API. If a metric is robust to dialects, we expect $\mathbb{E}[\sigma_m^{\text{dialect}}] \geq \mathbb{E}[\sigma_m^{\text{MT}}] > \mathbb{E}[\sigma_m^{\text{perturb}}]$. For quality estimation with reference, we keep the same setting as the quality estimation but use one of the two dialects ("zh-CN" for Mandarin and "pt-BR" for Portuguese) as reference. We then use {perturb, MT, the other dialect} as candidates and estimate their quality with regard to the selected dialects.

Results We show that success rates of $\sigma^{(\text{dialect})} > \sigma^{(\text{perturb})}$ in QE with and without references in Table 4. We show that 1) training with NANO outperforms existing metrics on dialect robustness for both Portuguese and Mandarin; 2) NANO is important to improve dialect robustness with a smaller model size (i.e., mT5$_{\text{XL}}$ in our case). The trends are consistent with our findings for the within-language evaluation. Figure 3 shows the coefficients from the regression model and confirms the dialect robustness after training with NANO by assigning higher scores to dialects than semantic perturbations.

6.4 Dialect Awareness

Following Definition 2.3, we test whether it is possible to build metrics that reward outputs in a desired dialect. Because existing metrics do not train

Candidate	Input Tag	-NANO $_{XL}$	+NANO $_{XL}$	-NANO $_{XXL}$	+NANO $_{XXL}$
zh-TW	zh-TW	0.70 ✗	0.71 ✓	0.80 ✓	0.75 ✗
	zh-CN	0.70	0.68	0.77	0.78
zh-CN	zh-TW	0.74	0.68	0.80	0.77
	zh-CN	0.75 ✓	0.82 ✓	0.76 ✗	0.81 ✓

Table 5: Dialect Awareness test on simplified Mandarin. We score each variant against translation of English to Mandarin, with the dialect tag as input. ✓ shows when the metric assign a higher score for the candidate with the matched dialect identifier, indicating the dialect awareness, and ✗ shows when it does not. NANO$_{\text{XL}}$ successfully assigns higher scores when the candidates matches with the input tags.

with dialect identifiers, we are only able to test NANO's dialect awareness, which can serve as a baseline for future works. We use the Mandarin dataset for sentence-level dialect rewrites for our experiments of dialect awareness, because Mandarin is covered during the pretraining of NANO.[10] We then score each dialect rewrite against its translation from the English sentence, written as, $\sigma_{m,i}^{j} = m(\text{tag}, y_i^{(\text{MT})}, y_i^{(j)})$. The models we use are the ones we trained for dialect robustness tests in Table 1, but we provide specific dialect tags (e.g., zh-CN for Mainland Mandarin) instead of the general language tags (e.g., zh-any) as inputs for inference. During the model inference, we either provide tags that agree or disagree with the candidate sentences. For example, for a candidate sentence in Taiwanese Mandarin, we run inference with both "zh-CN" and "zh-TW". A dialect-aware metric should assign higher scores for the input with the correct dialect tag.

Results Table 5 shows the results of dialect awareness of NANO. NANO$_{\text{XL}}$ assigns higher

[10] We provide the zero-shot result of dialect awareness of NANO on PT in Appendix G.

scores to the candidates with the correct dialect tag, compared to the finetuning-only setup ($-\text{NANO}_{\text{XL}}$). However, at the XXL scale the picture is more mixed: NANO_{XXL} successfully assigns higher scores for zh-CN inputs with zh-CN tag over the zh-TW tag, but it fails on zh-TW inputs. This is compatible with our mixed findings on the impact of NANO on dialect robustness at the XXL scale.

7 Related Work

Most popular NLP datasets and evaluation metrics do not take dialectal variation into consideration. For example, machine translation systems are usually evaluated by whether they match references in the target language, for which the dialect is generally unspecified (Gehrmann et al., 2022). The subcommunity that has attended most to dialect is the VarDial series of workshops, which has featured shared tasks such as dialect classification (Zampieri et al., 2014), translation between dialects (Akhbardeh et al., 2021), and transfer of NLP systems across dialects (Zampieri et al., 2017). Of this prior work, dialect classification is clearly relevant to the criterion of dialect awareness introduced in Definition 2.3 (see also Nerbonne et al., 2011), but our goal is to reward system outputs that match a target dialect rather than to classify the dialect of existing human-written texts. A related line of work has focused on inducing dialect features from corpora (Eisenstein et al., 2010; Jørgensen et al., 2015; Dunn, 2021) and on recognizing dialect features in text (Demszky et al., 2021; Masis et al., 2022). Following the feature-based view of dialect, we use cross-dialectal minimal pairs to measure dialect robustness in §4.1.

On the more specific topic of dialect-aware evaluation, classic approaches focused on the creation of dialect-specific test sets, e.g. for translation to and from Arabic dialects (e.g., Zbib et al., 2012). This idea has been extended to modern multi-task natural language understanding benchmarks by the VALUE project (Ziems et al., 2022), which used transformation rules to convert the GLUE benchmarks (Wang et al., 2018) into African-American English. Our evaluation in §4.2 builds on the FRMT dataset of multi-dialectal translations (Riley et al., 2022) to evaluate metrics for dialect robustness. However, in many cases it is not feasible to produce multi-dialect references or test sets. In these cases, dialect-robust and dialect-aware metrics can provide a more efficient solution, particularly if these capabilities can be achieved through a pretraining step like NANO, which can be transferred to multiple tasks and evaluation settings.

