Conformalizing Machine Translation Evaluation

Chrysoula Zerva[1,2,3] **André F. T. Martins**[1,2,3,4]
[1]Instituto de Telecomunicações, Portugal
[2]Instituto Superior Técnico, Universidade de Lisboa, Portugal
[3]ELLIS Unit Lisbon, Portugal [4]Unbabel, Portugal
{chrysoula.zerva, andre.t.martins}@tecnico.ulisboa.pt

Abstract

Several uncertainty estimation methods have been recently proposed for machine translation evaluation. While these methods can provide a useful indication of when not to trust model predictions, we show in this paper that the majority of them tend to underestimate model uncertainty, and as a result, they often produce misleading confidence intervals that do not cover the ground truth. We propose as an alternative the use of *conformal prediction*, a distribution-free method to obtain confidence intervals with a theoretically established guarantee on coverage. First, we demonstrate that split conformal prediction can ''correct'' the confidence intervals of previous methods to yield a desired coverage level, and we demonstrate these findings across multiple machine translation evaluation metrics and uncertainty quantification methods. Further, we highlight biases in estimated confidence intervals, reflected in imbalanced coverage for different attributes, such as the language and the quality of translations. We address this by applying conditional conformal prediction techniques to obtain calibration subsets for each data subgroup, leading to *equalized* coverage. Overall, we show that, provided access to a calibration set, conformal prediction can help identify the most suitable uncertainty quantification methods and adapt the predicted confidence intervals to ensure fairness with respect to different attributes.[1]

1 Introduction

Neural models for natural language processing (NLP) are able to tackle increasingly challenging tasks with impressive performance. However, their deployment in real-world applications does not come without risks. For example, systems that generate fluent text might mislead users with fabricated facts, particularly if they do not expose their confidence. High performance does not guarantee an accurate prediction for *every* instance—for example, the degradation tends to be more severe when instances are noisy or out of distribution. This makes uncertainty quantification methods more important than ever.

While most work on uncertainty estimation for NLP has focused on classification tasks, uncertainty quantification for text *regression* has recently gained traction, with applications in machine translation (MT) evaluation, semantic sentence similarity, or sentiment analysis (Wang et al., 2022; Glushkova et al., 2021). This line of work builds upon a wide range of methods proposed for estimating uncertainty (Kendall and Gal, 2017a; Kuleshov et al., 2018a; Amini et al., 2020; Ulmer et al., 2023). However, current uncertainty quantification methods suffer from three important limitations:

1. Most methods provide **confidence intervals without any theoretically established guarantees with respect to coverage**. In other words, while a representative confidence interval should include (cover) the ground truth target value for each instance (and ideally the bound of the confidence interval should be close in expectation to the ground truth as shown in Figure 1), the predicted interval is often much narrower and underestimates the model uncertainty. In fact, for the concrete problem of MT evaluation, we show that **the majority of uncertainty quantification methods achieve very low coverage** even after calibration, as can be observed in Figures 2 and 5.

2. Most proposed methods involve **underlying assumptions on the distribution** (e.g., Gaussianity) **or the source of uncertainty** (e.g., aleatoric or epistemic) which are often unrealistic and may lead to misleading (over-

[1]Code and data can be accessed on https://github.com/deep-spin/conformalizing_MT_eval.

Transactions of the Association for Computational Linguistics, vol. 12, pp. 1460–1478, 2024. https://doi.org/10.1162/tacl_a_00711
Action Editor: Joel Tetreault. Submission batch: 8/2023; Revision batch: 4/2024; Published 11/2024.

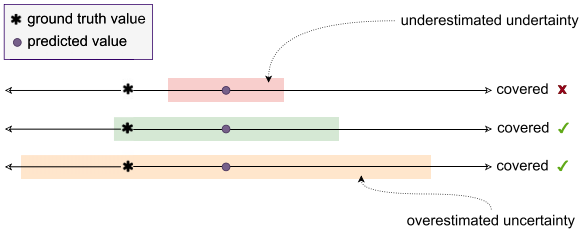


Figure 1: Predicted confidence intervals and coverage for the same ground truth/prediction points. We consider the middle (green) interval to be desired as it covers the ground truth but does not overestimate uncertainty.

or under-estimated) results (Izmailov et al., 2021; Zerva et al., 2022). Hence, choosing a suitable method for a dataset can be complicated.

3. While uncertainty quantification can shed light on model weaknesses and biases, **the uncertainty prediction methods themselves can suffer from biases** and provide unfair and misleading predictions for specific data subgroups or for examples with varying levels of difficulty (Cherian and Candès, 2023; Ding et al., 2020; Boström and Johansson, 2020).

To address the shortcomings above, we propose **conformal prediction** (§2) as a means to obtain more trustworthy confidence intervals on textual regression tasks, using MT evaluation as the primary paradigm. We rely on the fact that given a scoring or uncertainty estimation function, conformal prediction can provide statistically rigorous uncertainty intervals (Angelopoulos and Bates, 2021; Vovk et al., 2005, 2022). More importantly, the conformal prediction methodology provides theoretical guarantees about coverage over a test set, given a chosen coverage threshold. The predicted uncertainty intervals are thus valid in a **distribution-free** sense: They possess explicit, non-asymptotic guarantees even without distributional assumptions or model assumptions (Angelopoulos and Bates, 2021; Vovk et al., 2005), and they also allow for an intuitive interpretation of the confidence interval width.

We specifically show (§3) that previously proposed uncertainty quantification methods can be used to design non-conformity scores for split conformal prediction (Papadopoulos, 2008). We confirm that, regardless of the initially obtained coverage, the application of conformal prediction can increase coverage to the desired—user

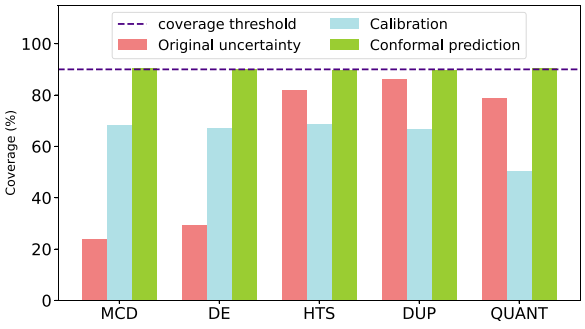


Figure 2: Coverage obtained by different uncertainty predictors. We compare originally obtained values (red), with values after calibration (light blue), and conformal prediction (green) for the desired coverage (dashed line) set to 0.9 (90%).

defined—value (see Figure 2). To this end, we compare four parametric uncertainty estimation methods (Monte Carlo dropout, deep ensembles, heteroscedastic regression, and direct uncertainty prediction) and one non-parametric method (quantile regression) with respect to coverage and distribution of uncertainty intervals. Additionally, we introduce a translation-inspired measure for referenceless quality estimation (QE) that uses the distance between quality estimates of translated and back-translated text to estimate non-conformity. We show that the estimated quantiles over each non-conformity score are indicative not only of the coverage but also of the overall suitability of the non-conformity score and the performance of the underpinning uncertainty quantification method (e.g. it aligns well with error correlation computed over the test set). Our experiments highlight the efficacy of quantile regression, a previously overlooked method for the MT evaluation task.

Moreover, we investigate the *fairness* of the obtained intervals (§4) for a set of different attributes: (1) translation language pair; (2) translation difficulty, as reflected by source sentence length and syntactic complexity, (3) estimated quality level and (4) uncertainty level. We highlight unbalanced coverage for all cases and demonstrate how **equalized conformal prediction** (Angelopoulos and Bates, 2021; Boström and Johansson, 2020; Boström et al., 2021) can address such imbalances effectively.

2 Conformal Prediction

In this section, we provide background on conformal prediction and introduce the notation used

throughout this paper. Later in §3 we show how this framework can be used for uncertainty quantification in MT evaluation.

2.1 Desiderata

Let $X \in \mathcal{X}$ and $Y \in \mathcal{Y}$ be random variables representing inputs and outputs, respectively; in this paper we focus on regression, where $\mathcal{Y} = \mathbb{R}$. We use upper case to denote random variables and lower case to denote their specific values.

Traditional machine learning systems use training data to learn predictors $\hat{y} : \mathcal{X} \to \mathcal{Y}$ which, when given a new test input x_{test}, output a point estimate $\hat{y}(x_{\text{test}})$. However, such point estimates lack uncertainty information. Conformal predictors (Vovk et al., 2005) depart from this framework by considering *set* functions $\mathcal{C} : \mathcal{X} \to 2^{\mathcal{Y}}$—given x_{test}, they return a **prediction set** $\mathcal{C}(x_{\text{test}}) \subseteq \mathcal{Y}$ with theoretically established guarantees regarding the coverage of the ground truth value. For regression tasks, this prediction set is usually a confidence interval (see Figure 1). Conformal prediction techniques have recently proved useful in many applications: for example, in the U.S. presidential election in 2020, the *Washington Post* used conformal prediction to estimate the number of outstanding votes (Cherian and Bronner, 2020).