Our work is additionally motivated by several papers that demonstrate the social impact of the failure to consider dialect variation in language technology. For example, literature shows that the out-of-the-box POS taggers (Jørgensen et al., 2015) and language identification and dependency parsing tools (Blodgett et al., 2016) perform poorly on AAVE texts. Other work has demonstrated large racial disparities in the performance of commercial speech recognition systems (DiChristofano et al., 2022; Koenecke et al., 2020). Our results contribute to this line of work by showing that metrics for text generation tend to penalize dialectal variants. We view the design of dialect-robust and dialect-aware metrics like NANO as a step towards making language technology that works more broadly across dialects.

8 Conclusion and Future Work

We introduce and formalize the dialect robustness and dialect awareness in the context of generation evaluation. Grounded by a suite of statistical tests, we find that existing evaluation methods are not robust to dialects. As a first step toward a solution to this problem, we propose NANO as a pretraining strategy. Our experiments demonstrate that NANO offers a size-efficient way to improve both the dialect robustness, shows the preliminary success towards dialect awareness and improves the metric performance of metrics on WMT benchmark.

Due to the limited availability of dialect-parallel corpora, our robustness tests are conducted in thousands of examples for Mandarin and Portuguese and hundreds of examples for English, which is insufficient to capture the full extent of these languages. We encourage future work to develop more resources, including benchmarks and corpora to conduct research on dialects for NLG evaluation. Due to this limitation, our work focuses on dialect robustness and only briefly evaluates dialect awareness. Future works may extend the details and criteria of the dialect-aware NLG evaluation, and we hope our work can serve as a baseline in this new research direction. Our encouraging preliminary results lead us to urge researchers to consider and improve the dialect diversity during pretraining.

Limitations

Besides the limited size of the evaluation corpora and a brevity of the exploration of dialect awareness that we point out as limitations in §8, we again acknowledge the data acquisition strategy as another limitation of our work. Our data acquisition of dialects requires country codes, which exclude many dialects. There is some work on getting dialectal data without country codes: Blodgett et al. (2016) build a dataset of tweets that are likely to include a high density of African-American English by linking geolocated Twitter data with demographic data from the U.S. census. However, this approach is limited to dialects that have strong geographic associations within the United States and which correlate with census demographics like race. Similarly, Abdul-Mageed et al. (2018) build a dataset of city-level Arabic dialects, again relying on Twitter geolocation. An alternative approach that does not rely on geolocation is to translate existing corpora into multiple dialects (e.g., Faisal et al., 2021; Ziems et al., 2022). However, this is labor intensive and therefore difficult to scale up to the amount of data needed for pretraining. We leave to future work the question of how to build large-scale corpora for dialects that do not align with easily-identifiable geographical indicators such as national boundaries.

References

Muhammad Abdul-Mageed, Hassan Alhuzali, and Mohamed Elaraby. 2018. You tweet what you speak: A city-level dataset of Arabic dialects. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

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.

R Harald Baayen. 2012. Mixed-effects models. *The Oxford handbook of laboratory phonology*, pages 668–677.

Su Lin Blodgett, Lisa Green, and Brendan O'Connor. 2016. Demographic dialectal variation in social media: A case study of African-American English. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1119–1130, Austin, Texas. Association for Computational Linguistics.

Jan A. Botha, Emily Pitler, Ji Ma, Anton Bakalov, Alex Salcianu, David Weiss, Ryan McDonald, and Slav Petrov. 2017. Natural language processing with small feed-forward networks. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2879–2885, Copenhagen, Denmark. Association for Computational Linguistics.

Asli Celikyilmaz, Elizabeth Clark, and Jianfeng Gao. 2020. Evaluation of text generation: A survey. *CoRR*, abs/2006.14799.

J.K. Chambers, P. Trudgill, and S.R. Anderson. 1998. *Dialectology*. Cambridge Textbooks in Linguistics. Cambridge University Press.

Aaron Daniel Cohen, Adam Roberts, Alejandra Molina, Alena Butryna, Alicia Jin, Apoorv Kulshreshtha, Ben Hutchinson, Ben Zevenbergen, Blaise Hilary Aguera-Arcas, Chung ching Chang, Claire Cui, Cosmo Du, Daniel De Freitas Adiwardana, Dehao Chen, Dmitry (Dima) Lepikhin, Ed H. Chi, Erin Hoffman-John, Heng-Tze Cheng, Hongrae Lee, Igor Krivokon, James Qin, Jamie Hall, Joe Fenton, Johnny Soraker, Kathy Meier-Hellstern, Kristen Olson, Lora Mois Aroyo, Maarten Paul Bosma, Marc Joseph Pickett, Marcelo Amorim Menegali, Marian Croak, Mark Díaz, Matthew Lamm, Maxim Krikun, Meredith Ringel Morris, Noam Shazeer, Quoc V. Le, Rachel Bernstein, Ravi Rajakumar, Ray Kurzweil, Romal Thoppilan, Steven Zheng, Taylor Bos, Toju Duke, Tulsee Doshi, Vinodkumar Prabhakaran, Will Rusch, YaGuang Li, Yanping Huang, Yanqi Zhou, Yuanzhong Xu, and Zhifeng Chen. 2022. Lamda: Language models for dialog applications. In *arXiv*.