Given a desired confidence level (e.g., 90%), these methods have a formal guarantee that, in expectation, $\mathcal{C}(X_{\text{test}})$ contains the true value Y_{test} with a probability equal to or higher than (but close to) that confidence level. Importantly, this is done in a *distribution-free* manner, i.e., without making any assumptions about the data distribution beyond **exchangeability**, a weaker assumption than independent and identically distributed (i.i.d.) data.[2]

In this paper, we use a simple inductive method called *split* conformal prediction (Papadopoulos, 2008), which requires the following ingredients:

- A mechanism to obtain **non-conformity scores** $s(x, y)$ for each instance, i.e., a way to estimate how ''unexpected'' an instance is with respect to the rest of the data. In this work, we do this by leveraging a pre-trained predictor $\hat{y}(x)$ together with *some* heuristic notion of uncertainty—our method is completely agnostic about which model is used for this. We describe in §2.2 the non-conformity scores we design in our work.
- A held-out **calibration set** containing n examples, $\mathcal{S}^{\text{cal}} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$. The underlying distribution from which the calibration set is generated is assumed unknown but it must be exchangeable (see footnote 2).
- A desired **error rate** α (e.g., $\alpha = 0.1$), such that the coverage level will be $1 - \alpha$ (e.g., 90%).

These ingredients are used to generate prediction sets for new test inputs. Specifically, let $(s_1, \ldots, s_n)$ be the non-conformity scores of each example in the calibration set, i.e., $s_i := s(x_i, y_i)$. Define $\hat{q}$ as the $\lceil (n+1)(1-\alpha) \rceil / n$ **empirical quantile** of these non-conformity scores, where $\lceil \cdot \rceil$ is the ceiling function. This quantile can be easily obtained by sorting the n non-conformity scores and examining the tail of the sequence. Then, for a new test input x_{test}, we output the prediction set

$$\mathcal{C}_{\hat{q}}(x_{\text{test}}) = \{y \in \mathcal{Y} : s(x_{\text{test}}, y) \leq \hat{q}\}. \quad (1)$$

We say that **coverage** holds if the true output y_{test} lies in the prediction set, i.e., if $y_{\text{test}} \in \mathcal{C}_{\hat{q}}(x_{\text{test}})$. This simple procedure has the following theoretical coverage guarantee:

Theorem 1. (Vovk et al. 1999, 2005) *Using the above quantities, the following bounds hold:*

$$\mathbb{P}\big(Y_{\text{test}} \in \mathcal{C}_{\hat{q}}(X_{\text{test}})\big) \in \left[1 - \alpha,\ 1 - \alpha + \frac{1}{n+1}\right].$$

This result tells us two important things: (i) the expected coverage is *at least* $1 - \alpha$, and (ii) with a large enough calibration set (large n), the procedure outlined above does not overestimate the coverage too much, so we can expect it to be *nearly* $1 - \alpha$.[3]

2.2 Non-conformity Scores

Naturally, the result stated in Theorem 1 is only practically useful if the prediction sets $\mathcal{C}_{\hat{q}}(X_{\text{test}})$

[2] Namely, the data distribution is said to be *exchangeable* iff, for any sample $(X_i, Y_i)_{i=1}^n$ and any permutation function π, we have $\mathbb{P}((X_{\pi(1)}, Y_{\pi(1)}), \ldots, (X_{\pi(n)}, Y_{\pi(n)})) = \mathbb{P}((X_1, Y_1), \ldots, (X_n, Y_n))$. If the data distribution is i.i.d., then it is automatically exchangeable, since $\mathbb{P}((X_1, Y_1), \ldots, (X_n, Y_n)) = \prod_{i=1}^n \mathbb{P}(X_i, Y_i)$ and multiplication is commutative. By De Finetti's theorem (De Finetti, 1929), exchangeable observations are conditionally independent relative to some latent variable.

[3] For most purposes, a reasonable size for the calibration set is $n \approx 1000$. See Angelopoulos and Bates (2021, § 3.2).

are small enough to be informative—to ensure this, we need a good heuristic to generate the non-conformity scores $s(x, y)$. In this paper, we are concerned with regression problems ($\mathcal{Y} = \mathbb{R}$), so we define the prediction sets to be **confidence intervals**. We assume we have a pretrained regressor $\hat{y}(x)$, and we consider two scenarios, one where we generate *symmetric* intervals (i.e., where $\hat{y}(x)$ is the midpoint of the interval) and a more general scenario where intervals can be *non-symmetric*.

Symmetric Intervals. In this simpler scenario, we assume that, along with $\hat{y}(x)$, we have a corresponding uncertainty heuristic $\delta(x)$, where higher $\delta(x)$ values signify higher uncertainty. An example—to be elaborated upon in §3.1.1—is where $\delta(x)$ is the quantile of a symmetric probability density, such as a Gaussian, which can be computed analytically from the variance. We then define the non-conformity scores as

$$s(x, y) = \frac{|y - \hat{y}(x)|}{\delta(x)} \tag{2}$$

and follow the procedure above to obtain the quantile $\hat{q}$ from the calibration set. Then, for a random test point $(X_{\text{test}}, Y_{\text{test}})$ and from (1) and (2), we have:

$$\mathbb{P}\big[|Y_{\text{test}} - \hat{y}(X_{\text{test}})| \leq \delta(X_{\text{test}})\hat{q}\big] \gtrsim 1 - \alpha, \tag{3}$$

which corresponds to the confidence interval

$$\mathcal{C}_{\hat{q}}(x) = \big[\hat{y}(x) - \hat{q}\delta(x),\ \hat{y}(x) + \hat{q}\delta(x)\big]. \tag{4}$$

We examine this procedure in §3.1.1 with various uncertainty heuristics (Monte Carlo dropout, deep ensembles, heteroscedastic regression, and direct uncertainty prediction estimates).

Non-symmetric Intervals. Sometimes, better heuristics can be obtained which are non-symmetric, i.e., where there is larger uncertainty in one of the sides of the interval—we will see a concrete example in §3.1.2 where we describe a non-parametric quantile regression procedure (although this might happen as well with parametric heuristics based on fitting non-symmetric distributions, such as the skewed beta distribution). In this case, we assume left and right uncertainty estimates δ_- and δ_+, both positive and satisfying $\delta_- \leq \delta_+$, and define the non-conformity scores as:

$$s(x, y) = \begin{cases} \frac{y - \hat{y}(x)}{\delta_+(x)} & \text{if } y \geq \hat{y}(x) \\ \frac{\hat{y}(x) - y}{\delta_-(x)} & \text{if } y < \hat{y}(x). \end{cases} \tag{5}$$

This leads to prediction sets

$$C_{\hat{q}}(x) = \big[\hat{y}(x) - \hat{q}\delta_-(x),\ \hat{y}(x) + \hat{q}\delta_+(x)\big], \tag{6}$$

which also satisfy Theorem 1. Naturally, when $\delta_- = \delta_+ := \delta$, this procedure recovers the symmetric case.

3 Conformal MT Evaluation

We now apply the machinery of conformal prediction to the problem of MT evaluation, which is a regression task, aiming to predict a numeric quality score over an (automatically) translated sentence. The input is a triplet of source segment s, automatic translation t, and (optionally) human reference r, $x := \langle s, t, r \rangle$, and the goal is to predict a scalar value $\hat{y}(x)$ that corresponds to the estimated quality of the translation t. We can also consider a reference-less MT evaluation scenario where the input is simply $x := \langle s, t \rangle$.[4] The ground truth is a quality score y manually produced by a human annotator, either in the form of a point on a quality scale called *direct assessment* (DA; Graham (2013)) or in the form of accumulated penalties called multidimensional quality metrics (MQM; Lommel et al., 2014). We use DA scores that are standardized for each annotator. An example instance is shown in Figure 3.

Subsequently, to apply conformal prediction we need to determine suitable non-conformity metrics, that can capture the divergence of a new test point x_{test} with respect to the seen data. To that end, we primarily experiment with a range of uncertainty quantification heuristics to generate $\delta(x)$ (or $\delta_-(x)$ and $\delta_+(x)$ in the non-symmetric case). With the symmetric parametric uncertainty methods, described in §3.1.1, we obtain heuristics to compute δ which we use to obtain non-conformity scores via (2), leading to the confidence intervals $\mathcal{C}_{\hat{q}}(x)$ in (4), for each x triplet. Alternatively, in §3.1.2 we describe a non-symmetric and non-parametric method which returns $\hat{y}$, δ_-, and δ_+, and which we will use to

[4] The reference-less scenario is also frequently referred to as quality estimation for machine translation.

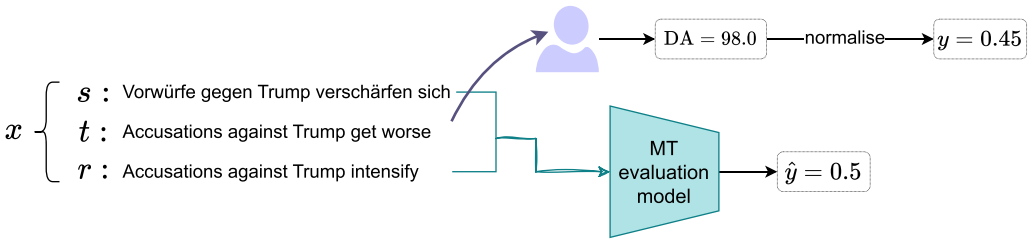


Figure 3: Example of MT evaluation instance: $x := \langle s, t, r \rangle$ input triplet (left) evaluated by human and then normalised to obtain the ground truth scores (top right) and model prediction (bottom right).

compute the non-conformity scores (5) and confidence intervals (6). Finally, in 3.1.3 we describe a heuristic non-conformity score which is inspired by the MT symmetry between s and t.