Dorottya Demszky, Devyani Sharma, Jonathan Clark, Vinodkumar Prabhakaran, and Jacob Eisenstein. 2021. Learning to recognize dialect features. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2315–2338, Online. Association for Computational Linguistics.

Alex DiChristofano, Henry Shuster, Shefali Chandra, and Neal Patwari. 2022. Performance disparities between accents in automatic speech recognition. *ArXiv*, abs/2208.01157.

Rotem Dror, Gili Baumer, Marina Bogomolov, and Roi Reichart. 2017. Replicability analysis for natural language processing: Testing significance with multiple datasets. *Transactions of the Association for Computational Linguistics*, 5:471–486.

Jonathan Dunn. 2021. Finding variants for construction-based dialectometry: A corpus-based approach to regional cxgs. *CoRR*, abs/2104.01299.

Jacob Eisenstein, Brendan T. O'Connor, Noah A. Smith, and Eric P. Xing. 2010. A latent variable model for geographic lexical variation. In *EMNLP*.

Fahim Faisal, Sharlina Keshava, Md Mahfuz Ibn Alam, and Antonios Anastasopoulos. 2021. SD-QA: Spoken dialectal question answering for the real world. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3296–3315, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Sebastian Gehrmann, Elizabeth Clark, and Thibault Sellam. 2022. Repairing the cracked foundation: A survey of obstacles in evaluation practices for generated text. *CoRR*, abs/2202.06935.

Anna Jørgensen, Dirk Hovy, and Anders Søgaard. 2015. Challenges of studying and processing dialects in social media. In *Proceedings of the Workshop on Noisy User-generated Text*, pages 9–18, Beijing, China. Association for Computational Linguistics.

Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R Rickford, Dan Jurafsky, and Sharad Goel. 2020. Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences*, 117(14):7684–7689.

Claudia Lange. 2012. *The syntax of spoken Indian English*. John Benjamins Publishing Company Amsterdam.

Beatriz R Lavandera. 1978. Where does the sociolinguistic variable stop? *Language in society*, 7(2):171–182.

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.

Tessa Masis, Anissa Neal, Lisa Green, and Brendan O'Connor. 2022. Corpus-guided contrast sets for morphosyntactic feature detection in low-resource english varieties. *arXiv preprint arXiv:2209.07611*.

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

John Nerbonne. 2009. Data-driven dialectology. *Lang. Linguistics Compass*, 3:175–198.

John Nerbonne, Rinke Colen, Charlotte Gooskens, Peter Kleiweg, and Therese Leinonen. 2011. Gabmap - a web application for dialectology. *Dialectologia*.

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.

Amy Pu, Hyung Won Chung, Ankur Parikh, Sebastian Gehrmann, and Thibault Sellam. 2021. Learning compact metrics for MT. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 751–762, Online and Punta Cana, Dominican Republic. 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.

Parker Riley, Timothy Dozat, Jan A. Botha, Xavier Garcia, Dan Garrette, Jason Riesa, Orhan Firat, and Noah Constant. 2022. Frmt: A benchmark for few-shot region-aware machine translation.

Adam Roberts, Hyung Won Chung, Anselm Levskaya, Gaurav Mishra, James Bradbury, Daniel Andor, Sharan Narang, Brian Lester, Colin Gaffney, Afroz Mohiuddin, Curtis Hawthorne, Aitor Lewkowycz, Alex Salcianu, Marc van Zee, Jacob Austin, Sebastian Goodman, Livio Baldini Soares, Haitang Hu, Sasha Tsvyashchenko, Aakanksha Chowdhery, Jasmijn Bastings, Jannis Bulian, Xavier Garcia, Jianmo Ni, Andrew Chen, Kathleen Kenealy, Jonathan H. Clark, Stephan Lee, Dan Garrette, James Lee-Thorp, Colin Raffel, Noam Shazeer, Marvin Ritter, Maarten Bosma, Alexandre Passos, Jeremy Maitin-Shepard, Noah Fiedel, Mark Omernick, Brennan Saeta, Ryan Sepassi, Alexander Spiridonov, Joshua Newlan, and Andrea Gesmundo. 2022. Scaling up models and data with `t5x` and `seqio`. *arXiv preprint arXiv:2203.17189*.

Suzanne Romaine. 1981. On the problem of syntactic variation: A reply to beatriz lavandera and william labov. sociolinguistic working paper number 82.

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.

Dirk Speelman, Kris Heylen, and Dirk Geeraerts. 2018. *Mixed-effects regression models in linguistics*. Springer.