3.1 Choice of Non-conformity Scores

For the application of conformal prediction on MT evaluation, we experiment with a diverse set of uncertainty prediction methods to obtain non-conformity scores, accounting both for parametric and non-parametric uncertainty prediction. We extensively compare all the parametric methods previously used in MT evaluation (Zerva et al., 2022), which return symmetric confidence intervals. In addition, we experiment with **quantile regression** (Koenker and Hallock, 2001), a simple non-parametric approach that has never been used for MT evaluation (to the best of our knowledge), and which can return non-symmetric intervals. Finally, we propose a new MT evaluation-specific non-conformity measure that relies on the symmetry between source and target in MT tasks.

3.1.1 Parametric Uncertainty

We compare a set of different parametric methods which fit the quality scores in the training data to an input-dependent Gaussian distribution $\mathcal{N}(\hat{\mu}(x), \hat{\sigma}^2(x))$. All these methods lead to **symmetric** confidence intervals (see Eq. 4). We use these methods to obtain estimates $\hat{y}(x) := \hat{\mu}(x)$. Then we use $\hat{\sigma}$ to extract the corresponding uncertainty estimates as $\delta(x) := \text{probit}(1 - \frac{\alpha}{2})\hat{\sigma}$, which correspond to the $\frac{\alpha}{2}$ and $1 - \frac{\alpha}{2}$ quantiles of the Gaussian, for a given confidence threshold $1 - \alpha$. For $\alpha = 0.1$ (i.e., a 90% confidence level) this results in $\delta(x) = 1.64 \times \hat{\sigma}$. We describe the concrete methods used to estimate $\hat{\mu}(x)$ and $\hat{\sigma}(x)$ below.

MC Dropout (MCD). This is a variational inference technique approximating a Bayesian network with a Bernoulli prior distribution over its weights (Gal and Ghahramani, 2016). By retaining dropout layers during multiple inference runs, we can sample from the posterior distribution over the weights. As such, we can approximate the uncertainty over a test instance x through a Gaussian distribution with the empirical mean $\hat{\mu}(x)$ and variance $\hat{\sigma}^2(x)$ of the quality estimates $\{\hat{y}_1, \ldots, \hat{y}_N\}$. We use 100 runs, following the analysis of Glushkova et al. (2021).

Deep Ensembles (DE). This method (Lakshminarayanan et al., 2017) trains an ensemble of neural models with the same architecture but different initializations. During inference, we collect the predictions of each single model and return $\hat{\mu}(x)$ and $\hat{\sigma}^2(x)$ as in MC dropout. We use $N = 5$ checkpoints obtained with different initialization seeds, following Glushkova et al. (2021).

Heteroscedastic Regression (HTS). We follow Le et al. (2005) and Kendall and Gal (2017b) and incorporate $\hat{\sigma}^2(x)$ as part of the training objective. This way, a regressor is trained to output two values: (1) a mean score $\hat{\mu}(x)$ and (2) a variance score $\hat{\sigma}^2(x)$. This predicted mean and variance parameterize a Gaussian distribution $\mathcal{N}(y; \hat{\mu}(x;\theta), \hat{\sigma}^2(x;\theta))$, where θ are the model parameters. The negative log-likelihood loss function is used:

$$\mathcal{L}_{\text{HTS}}(\hat{\mu}, \hat{\sigma}^2; y) = \frac{(y - \hat{\mu})^2}{2\hat{\sigma}^2} + \frac{1}{2}\log \hat{\sigma}^2. \quad (7)$$

This framework is particularly suitable to express aleatoric uncertainty due to heteroscedastic noise, as the framework allows larger variance to be assigned to ''noisy'' examples which will result in down-weighting the squared term in the loss.

Direct Uncertainty Prediction (DUP). This is a two-step procedure which relies on the assumption that the total uncertainty over a test instance

is equivalent to the generalization error of the regression model (Lahlou et al., 2021). A standard regression model $\hat{y}(x)$ is first fit on the training set and then applied to a held-out validation set $\mathcal{S}^{\text{val}}$. Then, a second model is trained on this held-out set to regress on the error $\epsilon = |\hat{y}(x) - y|$ incurred by the first model predictions, approximating its uncertainty. To train the error predicting model, we follow the setup of Zerva et al. (2022), using as inputs the $x^{\text{val}} = \langle s, t, r \rangle$ triplets combined with the predictions $\hat{y}^{\text{val}}$ of the first model, which are used as bottleneck features in an intermediate fusion fashion. The loss function is

$$\mathcal{L}_{\text{DUP}}(\hat{\epsilon}; \epsilon) = \frac{\epsilon^2}{2\hat{\epsilon}^2} + \frac{1}{2}\log(\hat{\epsilon})^2. \quad (8)$$

We use $\hat{\epsilon}(x)$ as the uncertainty heuristic.

3.1.2 Non-parametric Uncertainty: Quantile Regression (QNT)

Quantile regression is a statistical method used to model input-dependent quantiles within a regression framework (Koenker and Bassett Jr, 1978). As opposed to regular (linear) regression that models the mean of a target variable Y conditioned on the input X, quantile regression models a *quantile* of the distribution of Y (e.g., the median, the 95%, or the 5% percentile scores). By definition, quantile regression does not require any parametric assumptions on the distribution of Y and is less sensitive to outliers. Quantiles provide an attractive representation for uncertainty: They allow for easy construction of prediction intervals, at chosen confidence levels. Learning the quantile for a particular quantile level involves optimizing the **pinball loss**, a tilted transformation of the absolute value function (see Figure 4). Given a target y, a prediction $\hat{y}$, and quantile level $\tau \in (0, 1)$, the pinball loss $\mathcal{L}_\tau$ is defined as:

$$\mathcal{L}_\tau(\hat{y}; y) = (\hat{y} - y)(\mathbb{1}\{y \leq \hat{y}\} - \tau). \quad (9)$$

We can select τ to correspond to the error rate α that we want to achieve. Note that for $\tau = 0.5$ the loss function reduces to (half) the mean absolute error loss $\mathcal{L}_{\text{MAE}}(\hat{y}; y) = \frac{1}{2}|\hat{y} - y|$.

We use $\tau = \alpha$ to train our models to predict the $\hat{Q}_{1-\tau/2}$ and $\hat{Q}_{\tau/2}$ quantiles, as well as the $\hat{Q}_{0.5}$ quantile, which corresponds to the median (see below), but there are extensions that either optimize multiple quantiles that cover the full predictive distribution (Tagasovska and Lopez-Paz, 2019) or explore asymmetric loss extensions to account for overestimating or underestimating the confidence intervals (Beck et al., 2016).

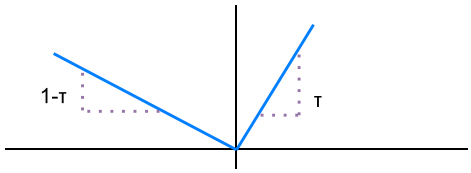


Figure 4: The pinball loss objective used for quantile regression. The slope of the lines is determined by the desired quantile level τ.

Method	Orig.	Calib.	Conform.	$\hat{q}$
MCD	23.82	66.60	90.01	8.08
DE	29.10	66.23	91.31	6.99
HTS	82.02	68.29	89.89	1.28
DUP	86.01	66.13	89.88	1.11
QUANT-NS	77.83	–	90.21	1.29
QUANT-S	78.66	49.03	90.54	1.28

Table 1: Coverage percentages for $\alpha = 0.1$ over different uncertainty methods. Values reported correspond to the mean over 10 runs. The second, third, and fourth columns refer respectively to the coverage obtained by original methods without calibration, after the ECE calibration described in §3.2, and with conformal prediction as described in §2.

Unlike the parametric methods covered in §3.1.1, the quantile regression method can be used to return **asymmetric** confidence intervals. This is done by fitting 0.5, $1 - \frac{\tau}{2}$, and $\frac{\tau}{2}$ quantile predictors to the data, and setting $\hat{y}(x) := \hat{Q}_{0.5}(x)$, $\hat{\delta}_+(x) := \hat{Q}_{1-\frac{\tau}{2}}(x) - \hat{y}(x)$, and $\hat{\delta}_-(x) := \hat{y}(x) - \hat{Q}_{\frac{\tau}{2}}(x)$.

For completeness, we also consider a **symmetric** variant of quantile regression where we do not estimate the median $\hat{Q}_{0.5}(x)$, but rather set $\hat{y}(x) = \frac{1}{2}\big(\hat{Q}_{1-\frac{\tau}{2}}(x) + \hat{Q}_{\frac{\tau}{2}}(x)\big)$. We report coverage for both the non-symmetric (**QNT-NS**) and the symmetric case (**QNT-S**) later in Table 1.

3.1.3 Back-translation-inspired Non-conformity

The aforementioned uncertainty quantification metrics are based on well-established methods that could be applied on other regression problems with minimal modifications. However, conformal prediction is quite flexible with respect to the choice of the underlying non-conformity measure,

allowing us to tailor the definition of conformity to the task at hand. Thus, we also experiment with a back-translation-inspired setup for the referenceless COMET metric (COMET-QE).