Brian Thompson and Matt Post. 2020. Automatic machine translation evaluation in many languages via zero-shot paraphrasing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online. Association for Computational Linguistics.

Daan van Esch, Tamar Lucassen, Sebastian Ruder, Isaac Caswell, and Clara Rivera. 2022. Writing system and speaker metadata for 2,800+ language varieties. In *Proceedings of LREC*.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

M Farr Whiteman. 2013. *Writing: The nature, development, and teaching of written communication*. Routledge.

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.

Marcos Zampieri, Shervin Malmasi, Nikola Ljubešić, Preslav Nakov, Ahmed Ali, Jörg Tiedemann, Yves Scherrer, and Noëmi Aepli. 2017. Findings of the vardial evaluation campaign 2017. In *Proceedings of the fourth workshop on NLP for similar languages, varieties and dialects*.

Marcos Zampieri, Liling Tan, Nikola Ljubešić, and Jörg Tiedemann. 2014. A report on the dsl shared task 2014. In *Proceedings of the first workshop on applying NLP tools to similar languages, varieties and dialects*, pages 58–67.

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.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Yuan Zhang, Jason Riesa, Daniel Gillick, Anton Bakalov, Jason Baldridge, and David Weiss. 2018. A fast, compact, accurate model for language identification of codemixed text. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 328–337, Brussels, Belgium. Association for Computational Linguistics.

Caleb Ziems, Jiaao Chen, Camille Harris, Jessica Anderson, and Diyi Yang. 2022. VALUE: Understanding dialect disparity in NLU. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3701–3720, Dublin, Ireland. Association for Computational Linguistics.

A Example Semantic Perturbations

Table 6 shows the task instruction and examples we used to prompt LaMDA for the automatic semantic perturbation on English sentences, for both micro-level (§4.1) and sentence-level (§4.2) studies. During decoding, we use greedy decoding.

B Languages and Variants

Table 7 shows the language codes and region codes that we cover during NANO pretraining. We cover 10 WMT languages and 95 language variants, presented as BCP language codes. Although `iu` is one of the WMT languages, it is not supported by LangID model that we are using and we thus do not include it in our pretraining. Portuguese (PT) is not included because it is not a WMT language. Therefore, all NANO dialect robustness results on PT are fully through zero-shot transfer. We report additional experiments that include Portuguese during pretraining in Appendix E. Our experiments show that pretraining with all languages leads to better dialect robustness on both PT and ZH.

C Metric Implementations

We use the official releases of Prism (Thompson and Post, 2020), COMET (Rei et al., 2020) and BLEURT (Pu et al., 2021) in our work. For YiSi, we use an internal implementation. Table 9 presents the supported setups for each model in their latest released versions. Although all metrics could in theory be adapted to different use cases, their existing capabilities restrict the experiments we can run with them. For BLEURT,[11] we use the latest checkpoint `BLEURT-20`. We use Prism[12] (`m39v1` checkpoint) for quality estimation with and without references. Lastly, there are two models that we use for COMET[13]. Model `wmt21-comet-qe-mqm` is for reference-free quality estimation and `wmt20-comet-da` for reference-based quality estimation. For our experiments, if a language is not supported by the model, we exclude it from the results.

D Training Details and Hyperparameters of NANO

Hyperparameters We implement NANO using T5X and SeqIO (Roberts et al., 2022). We experimented with the following hyperparemeters during training: learning rate of {1e-3, 1e-4, 1e-5, 3e-5, 5e-5} × sequence length of {512, 1024}. The reported results are based on a learning rate of $1e-4$ and sequence length of 1024. We train for 200,000 steps for pretraining and another 20,000 steps for finetuning. We set the drop out rate as 0.1 and optimizer warm up steps as 10,000. We train with a batch size of 128.

Choosing Checkpoints We calculate the Kendall-Tau correlation on the development set every 1000 steps throughout training and choose the checkpoint with the highest correlation as the final checkpoint for evaluation.

Compute Time Our models are trained on 64 TPUs, pretraining step normally takes one day to finish across different sizes. While mT5$_{\text{small}}$ can be trained within a single day, finetuning mT5$_{\text{XL}}$ and mT5$_{\text{XXL}}$ takes three and nine days respectively to reach 20,000 steps, but the models converge before they finish training.

E NANO Design Choices

Table 10 shows different variations of NANO and their performances. We studied:

- Comparing pretraining on all WMT language variants to only prertaining on zh/pt or zh/pt/en.
- Comparing $\lambda = 1$ to $\lambda = 0$ and $\lambda = 2$, i.e., the balance in pretraining between dialect-tags and language-tags.
- Variations of the ratio of positive vs. negative instances during pretraining. We compare a balanced set to a setup where we have twice as many positive as negative examples.

We gain the following insights: 1) using all WMT languages for pretraining performs better than using partial data; 2) An equal balance between dialect-tags and general language tags ($\lambda = 1$) during pretraining improves upon a higher fraction of dialect-tags ($\lambda = 2$). However, using *only* data with general language tags ($\lambda = 0$) surprisingly leads to an even better **dialect-robustness**, although the model will lose its potential for **dialect-awareness** since it never sees dialect tags; 3) A balanced set of positive and negative instances during pretraining is better than oversampling positive instances.