Our intuition for this metric is previous work that exploiting the symmetry between source and target, e.g. via back translation, can be used as an indicator of translation quality (Agrawal et al., 2022; Moon et al., 2020). In other words, a metric that computes the distance (on semantic or surface level) between the original source sentence and the one obtained upon translating the target, correlates well with translation quality. In this work, we hypothesize that we can exploit the symmetry between translation directions to infer a non-conformity measure as follows:

- We compute $\hat{y}$ as described in §3 for $x := \langle s, t \rangle$.
- We compute $\hat{y}_{\mathrm{T}}$ for inverted inputs $x := \langle t, s \rangle$.
- We finally compute the non-conformity score s as in Eq. 2 with $\delta(x) = |\hat{y} - \hat{y}_{\mathrm{T}}|$.

We henceforth refer to this score as the BT non-conformity score.

3.2 Comparison with Calibration

We compare the coverage obtained by our proposed ''conformalized'' uncertainty scores with that of a vanilla calibration approach that minimizes the **expected calibration error** (ECE; Naeini et al. 2015; Kuleshov et al. 2018b). ECE has been proposed as a measure of how well aligned the model confidence is with the model accuracy, based on the simple desideratum that a model with, e.g., 80% confidence over a set of examples should achieve an accuracy of 80% over the same examples to be well-calibrated. It is defined as

$$\mathrm{ECE} = \frac{1}{M} \sum_{b=1}^{M} |\mathrm{acc}(\gamma_b) - \gamma_b|, \qquad (10)$$

where each b is a bin representing a confidence level γ_b, and $\mathrm{acc}(\gamma_b)$ is the fraction of times the ground truth y falls inside the confidence interval associated to that bin. Several variants of uncertainty calibration have been proposed to correct unreliable uncertainty estimates that do not correlate with model accuracy (Kuleshov et al., 2018b; Amini et al., 2020; Levi et al., 2022). We follow Glushkova et al. (2021) who find that computing a simple affine transformation of the original uncertainty distribution that minimises the ECE is effective to quantify uncertainty in MT evaluation.

3.3 Experimental Setup

Models. We experiment with a range of different models for the task of MT quality evaluation. We specifically use two models that employ source, translation, and reference in their input, namely UniTE (Wan et al., 2022) and COMET (Rei et al., 2020). We also experiment with BLEURT (Sellam et al., 2020), a metric that relies only on translation and reference comparisons, and finally, we explore a reference-less setup using COMET-QE, receiving only the source and translation sentences as input (Rei et al., 2021; Zerva et al., 2021). We provide model training hyperparameters in Appendix B.

Data. For training, we use the direct assessment (DA) data from the WMT17-19 metrics shared tasks (Ma et al., 2018, 2019). We evaluate our models on the WMT20 metrics dataset (Mathur et al., 2020). For the calibration set $\mathcal{S}^{\mathrm{cal}}$, we use repeated random sub-sampling for $k = 20$ runs. The WMT20 test data includes 16 language pairs, of which 9 pairs are into-English and 7 pairs are out-of-English translations. For the calibration set sub-sampling, we sample uniformly from each language pair. For metrics for which we report averaged performance, we use micro-average over all of the language pairs.

3.4 Results

We first compare the uncertainty methods described in §3.1 with respect to coverage percentage as shown in Table 1. We select a desired coverage level of 90%, i.e., we set $\alpha = 0.1$. We also align the uncertainty estimates with respect to the same α value: for the parametric uncertainty heuristics, we select the $\delta(x)$ that corresponds to a $1 - \alpha$ coverage of the distribution, by using the probit function as described in §3.1.1; and for the non-parametric approach, we train the quantile regressors by setting $\tau = \alpha/2$, as described in §3.1.2.

Table 1 shows that coverage varies significantly across methods for the COMET metric,

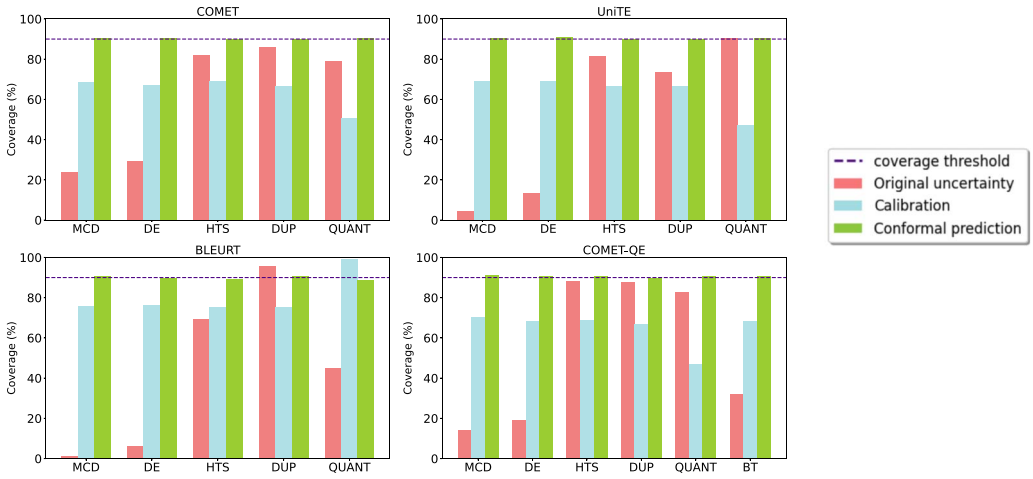


Figure 5: Coverage obtained by different uncertainty predictors for different MT evaluation metrics. We compare originally obtained values (red), with values after calibration (light blue) and after conformal prediction (green) with the desired coverage threshold (dashed line) set to 0.9 (90%).

while Figure 5 shows that the same trend follows the coverage for BLEURT, UniTE, and COMET-QE metrics respectively. We can see that the sampling-based methods such as MC dropout and deep ensembles achieve coverage much below the desired $1-\alpha$ level. In contrast, direct uncertainty prediction and heteroscedastic regression achieve comparatively high coverage even before the application of conformal prediction. This could be related to the fact that by definition, they try to model uncertainty in relation to model error (DUP explicitly tries to predict uncertainty modelled as $\epsilon = |\hat{y} - y|$, while based on Eq. 7, the model needs to predict larger variance for larger errors). The quantile regression method also performs competitively to DUP achieving high coverage across metrics, with the exception of BLEURT (the only metric that does not use the source sentence), where coverage is significantly lower to DUP and HTS. Finally, while the back-translation-inspired (BT) score does not achieve high coverage, it outperforms sampling-based methods, providing a low-cost solution even in the absence of trained uncertainty quantifiers.

Calibration helps improve coverage in the cases of MC dropout and deep ensembles—albeit still without reaching close to 0.9. Instead, it seems that minimizing the ECE is not well aligned to optimizing coverage as for most cases calibration leads to less than 70% coverage. In contrast, we can see that conformal prediction approximates the desired coverage level best for all methods, regardless of the initial coverage they obtain, in line with the guarantees provided by Theorem 1.

		MCD	DE	HTS	DUP	QNT	BT
COMET	q	8.08	6.99	1.29	1.11	1.28	–
	UPS	0.04	0.07	0.24	0.27	0.34	–
BLEURT	q	38.72	13.92	1.31	0.89	1.23	–
	UPS	0.29	0.18	0.28	0.33	0.33	–
UniTE	q	45.50	18.02	1.36	1.66	1.56	–
	UPS	0.02	0.04	0.17	0.23	0.35	–
COM-QE	q	15.00	11.75	1.12	0.09	1.28	11.77
	UPS	−0.11	0.02	0.23	0.29	0.20	0.13

Table 2: Conformal prediction quantiles $\hat{q}$ versus UPS correlation coefficients over the test set, for each metric.

In addition, as shown in Table 2 the $\hat{q}$ value seems to correlate well with the performance of each uncertainty quantification metric, as measured by *uncertainty Pearson correlation* (UPS) (Glushkova et al., 2021).[5] We can specifically see that methods with low $\hat{q}$ correspond to uncertainty quantification methods that yield better performance and correlate better with the residuals of the MT evaluation metric. Hence, conformal prediction can be used to efficiently guide the selection

[5]Note that unlike (Glushkova et al., 2021) we compute UPS over the full test set, instead of taking the macro-average over each language-pair. However, looking at the values reported in that work we can see that our findings hold for the macro-averaged UPS values as well.

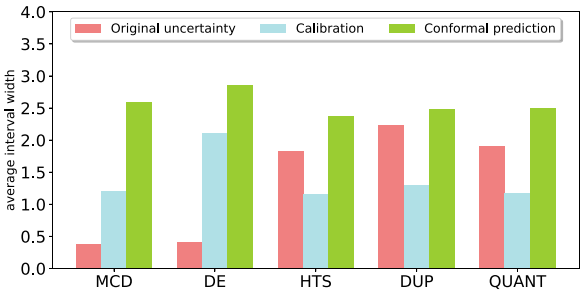


Figure 6: Width for each uncertainty quantifier for the COMET metric showing the original intervals (red), the intervals after calibration (light blue) and the intervals after conformal prediction (green).

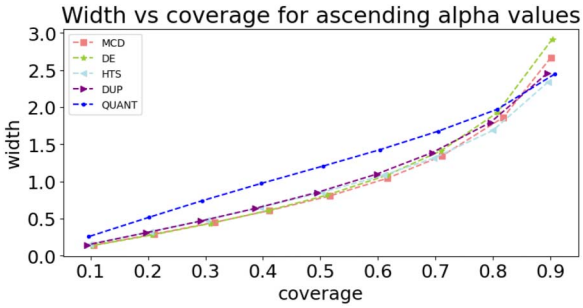


Figure 7: Coverage (x-axis) vs width (y-axis) for increasing α values, for the COMET metric.

of a suitable uncertainty quantification method, using only a small amount of data (calibration set).

Besides measuring the coverage achieved by the several methods, it is also important to examine how wide the predicted intervals are. To that end, we also compute the confidence interval **width** as the average width of the predicted confidence intervals after the conformal prediction application (Kuleshov and Deshpande, 2022),[6]

$$\text{avg.width} = \frac{1}{|\mathcal{S}^{\text{test}}|} \sum_{x \in \mathcal{S}^{\text{test}}} |\mathcal{C}_{\hat{q}}(x)|. \qquad (11)$$

We show the width in Figure 6, where we can see that especially for MCD and DE methods the average width increases significantly to reach the desired 90% coverage. The direct uncertainty prediction method is the one that shows a smaller increase in width, with quantile and heteroscedastic regression following.

[6]In related work (Glushkova et al., 2021), sharpness is computed with respect to σ^2, but this cannot be applied to non-parametric uncertainty cases, so we use the confidence interval length, henceforth referred to as *width* to be able to compare conformal prediction for all uncertainty quantification methods.

We also plot the average width of the conformalised confidence intervals with respect to coverage for increasing α values (see Figure 7). We can see that as the desideratum on coverage relaxes confidence interval widths reduce accordingly, and that depending on the chosen α value the optimal method can vary. For example, quantile regression performs much better for $\alpha \leq 0.2$ but for more ‘‘relaxed" values the width-coverage balance deteriorates.

4 Conditional Coverage

The coverage guarantees stated in Theorem 1 refer to *marginal* coverage—the probabilities are not conditioned on the input points, they are averaged (marginalized) over the full test set. In several practical situations it is desirable to assess the **conditonal** coverage $\mathbb{P}[Y_{\text{test}} \in \mathcal{C}(X_{\text{test}}) \mid X_{\text{test}} \in \mathcal{G}]$ where $\mathcal{G} \subseteq \mathcal{X}$ denotes a region of the input space, e.g., inputs containing some specific attributes or pertaining to some group of the population.

In fact, evaluating the conditional coverage with respect to different data attributes may reveal biases of the uncertainty estimation methods towards specific data subgroups which are missed if we only consider marginal coverage. In the next experiments, we follow the feature stratified coverage described in Angelopoulos and Bates (2021); we use conformal prediction with MC dropout as our main paradigm. We demonstrate five examples of imbalanced coverage in Figure 8 and Table 3 with respect to different attributes: language pairs, estimated source difficulty, and predicted quality and uncertainty scores.

We can see that, coverage varies significantly across groups, revealing biases towards specific attribute values. For example, the plots show that into-English translations are under-covered for most uncertainty quantifiers (coverage $\leq$ 0.9), i.e., we consistently underestimate the uncertainty over the predicted quality for these language pairs. More importantly, we can see that examples with low predicted quality are significantly under-covered, as coverage for quality scores where $y \leq -1.5$ drops below 50%. For MCD-based uncertainty scores on the other hand, the drop in coverage seems to be related to the low uncertainty scores, indicating that due to the skewed distribution of uncertainty scores, the calculation of the $\hat{q}$ quantile is not well tuned to lower uncertainty values (i.e., higher non-conformity scores). Instead,

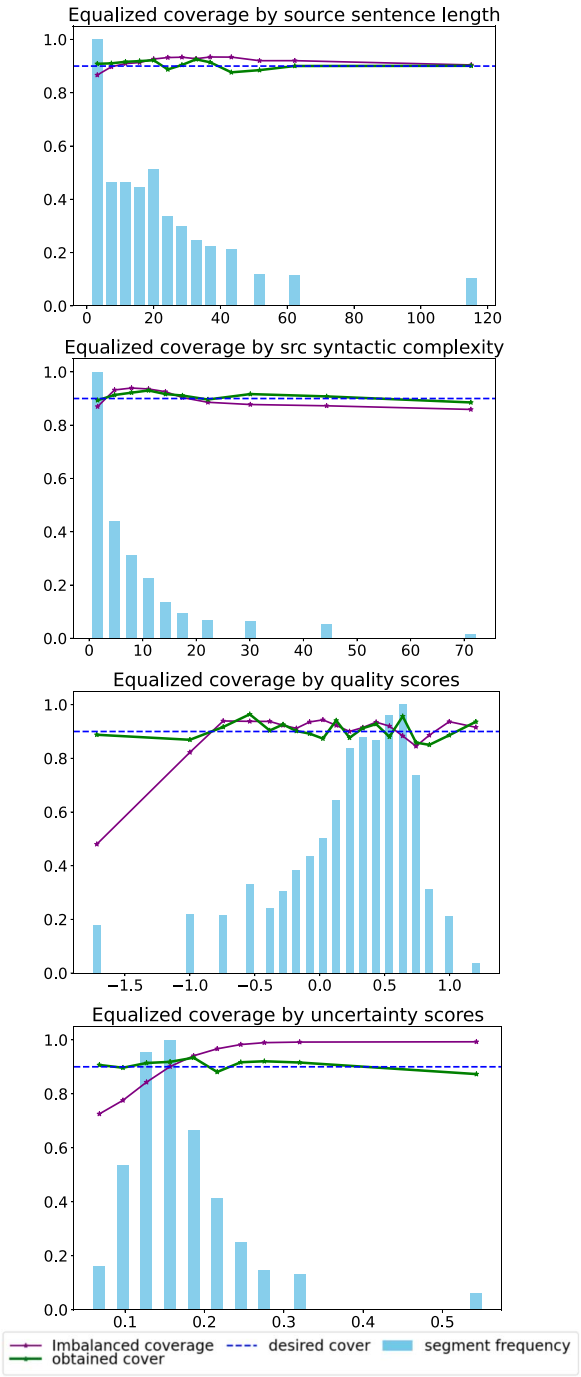


Figure 8: Conditional coverage imbalance per (top-to-bottom): sentence length, syntactic complexity, estimated quality score level and uncertainty score for conformal prediction with MCD-based non-conformity scores. To facilitate plotting, the segment frequencies are re-scaled with respect to the maximum bin frequency (so that the bin with the maximum frequency equals 1).

our two proxies for source difficulty reveal better balanced behaviour, with small deviations for very small sentences or high syntactic complexity. Similar patterns for these dimensions are also observed for the other uncertainty quantification methods shown in Appendix C.

	QNT	MCD	DE	HTS	DUP
En-Cs	0.982	0.959	0.939	0.875	0.931
En-De	0.973	0.971	0.925	0.863	0.927
En-Ja	0.990	0.978	0.987	0.886	0.972
En-Pl	0.977	0.948	0.914	0.882	0.914
En-Ru	0.974	0.958	0.936	0.862	0.926
En-Ta	0.970	0.952	0.949	0.892	0.858
En-Zh	0.934	0.983	0.991	0.919	0.945
Cs-En	0.890	0.871	0.884	0.898	0.875
De-En	0.880	0.888	0.867	0.896	0.902
Ja-En	0.883	0.856	0.921	0.910	0.887
Kn-En	0.881	0.875	0.948	0.943	0.840
Pl-En	0.862	0.833	0.825	0.873	0.849
Ps-En	0.851	0.854	0.932	0.922	0.786
Ru-En	0.851	0.828	0.831	0.879	0.888
Ta-En	0.793	0.809	0.878	0.898	0.883
Zh-En	0.861	0.833	0.868	0.886	0.827
En-Cs	0.893	0.917	0.888	0.892	0.902
En-De	0.902	0.902	0.902	0.896	0.893
En-Ja	0.909	0.891	0.900	0.891	0.904
En-Pl	0.882	0.905	0.895	0.900	0.898
En-Ru	0.900	0.898	0.908	0.906	0.903
En-Ta	0.903	0.895	0.883	0.886	0.903
En-Zh	0.880	0.890	0.884	0.896	0.896
Cs-En	0.890	0.917	0.909	0.904	0.894
De-En	0.897	0.901	0.901	0.897	0.903
Ja-En	0.900	0.912	0.899	0.894	0.902
Kn-En	0.896	0.903	0.902	0.904	0.894
Pl-En	0.900	0.905	0.893	0.894	0.877
Ps-En	0.905	0.899	0.900	0.884	0.907
Ru-En	0.910	0.896	0.907	0.900	0.900
Ta-En	0.884	0.901	0.886	0.901	0.908
Zh-En	0.900	0.910	0.908	0.900	0.905

Table 3: Conditional coverage over different language pairs of WMT 2020 DA data, before (top) and after (bottom) balanced conformal prediction. Red coloured entries signify coverage < 0.9.

Ensuring that we do not overestimate confidence for such examples is crucial for MT evaluation, in particular for applications where MT is used on the fly and one needs to decide if human editing is needed. Hence, in the rest of this section, we elaborate approaches to assess and mitigate coverage imbalance in the aforementioned examples, towards **equalized coverage** (Romano et al., 2020).