Following Equation 3, we use $\sigma^{(\text{perturb})}_{m,i}$, $\sigma^{(\text{dialect})}_{m,i}$, $\sigma^{(\text{MT})}_{m,i}$ as conditions and model each metric as a mixed-effects regression. Table 11 shows ϕ_{dialect}

[11] https://github.com/google-research/bleurt.

[12] https://github.com/thompsonb/prism.

[13] https://github.com/Unbabel/COMET

	Task Instruction	Examples	Output Prefix
Delete	Generate a sentence by deleting one word from the original sentence and change its meaning.	**Original Sentence:**	**\nDelete one word from original sentence:**
		the person I like the most is from the mechanical department	the person I like is from the mechanical department
		a recipe is a simple thing	it is a simple thing
		the union person contacted his representative at the school	the union person contacted his representative
		we have two tailors who can make them for us	we have two tailors who can make them
		So if you're not good at communication you may get filtered at even the first level	So if you're good at communication you may get filtered at even the first level
Replace	Generate a sentence by replacing one word from the original sentence and change its meaning.	**Original Sentence:**	**\nReplace one word from original sentence:**
		the person I like the most is from the mechanical department	the person I like the least is from the mechanical department
		a recipe is a simple thing	a recipe is a complicated thing
		the union person contacted his representative at the school	the union person contacted his representative at the factory
		we have two tailors who can make them for us	we have three tailors who can make them for us
		he didn't give it to me	he didn't give it to anyone
Insert	Add one word to a sentence and change its meaning.	**Original Sentence:**	**\nAdd a word to it:**
		it was the first day of term	it was the first day of spring term
		the person I like the most is from the mechanical department	the person I like to talk to the most is from the mechanical department
		he does a lot of things	he does a lot of funny things
		my brother said that one of his favorite places is the beach nearby	my brother said that one of his least favorite places is the beach nearby
		I think you should start going to the gym from now on	I think you should start going to the other gym from now on

Table 6: The prompts, prefix and five examples that we use to prompt LaMDA for automatic semantic perturbation on English sentences. We include three types of semantic perturbation: replace (highlighted in yellow), delete (highlighted in blue) and insert (highlighted in purple).

Language	Region Code
en	AU, BZ, BM, BR, CA, KY, CK, CU, DO, FK, GI, GP, GT, GY, HN, IE, LR, MX, NF, PN, SH, ZA, SR, GB, US, VE, IN
cs	AT, CZ, PL, SK
de	AT, BE, CZ, DK, FR, DE, HU, IT, LI, LU, NL, PL, RO, SK, SI, CH
ja	JP
km	KH, LA, TH, VN
pl	BY, CZ, DE, LT, PL, RU, SK, UA
ps	PK
ru	BY, CN, EE, FI, GE, KZ, KP, KG, LV, LT, MD, MN, NO, PL, RO, RU, TM, UA, UZ
ta	IN, LK
zh-cmn-Hans	CN, KP, LA, VN, TW, MM, MN, RU
zh-yue	CN, VN, HK
zh-cmn-Hant	CN, TW

Table 7: The language code and region code that we cover. We consider 10 WMT languages and use BCP language codes. We underline selected English dialects under the increasing noise setup `zh, pt, en` in §5.1.

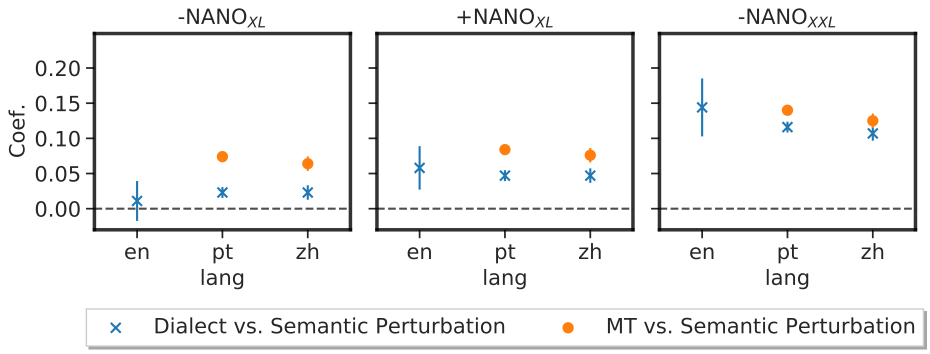


Figure 4: Coefficients from the regression model for *Dialect vs. Semantic Perturbation* ($\phi_{\text{dialect vs. perturb}}$) and *MT vs. Semantic Perturbation* of NANO across XL and XXL model sizes. Training with NANO improves the dialect robustness for both model sizes. This figure is complementary to Figure 2.

	EN	PT	ZH
All	148	2616	2227
Replace	96	962	866
Insert	89	550	528
Delete	63	693	614
AGG.	115	1415	1252

Table 8: Number of evaluation examples per language before and after semantic perturbation. The middle three rows are the number of examples to which each perturbation was applicable, and the final row AGG. is the number of examples to which at least one perturbation is applicable, which we use in our final analysis.