4.1 Conditioning on Categorical Attributes: Language-pairs

To deal with imbalanced coverage for discrete data attributes we use an equalized conformal prediction approach, i.e., we compute the conditional coverage for each attribute value and, upon observing imbalances, **we compute conditional quantiles instead of a single one** on the calibration set.

Let $\{1, \ldots, K\}$ index the several attributes (e.g., language pairs). We partition the calibration set according to these attributes, $\mathcal{S}^{\text{cal}} = \bigcup_{k=1}^{K} \mathcal{S}_k^{\text{cal}}$, where $\mathcal{S}_k^{\text{cal}}$ denotes the partition corresponding to the k^{th} attribute and $\mathcal{S}_k^{\text{cal}} \cap \mathcal{S}_{k'}^{\text{cal}} = \varnothing$ for every $k \neq k'$. Then, we follow the procedure described in §2 to fit attribute-specific quantiles $\hat{q}_k$ to each calibration set $\mathcal{S}_k^{\text{cal}}$.

We demonstrate the application of this process on language pairs in Table 3 for all uncertainty quantification methods examined in the previous section. The top part of Table 3 shows the language-based conditional coverage, using a heatmap coloring to highlight the language pairs that fall below the guaranteed marginal coverage of $1 - \alpha = 0.9$. We can see that for all language pairs we achieve coverage >75% but some are below the 90% target. For all methods except for DUP, the coverage is high for out-of-English translations and drops for the majority of into-English cases. Applying the equalizing approach described above, we successfully rectify the imbalance for all uncertainty quantification methods, as shown in the bottom heatmap of Table 3.

4.2 Conditioning on Numerical Attributes: Quality, Difficulty and Uncertainty Scores

With some additional constraints on the equalized conformal prediction process described in §4.1 we can generalize this approach to account for attributes with numerical discrete or continuous values, such as the MT quality scores (ground truth quality y) or the uncertainty scores obtained by different uncertainty quantification methods. To that end, we adapt the **Mondrian conformal regression** methodology (Vovk et al., 2005; Boström et al., 2021). Mondrian conformal predictors have been used initially for classification and later for regression, where they have been used to partition the data with respect to the residuals $|y - \hat{y}(x)|$ (Johansson et al., 2014; Boström et al., 2021). Boström and Johansson (2020) proposed a Mondrian conformal predictor that partitions along the expected ''difficulty'' of the data as estimated by the non-conformity score $s(x, y)$ or the uncertainty score $\delta(x)$.

In all the above cases, the calibration instances are sorted according to a continuous variable of interest and then partitioned into calibration bins. While the bins do not need to be of equal size, they need to satisfy a minimum length condition that depends on the chosen α threshold for the error rate (Johansson et al., 2014). Upon obtaining a partition into calibration bins, and similarly to what was described in §4.1 for discrete attributes, we compute bin-specific quantiles $\hat{q}_b$, where $b \in \{1, \ldots, B\}$ indexes a bin.

We apply the aforementioned approach to the MT evaluation for the estimated translation quality scores, $\hat{y}$, and uncertainty scores, as well as two different proxies for sentence translation difficulty, namely sentence length and syntactic complexity (computed on the source language) (Mishra et al., 2013). We compute the **source sentence length**, as the number of tokens in the sentence, while for syntactic complexity we consider the sum of subtrees that constitute grammatical phrases[7] and sort the calibration and test samples accordingly.

We then split the ordered calibration set into bins[8] and compute the quantiles over the calibration set bins. Subsequently, to apply the conformal prediction on a test instance x_{test}, we check the attribute value and identify which bin $\hat{b}$ of the S_{cal} set it falls into, to use the corresponding quantile $\hat{q}_{\hat{b}}$.

The equalized coverage for COMET-MCD is shown in Figure 8, compared with the original coverage. We can see that for estimated quality and uncertainty scores, the previously observed coverage drop for the lower values is successfully rectified by the equalized conformal prediction approach, achieving balanced coverage across bins, as desired. Between the two difficulty approximation methods, we see that for MCD the obtained bins are fairly balanced concerning coverage, with only a small drop for higher difficulty in terms of syntactic complexity. We provide additional results for the remaining uncertainty quantifiers in Appendix C.

[7] We employ an `nltk`-based dependency parser.

[8] We use a threshold of 100 instances per bin.

5 Related Work

5.1 Conformal Prediction

We build on literature on conformal prediction that has been established by Vovk et al. (2005). Subsequent works focus on improving the predictive efficiency of the conformal sets or relaxing some of the constraints (Angelopoulos and Bates, 2021; Jin and Candès, 2022; Tibshirani et al., 2019). Most relevant to our paper are works that touch conformal prediction for regression tasks, either via the use of quantile regression (Romano et al., 2019) or using other scalar uncertainty estimates (Angelopoulos and Bates, 2021; Johansson et al., 2014; Papadopoulos et al., 2011). Other strands of work focus on conditional conformal prediction and methods to achieve balanced coverage (Angelopoulos and Bates, 2021; Romano et al., 2020; Boström et al., 2021; Lu et al., 2022).

There are few studies that use conformal prediction in NLP, so far focusing only on classification or generation, with applications to sentiment and relation classification and entity detection (Fisch et al., 2021, 2022; Maltoudoglou et al., 2020). Recently, Ravfogel et al. (2023) and Ulmer et al. (2024) considered natural language generation, with the former proposing the use of conformal prediction applied to top-p nucleus sampling, and the latter proposing applying non-exchangeable conformal prediction with k-nearest neighbors to obtain better prediction sets for generation. Some other works apply conformal prediction on the sentence level to rank generated sentences for different tasks (Kumar et al., 2023; Ren et al., 2023; Liang et al., 2024). Concurrent to this work, Giovannotti (2023) proposed the use of conformal prediction to quantify MT quality estimation, using a k-nearest neighbor (kNN) quality estimation model to obtain non-conformity scores, proposing the use of conformal prediction as a new standalone uncertainty quantification method for this task. They empirically demonstrate the impact of violating the i.i.d. assumption on the obtained performance and compare to a fixed-variance baseline regarding ECE, AUROC, and sharpness, but they consider neither the aspect of marginal or conditional coverage for the estimated confidence intervals, nor any other uncertainty quantification methods.

Our work complements the aforementioned efforts, as it focuses on a regression task (MT evaluation) and investigates the impact of conformal prediction on the estimated confidence intervals. Contrary to previous approaches, however, we provide a detailed analysis of conformal prediction for an NLP regression task and investigate a wide range of uncertainty methods that can be used to design non-conformity scores. Additionally, we elaborate different aspects of equalized coverage for MT evaluation, revealing biases for different data attributes, and providing an effective method that corrects these biases.

5.2 Uncertainty Quantification

Several uncertainty methods have been previously proposed for regression tasks in NLP and the task of MT evaluation specifically. Beck et al. (2016) focused on the use of Gaussian processes to obtain uncertainty predictions for the task of quality estimation, with emphasis on cases of asymmetric risk. Wang et al. (2022) also explored Gaussian processes but provided a comparison of multiple NLP regression tasks (semantic sentence similarity, MT evaluation, sentiment quantification) investigating end-to-end and pipeline approaches to apply Bayesian regression to large language models. Focusing on MT evaluation, Glushkova et al. (2021) proposed the use of MC dropout and deep ensembles as efficient approximations of Bayesian regression, inspired by work in computer vision (Kendall and Gal, 2017a). Zerva et al. (2022) proposed additional methods of uncertainty quantification for MT evaluation, focusing on methods that targ et aleatoric or epistemic uncertainties under specific assumptions. They specifically investigated heteroscedastic regression and KL-divergence for aleatoric uncertainty and direct uncertainty prediction for epistemic uncertainty, highlighting the performance benefits of these methods, when compared to MC dropout and deep ensembles, with respect to the correlation of uncertainties to model error. However, none of the previous works in uncertainty for NLP regression considered coverage. We compare several of the aforementioned uncertainty quantification methods with respect to coverage and focus on the impact of applying conformal prediction to each uncertainty method.

6 Conclusions

In this work, we apply conformal prediction to the important problem of MT evaluation. We

show that most existing uncertainty quantification methods significantly underestimate uncertainty, achieving low coverage, and that the application of conformal prediction can help rectify this and guarantee coverage tuned to a user-specified threshold. We further show that the estimated quantiles provide a way to choose the most suitable uncertainty quantification methods, aligning well with other metrics such as UPS (Glushkova et al., 2021).

We also use conformal prediction tools to assess the conditional coverage for five different attributes: language pairs, sentence length and syntactic complexity, predicted translation quality, and estimated uncertainty level. We highlight inconsistencies and imbalanced coverage for the different cases, and we show that equalized conformal prediction can correct the initially unfair confidence predictions to obtain more balanced coverage across attributes.

Overall, our work aims to highlight the potential weaknesses of using uncertainty estimation methods without a principled calibration procedure. To this end, we propose a methodology that can guarantee more meaningful confidence intervals. In future work, we aim to further investigate the application of conformal prediction across different data dimensions as well as different regression tasks in NLP.

Acknowledgments

This work was supported by the Portuguese Recovery and Resilience Plan through project C645008882- 00000055 (NextGenAI - Center for Responsible AI), by the EU's Horizon Europe Research and Innovation Actions (UTTER, contract 101070631), by the project DECOLLAGE (ERC-2022-CoG 101088763), and by Fundação para a Ciência e Tecnologia through contract UIDB/50008/2020.