	BLEURT	PRISM	YiSi	COMET	NANO
Within	✓	✓	✓		✓
QE		✓		✓	✓
QE w/ Ref				✓	✓

Table 9: Supported setups for different metrics.

with its standard errors against the ϕ_{perturb} condition. Take ϕ_{perturb} for BLEURT under EN as an example, -0.09 with an error smaller than 0.05 means that semantic perturbation would result in a decrease of 0.09 point for BLEURT compared to the dialect condition, and the result is significant. For a dialect-robust metric, we expect its ϕ_{perturb} to be positive. However, this is not always true during our observations. BLEURT performs the best among existing evaluation metrics and all other existing metrics have positive ϕ_{perturb} for at least one language of our test data. This indicates that existing evaluation metrics wrongly assign a higher score to semantically-perturbed rewriting than the dialects in at least one of the three languages, suggesting that they should not be used to assess dialects they were not trained for.

F Versatility of NANO

F.1 Input Format

We use the following input format to adapt NANO to different use cases.

- For within-language assessment, we format the input as `candidate:` *{sentence}* `reference:` *{reference}* `language:` *{language_tag}*.
- For quality estimation without reference, we format the input as `candidate:` *{sentence}* `source:` *{source}* `language:` *{language_tag}*.
- For quality estimation with reference, we format the input as `candidate:` *{sentence}* `reference:` *{reference}* `source:` *{source}* `language:` *{language_tag}*.

The *{language_tag}* during fine-tuning indicates the language where the candidate sentence comes from, but it is the general language tag (e.g., “en-any”) and does not contain the dialect information. We finetune one model for each setting.

F.2 Dialect Robustness

We show additional results of coefficients from the regression model across XL and XXL sizes in Figure 2, which shows that training with NANO improves the dialect robustness across both sizes and for all languages. In addition, we compare three pretraining settings: 1) Mandarin and Portuguese only; 2) Mandarin, Portuguese and selected English dialects and 3) ten languages with metric finetuning data evaluated during the WMT benchmark with ninety-five language variants following the classification by van Esch et al. (2022). Table 10 shows the exact numbers for both coefficients and success rates. NANO performs the best with the

		EN		PT				ZH			
		$\phi_{\text{dl vs. pb}}\uparrow$	$\underline{R_{\text{pb}}}\uparrow$	$\phi_{\text{dl vs. pb}}\uparrow$	$\phi_{\text{dl vs. MT}}$	$\underline{R_{\text{pb}}}\uparrow$	$\underline{R_{\text{MT}}}$	$\phi_{\text{dl vs. pb}}\uparrow$	$\phi_{\text{dl vs. MT}}$	$\underline{R_{\text{pb}}}\uparrow$	$\underline{R_{\text{MT}}}$
	BLEURT	$0.09_{0.01}$	$0.53^{*\dagger}$	$0.03_{0.01}$	$-0.07_{0.01}$	0.59	0.19	$0.04_{0.01}$	$-0.04_{0.01}$	0.59	0.33
mT5$_{\text{base}}$	Finetuning	$0.01^{*}_{0.01}$	$0.50^{*\dagger}$	$-0.02_{0.00}$	$-0.09_{0.00}$	0.39	0.13	$-0.02_{0.00}$	$-0.08_{0.00}$	$0.46^{\dagger}$	0.31
	NANO $_{\text{all}\,\vert\,\lambda=1}$	$0.04_{0.01}$	$0.50^{*\dagger}$	$-0.01_{0.00}$	$-0.08_{0.00}$	0.44	0.16	$0.00^{*}_{0.00}$	$-0.08_{0.00}$	$0.45^{\dagger}$	0.28
mT5$_{\text{XL}}$	Finetuning	$0.01^{*}_{0.01}$	$0.55^{*\dagger}$	$0.02_{0.00}$	$-0.05_{0.00}$	0.57	0.21	$0.02_{0.00}$	$-0.04_{0.00}$	0.51^{*}	0.31
	♦ NANO $_{\text{all}\,\vert\,\lambda=1}$	$0.06_{0.01}$	$\mathbf{0.54}^{*\dagger}$	$0.05_{0.00}$	$-0.04_{0.00}$	**0.65**	**0.25**	$0.05_{0.00}$	$-0.03_{0.00}$	**0.59**	**0.35**
	NANO $_{\text{zh/pt}\,\vert\,\lambda=1}$	$0.03_{0.01}$	$0.53^{*\dagger}$	$0.03_{0.00}$	$-0.04_{0.00}$	0.59	0.23	$0.03_{0.00}$	$-0.03_{0.00}$	$0.54^{\dagger}$	0.32
	NANO $_{\text{zh/pt/en}\,\vert\,\lambda=1}$	$0.06_{0.01}$	$0.53^{*\dagger}$	$0.04_{0.00}$	$-0.04_{0.00}$	0.64	0.24	$0.04_{0.00}$	$-0.03_{0.00}$	0.57	0.33
	NANO $_{\text{all}\,\vert\,\text{pos:neg}=2}$	$0.21_{0.02}$	$0.53^{*\dagger}$	$0.04_{0.00}$	$-0.04_{0.00}$	0.60	0.23	$0.04_{0.00}$	$-0.03_{0.00}$	0.56	0.33
mT5$_{\text{XXL}}$	Finetuning	$0.15_{0.02}$	$0.57^{*\dagger}$	$0.12_{0.00}$	$-0.02_{0.00}$	0.82	0.32	$0.11_{0.00}$	$-0.02_{0.00}$	0.74	0.38
	🏆 NANO $_{\text{all}\,\vert\,\lambda=1}$	$0.19_{0.02}$	$\mathbf{0.57}^{*\dagger}$	$0.15_{0.00}$	$-0.02_{0.02}$	**0.81**	**0.35**	$0.13_{0.00}$	$-0.01_{0.00}$	**0.74**	**0.38**
	NANO $_{\text{zh/pt}\,\vert\,\lambda=1}$	$0.19_{0.02}$	$0.54^{*\dagger}$	$0.13_{0.00}$	$-0.02_{0.00}$	0.80	0.33	$0.12_{0.00}$	$-0.01_{0.00}$	0.73	0.41
	NANO $_{\text{zh/pt/en}\vert\lambda=1}$	$-0.18_{0.02}$	$0.56^{*\dagger}$	$0.13_{0.00}$	$-0.02_{0.00}$	0.80	0.34	$0.12_{0.00}$	$-0.02_{0.00}$	0.73	0.39
	NANO $_{\text{all}\,\vert\,\lambda=0}$	$0.20_{0.02}$	$0.53^{*\dagger}$	$0.15_{0.00}$	$-0.02_{0.02}$	0.82	0.35	$0.13_{0.00}$	$-0.01_{0.02}$	0.76	0.40
	NANO $_{\text{all}\,\vert\,\lambda=2}$	$0.20_{0.02}$	$0.56^{*\dagger}$	$0.15_{0.00}$	$-0.02_{0.02}$	0.81	0.34	$0.13_{0.00}$	$-0.01_{0.00}$	0.75	0.40