References

Sweta Agrawal, Nikita Mehandru, Niloufar Salehi, and Marine Carpuat. 2022. Quality estimation via backtranslation at the wmt 2022 quality estimation task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 593–596.

Alexander Amini, Wilko Schwarting, Ava Soleimany, and Daniela Rus. 2020. Deep evidential regression. *Advances in Neural Information Processing Systems*, 33:14927–14937.

Anastasios N. Angelopoulos and Stephen Bates. 2021. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*.

Daniel Beck, Lucia Specia, and Trevor Cohn. 2016. Exploring prediction uncertainty in machine translation quality estimation. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 208–218. https://doi.org/10.18653/v1/K16-1021

Henrik Boström and Ulf Johansson. 2020. Mondrian conformal regressors. In *Conformal and Probabilistic Prediction and Applications*, pages 114–133. PMLR.

Henrik Boström, Ulf Johansson, and Tuwe Löfström. 2021. Mondrian conformal predictive distributions. In *Conformal and Probabilistic Prediction and Applications*, pages 24–38. PMLR.

John Cherian and Lenny Bronner. 2020. How the washington post estimates outstanding votes for the 2020 presidential election.

John J. Cherian and Emmanuel J. Candès. 2023. Statistical inference for fairness auditing. *arXiv preprint arXiv:2305.03712*.

Bruno De Finetti. 1929. Funzione caratteristica di un fenomeno aleatorio. In *Atti del Congresso Internazionale dei Matematici: Bologna del 3 al 10 de settembre di 1928*, pages 179–190.

Yukun Ding, Jinglan Liu, Jinjun Xiong, and Yiyu Shi. 2020. Revisiting the evaluation of uncertainty estimation and its application to explore model complexity-uncertainty trade-off. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. https://doi.org/10.1109/CVPRW50498.2020.00010

Adam Fisch, Tal Schuster, Tommi Jaakkola, and Dr. Regina Barzilay. 2021. Few-shot conformal prediction with auxiliary tasks. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 3329–3339. PMLR.

Adam Fisch, Tal Schuster, Tommi Jaakkola, and Regina Barzilay. 2022. Conformal prediction

sets with limited false positives. In *International Conference on Machine Learning*, pages 6514–6532. PMLR.

Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning*, pages 1050–1059. PMLR.

Patrizio Giovannotti. 2023. Evaluating machine translation quality with conformal predictive distributions.

Taisiya Glushkova, Chrysoula Zerva, Ricardo Rei, and André F. T. Martins. 2021. Uncertainty-aware machine translation evaluation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3920–3938. https://doi.org/10.18653/v1/2021.findings-emnlp.330

Yvette Graham. 2013. Continuous measurement scales in human evaluation of machine translation. Association for Computational Linguistics.

Pavel Izmailov, Patrick Nicholson, Sanae Lotfi, and Andrew G. Wilson. 2021. Dangers of Bayesian model averaging under covariate shift. *Advances in Neural Information Processing Systems*, 34:3309–3322.

Ying Jin and Emmanuel J. Candès. 2022. Selection by prediction with conformal p-values. *arXiv preprint arXiv:2210.01408*.

Ulf Johansson, Cecilia Sönströd, Henrik Linusson, and Henrik Boström. 2014. Regression trees for streaming data with local performance guarantees. In *2014 IEEE International Conference on Big Data (Big Data)*, pages 461–470. IEEE. https://doi.org/10.1109/BigData.2014.7004263

Alex Kendall and Yarin Gal. 2017a. What uncertainties do we need in Bayesian deep learning for computer vision? *Advances in Neural Information Processing Systems*, 30.

Alex Kendall and Yarin Gal. 2017b. What uncertainties do we need in Bayesian deep learning for computer vision? In *NIPS*.

Roger Koenker and Gilbert Bassett Jr. 1978. Regression quantiles. *Econometrica: Journal of the Econometric Society*, pages 33–50. https://doi.org/10.2307/1913643

Roger Koenker and Kevin F. Hallock. 2001. Quantile regression. *Journal of Economic Perspectives*, 15(4):143–156. https://doi.org/10.1257/jep.15.4.143

Volodymyr Kuleshov and Shachi Deshpande. 2022. Calibrated and sharp uncertainties in deep learning via density estimation. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 11683–11693. PMLR.

Volodymyr Kuleshov, Nathan Fenner, and Stefano Ermon. 2018a. Accurate uncertainties for deep learning using calibrated regression. In *International Conference on Machine Learning*, pages 2796–2804. PMLR.

Volodymyr Kuleshov, Nathan Fenner, and Stefano Ermon. 2018b. Accurate uncertainties for deep learning using calibrated regression. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2796–2804. PMLR.

Bhawesh Kumar, Charlie Lu, Gauri Gupta, Anil Palepu, David Bellamy, Ramesh Raskar, and Andrew Beam. 2023. Conformal prediction with large language models for multi-choice question answering. *arXiv preprint arXiv:2305.18404*.

Salem Lahlou, Moksh Jain, Hadi Nekoei, Victor Ion Butoi, Paul Bertin, Jarrid Rector-Brooks, Maksym Korablyov, and Yoshua Bengio. 2021. Deup: Direct epistemic uncertainty prediction. *arXiv preprint arXiv:2102.08501*.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Quoc V. Le, Alex J. Smola, and Stéphane Canu. 2005. Heteroscedastic gaussian process regression. In *Proceedings of the 22nd International Conference on Machine Learning*, pages 489–496. https://doi.org/10.1145/1102351.1102413

Dan Levi, Liran Gispan, Niv Giladi, and Ethan Fetaya. 2022. Evaluating and calibrating

uncertainty prediction in regression tasks. *Sensors*, 22(15):5540. https://doi.org/10.3390/s22155540

Kaiqu Liang, Zixu Zhang, and Jaime Fernández Fisac. 2024. Introspective planning: Guiding language-enabled agents to refine their own uncertainty.

Arle Lommel, Hans Uszkoreit, and Aljoscha Burchardt. 2014. Multidimensional quality metrics (mqm): A framework for declaring and describing translation quality metrics. *Tradumàtica*, (12):455–463. https://doi.org/10.5565/rev/tradumatica.77

Charles Lu, Andréanne Lemay, Ken Chang, Katharina Höbel, and Jayashree Kalpathy-Cramer. 2022. Fair conformal predictors for applications in medical imaging. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 12008–12016. https://doi.org/10.1609/aaai.v36i11.21459

Qingsong Ma, Ondřej Bojar, and Yvette Graham. 2018. Results of the wmt18 metrics shared task: Both characters and embeddings achieve good performance. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 671–688. https://doi.org/10.18653/v1/W18-6450

Qingsong Ma, Johnny Tian-Zheng Wei, Ondřej Bojar, and Yvette Graham. 2019. Results of the wmt19 metrics shared task: Segment-level and strong mt systems pose big challenges. Association for Computational Linguistics. https://doi.org/10.18653/v1/W19-5302

Lysimachos Maltoudoglou, Andreas Paisios, and Harris Papadopoulos. 2020. Bert-based conformal predictor for sentiment analysis. In *Conformal and Probabilistic Prediction and Applications*, pages 269–284. PMLR.

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.

Abhijit Mishra, Pushpak Bhattacharyya, and Michael Carl. 2013. Automatically predicting sentence translation difficulty. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 346–351.

Jihyung Moon, Naver Papago, Hyunchang Cho, and Eunjeong L. Park. 2020. Revisiting round-trip translation for quality estimation. In *22nd Annual Conference of the European Association for Machine Translation*, page 91.

Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. 2015. Obtaining well calibrated probabilities using Bayesian binning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29. https://doi.org/10.1609/aaai.v29i1.9602

Harris Papadopoulos. 2008. Inductive conformal prediction: Theory and application to neural networks. In *Tools in Artificial Intelligence*. Citeseer. https://doi.org/10.5772/6078

Harris Papadopoulos, Vladimir Vovk, and Alexander Gammerman. 2011. Regression conformal prediction with nearest neighbours. *Journal of Artificial Intelligence Research*, 40:815–840. https://doi.org/10.1613/jair.3198

Shauli Ravfogel, Yoav Goldberg, and Jacob Goldberger. 2023. Conformal nucleus sampling. *arXiv preprint arXiv:2305.02633*.

Ricardo Rei, Ana C. Farinha, Chrysoula Zerva, Daan van Stigt, Craig Stewart, Pedro Ramos, Taisiya Glushkova, André F. T. Martins, and Alon Lavie. 2021. Are references really needed? Unbabel-ist 2021 submission for the metrics shared task. In *Proceedings of the Sixth Conference on Machine Translation*, pages 1030–1040. https://doi.org/10.18653/v1/2020.emnlp-main.213

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. https://doi.org/10.18653/v1/2020.emnlp-main.213

Allen Z. Ren, Anushri Dixit, Alexandra Bodrova, Sumeet Singh, Stephen Tu, Noah Brown, Peng

Xu, Leila Takayama, Fei Xia, Jake Varley, Zhenjia Xu, Dorsa Sadigh, Andy Zeng, and Anirudha Majumdar. 2023. Robots that ask for help: Uncertainty alignment for large language model planners. In *7th Annual Conference on Robot Learning*.