Table 10: Dialect Robustness Tests for metrics with and without NANO. "pb" and "dl" are short for "perturb" and "dialect". R_{pb} and R_{MT} are the success rates of $\sigma^{(\text{dialect})} > \sigma^{(\text{perturb})}$ and $\sigma^{(\text{dialect})} > \sigma^{(\text{MT})}$ correspondingly. Standard errors of coefficients are in the subscript. We can observe that that 1) pretraining improves the dialect robustness compared to the finetuning-only setting and 2) Pretraining on more languages improves the dialect robustness. ↑ means higher is better. The success rates ($\underline{R}$) are comparable across metrics, but co-efficients from regression models are only comparable within the same metric. NANO based on mT5$_{\text{XL}}$ with full data improves upon the strongest baseline ♦ and achieves the best performance with mT5$_{\text{XXL}}$. This is complementary to Table 1.

		BLEURT	**PRISM**	**YiSi**	**BLEU**	**CHRF**
$\phi_{\text{dl vs. pb}}$	EN	$0.10_{0.01}$	$0.34_{0.05}$	$-0.05_{0.01}$	$-12.01_{1.91}$	$0.03^{*}_{0.01}$
	PT	$0.03_{0.00}$	$0.06_{0.01}$	$-0.02_{0.00}$	$-8.39_{0.53}$	$-0.05_{0.00}$
	ZH	$0.04_{0.00}$	$-0.02^{*}_{0.02}$	$-0.00_{0.00}$	$-0.34^{*}_{0.49}$	$-0.05_{0.00}$

Table 11: Coefficients from Equation 3 with standard errors in the subscript. We mark the ones that have p-value $\geq$ 0.05/5 = 0.01 with * using Bonferroni correction per metric. $\phi_{\text{dl vs. pb}}$ indicates the corresponding score increase (positive value) for dialect (dl) edits compared to the semantic perturbation (pb). For a dialect-robust metric, we expect $\phi_{\text{dl vs. pb}}$ to be positive.

	en-*	**en-cs**	**en-de**	**en-ja**	**en-pl**	**en-ru**	**en-ta**	**en-zh**
COMET	51.4	70.9	37.3	51.5	48.9	39.4	61.3	50.3
Prism	-	48.3	26.5	38.2	18.8	11.6	-	11.3
-NANO $_{\text{XL}}$	51.4	68.7	40.6	59.6	44.3	28.2	66.3	51.8
+NANO $_{\text{XL}}$	53.8	69.5	42.7	62.6	47.1	31.5	68.4	54.8
-NANO $_{\text{XXL}}$	57.4	71.4	47.1	65.5	52.4	36.3	70.3	58.7
+NANO $_{\text{XXL}}$	57.6	71.8	46.6	66.3	51.0	38.5	70.4	58.8

Table 12: Segment-level agreement with human ratings for metrics as quality estimation without references.

full set of languages. NANO improves the success rates under the XL size, but reach comparable results with training without NANO under the XXL size. We suspect the discrepancy between getting a higher coefficients but having nearly the same success rates is because some big increase of score after applying NANO which does not influence the success rates.

F.3 Transfer on Reference-based QE

For a system that produces equally-good quality translations that are in different dialects j and k, we expect $\mathbb{E}[\sigma_m^j] \approx \mathbb{E}[\sigma_m^k] > \mathbb{E}[\sigma_m^{\text{perturb}}]$ for a metric that is robust to dialect variations.

Quality Estimation with Reference For the quality estimation, we can also use one dialect (k) as reference and evaluate other conditions (e.g., perturb, MT, dialect j) against dialect k as candidates for evaluation, written as:

$$\sigma_{m,i}^{j} = m(\text{EN}_i, y_i^{(j)}, y_i^{(k)}) \tag{6}$$

$$\sigma_{m,i}^{\text{perturb}} = m(\text{EN}_i, y_i^{(\text{perturb})}, y_i^{(k)}). \tag{7}$$

For a metric that is robust to dialect variations, we expect $\mathbb{E}[\sigma_m^j] > \mathbb{E}[\sigma_m^{\text{perturb}}]$. The candidate can also be $y_i^{(\text{MT})}$. We can use all statistical methods in §4.3 to understand the difference in outputs from evaluation metrics.

We report NANO's performance on dialect robustness as the reference-based quality estimation in Table 13 and its corresponding WMT performance in Table 14. In the XL setting, NANO improves upon both COMET and the finetuning only setup for the dialect robustness and perfor-

		COMET	-NANO $_{\text{XL}}$	NANO $_{\text{XL}}$	-NANO $_{\text{XXL}}$	NANO $_{\text{XXL}}$
PT	R_{pb}	0.54	0.67	0.76	0.84	0.85
	R_{MT}	0.52	0.64	0.65	0.69	0.67
ZH	R_{pb}	0.53	0.67	0.75	0.84	0.84
	R_{MT}	0.50*	0.54	0.64	0.74	0.75

Table 13: NANO performance on reference-based QE.

	en-*	en-cs	en-de	en-ja	en-pl	en-ru	en-ta	en-zh
COMET	51.4	70.9	37.3	51.5	48.9	39.4	61.3	50.3
FT_{XL}	51.4	68.7	40.6	59.6	44.3	28.2	66.3	51.8
NANO $_{\text{XL}}$	53.8	69.5	42.7	62.6	47.1	31.5	68.4	54.8
FT_{XXL}	57.4	71.4	47.1	65.5	52.4	36.3	70.3	58.7
NANO $_{\text{XXL}}$	57.6	71.8	46.6	66.3	51.0	38.5	70.4	58.8

Table 14: Segment-level agreement with human ratings for reference-based quality estimation on WMT.

mance on WMT benchmark. However, NANO achieves comparable performances with finetuning-only setting with XXL models. The findings are consistent with our findings for within-language and reference-free quality estimation settings in the main content: NANO provides a size-efficient way for models to improve the dialect robustness and their performance on the WMT metrics benchmark.

F.4 Performance on WMT Tasks

We have shown that NANO is more robust to dialects. Is the robustness at the cost of sacrificing the metrics' performance on standard benchmark of evaluation metrics? To study this, we evaluate on the test set of WMT 2020.

Metrics We calculate the segment-level agreement with human ratings and report DaRR (Mathur et al., 2020), a robust variant of Kendall Tau. We follow Pu et al. (2021) and omit `*-en` results because of inconsistencies between benchmark implementations.

Results Table 3 and Table 12 show the performance of existing methods and NANO on WMT 2020 test sets for within the same language and quality estimation settings respectively. In both settings, adding NANO improves mT5_{XL} model's performance on WMT benchmark tasks compared to the finetuning-only setup. As in the dialect robustness tests, NANO does not help much for the model size XXL and achieves comparable results to finetuning-only settings. Moreover, our results are on par with or exceed those of prior metrics, demonstrating that mT5 is an effective base model for developing new metrics.

Candidate	Input Tag	FT_{XL}	NANO $_{\text{XL}}$	FT_{XXL}	NANO $_{\text{XXL}}$
	perturb	0.89	0.85	0.78	0.79
pt-BR	pt-BR	0.89	0.88	0.88	0.85
	pt-PT	0.88	0.87	0.88	0.93
pt-PT	pt-BR	0.85	0.84	0.85	0.84
	pt-PT	0.84	0.84	0.85	0.91

Table 15: Dialect Awareness test of NANO on Portuguese. We score each variant against a translation of English to Portuguese, with the dialect tag as input.

G Dialect Awareness on PT

Table 15 shows the dialect awareness test of NANO on Portuguese. As Portuguese and its language variants are not covered in pretraining, we expect NANO to not perform well in terms of dialect awareness because it has never seen the input dialect tags during training. Table 15 confirms our expectation. We observe that both finetuning-only and pretraining with NANO fail to assign higher scores to candidates with matched input language tags over mismatched dialect tags.

ACL 2023 Responsible NLP Checklist

A For every submission:

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

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

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

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

B ☑ Did you use or create scientific artifacts?

Section 5.1

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

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

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

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

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

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

C ☑ Did you run computational experiments?

Left blank.

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

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

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

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

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

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

Left blank.

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

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

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

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

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