Yaniv Romano, Rina Foygel Barber, Chiara Sabatti, and Emmanuel Candès. 2020. With malice toward none: Assessing uncertainty via equalized coverage. *Harvard Data Science Review*, 2(2):4. https://doi.org/10.1162/99608f92.03f00592

Yaniv Romano, Evan Patterson, and Emmanuel Candes. 2019. Conformalized quantile regression. *Advances in Neural Information Processing Systems*, 32.

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. https://doi.org/10.18653/v1/2020.acl-main.704

Natasa Tagasovska and David Lopez-Paz. 2019. Single-model uncertainties for deep learning. *Advances in Neural Information Processing Systems*, 32.

Ryan J. Tibshirani, Rina Foygel Barber, Emmanuel Candes, and Aaditya Ramdas. 2019. Conformal prediction under covariate shift. *Advances in Neural Information Processing Systems*, 32.

Dennis Ulmer, Chrysoula Zerva, and Andre Martins. 2024. Non-exchangeable conformal language generation with nearest neighbors. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1909–1929, St. Julian's, Malta. Association for Computational Linguistics.

Dennis Thomas Ulmer, Christian Hardmeier, and Jes Frellsen. 2023. Prior and posterior networks: A survey on evidential deep learning methods for uncertainty estimation. *Transactions on Machine Learning Research*.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. 2005. *Algorithmic Learning in a Random World*, volume 29. Springer.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. 2022. Conformal prediction: General case and regression. In *Algorithmic Learning in a Random World*, pages 19–69. Springer. https://doi.org/10.1007/978-3-031-06649-8_2

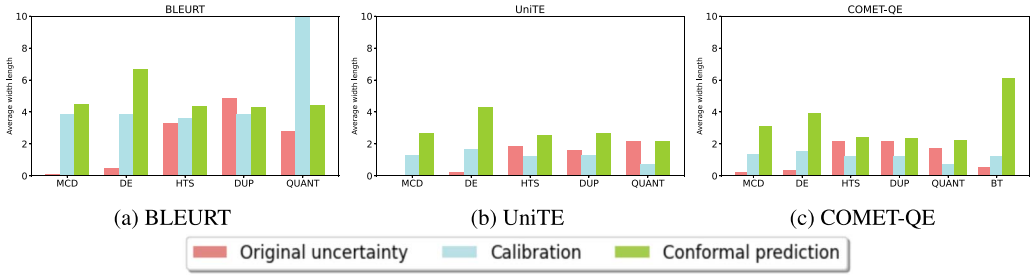
Volodya Vovk, Alexander Gammerman, and Craig Saunders. 1999. Machine-learning applications of algorithmic randomness. In *Proceedings of the Sixteenth International Conference on Machine Learning*, pages 444–453.

Yu Wan, Dayiheng Liu, Baosong Yang, Haibo Zhang, Boxing Chen, Derek Wong, and Lidia Chao. 2022. UNITE: Unified translation evaluation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8117–8127. https://doi.org/10.18653/v1/2022.acl-long.558

Yuxia Wang, Daniel Beck, Timothy Baldwin, and Karin Verspoor. 2022. Uncertainty estimation and reduction of pre-trained models for text regression. *Transactions of the Association for Computational Linguistics*, 10:680–696. https://doi.org/10.1162/tacl_a_00483

Chrysoula Zerva, Taisiya Glushkova, Ricardo Rei, and André F. T. Martins. 2022. Disentangling uncertainty in machine translation evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8622–8641. https://doi.org/10.18653/v1/2022.emnlp-main.591

Chrysoula Zerva, Daan Van Stigt, Ricardo Rei, Ana C. Farinha, Pedro Ramos, José G. C. de Souza, Taisiya Glushkova, Miguel Vera, Fabio Kepler, and André F. T. Martins. 2021. Ist-unbabel 2021 submission for the quality estimation shared task. In *Proceedings of the Sixth Conference on Machine Translation*, pages 961–972.



(a) BLEURT (b) UniTE (c) COMET-QE

Figure 9: Widths obtained for the BLEURT, UniTE, and COMET-QE metrics showing the original intervals (red), the intervals after calibration (light blue) and the intervals after conformal prediction (green).

Hyperparameter	**COMET**	**BLEURT**	**UniTE**	**COMET-QE**
Encoder Model	XLM-R (large)	RemBERT (large)	Info-XLM (large)	XLM-R (large)
Optimizer	Adam	Adam	Adam	Adam
No. frozen epochs	0.3	0.3	0.3	0.3
Learning rate	3e-05	3e-05	3e-05	3e-05
Encoder Learning Rate	1e-05	1e-05	1e-05	1e-05
Layerwise Decay	0.95	0.95	0.95	0.95
Batch size	4	4	4	4
Dropout	0.15	0.15	0.15	0.15
Hidden sizes	[3072, 1024]	[2048, 1024]	[3072, 1024]	[2048, 1024]
Encoder Embedding layer	Frozen	Frozen	Frozen	Frozen
FP precision	32	32	32	32
No. Epochs (training)	2	2	2	2

Table 4: Hyperparameters for MT evaluation metrics used.

A Average Width Across Metrics and Uncertainty Quantifiers

In this section we are presenting the average width of the confidence intervals calculated by the original uncertainty quantification methods, as well as the adapted width when using calibration either by minimising the ECE or by applying conformal prediction. Expanding the analysis on COMET as presented in Figure 6, we are presenting results for BLEURT, UniTE, and COMET-QE. As shown in Figure 9, a similar pattern can be observed for all metrics, where, upon conformalizing, the width increases significantly for MCD, DE, and BT, while changes for the other methods are more moderate.

B Model Implementation and Parameters

Table 4 shows the hyperparameters used to train the following metrics: BLEURT, UniTE, COMET, and COMET-QE. We implemented the models using the COMET codebase[9] and implementation from Zerva et al. (2022) for the uncertainty quantification methods. For deep ensembles, we trained 5 models with different seeds. For MCD we used a total of 100 runs following Glushkova et al. (2021) and Zerva et al. (2022). For the DUP method, we used a bottleneck layer with dimensionality 256, and we maintained the same setup across metrics.

C Equalized Conformal Prediction Across Uncertainty Quantification Methods

In this section, we extend the analysis discussed in Section 4 of the main paper, to the rest of the quantification methods for the COMET metric, shown in Figures 10 to 13. We can see that direct uncertainty prediction (Figure 12) and quantile regression (Figure 13) are the two methods that suffered less from imbalanced coverage, even for extreme values of quality and uncertainty, supporting their suitability for MT evaluation, as also shown for the general results in Section 3.4. We can also observe that when the initial calibration step yields balanced results around the desired α, the recalibration brings no significant benefits and may even result in slightly lower coverage. Hence, it is important to first detect for which, if any, attributes we may need to recalibrate.

[9] `https://github.com/Unbabel/COMET`, version 2.1.0.

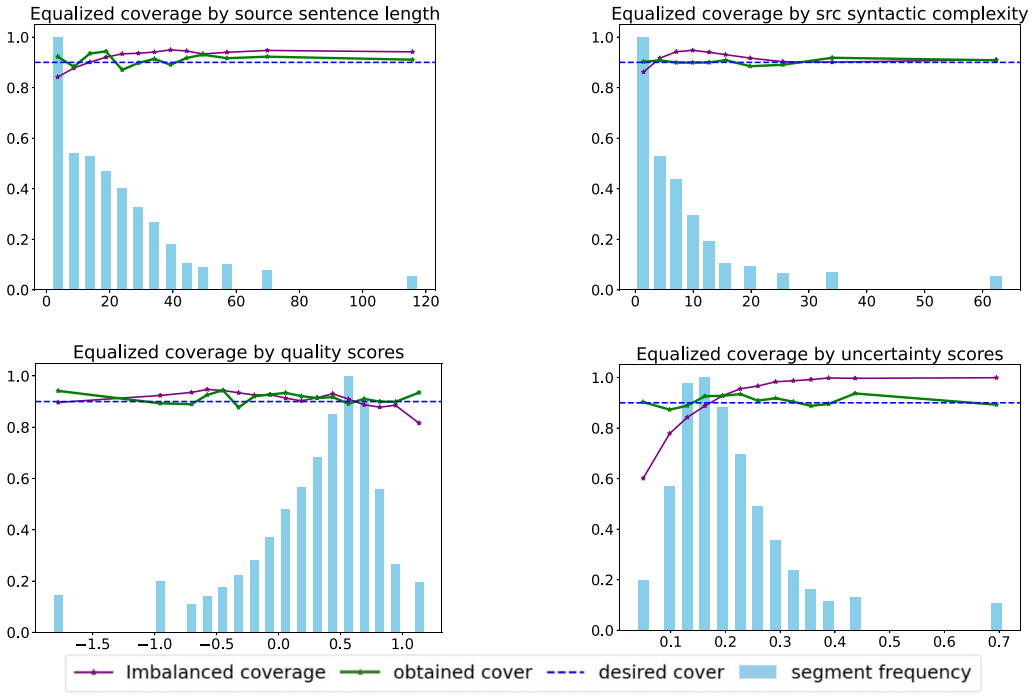


Figure 10: Equalized prediction for COMET using deep ensembles.

Figure 11: Equalized prediction for COMET using heteroscedastic regression.

Figure 12: Equalized prediction for COMET using direct uncertainty prediction.

Figure 13: Equalized prediction for COMET using quantile regression.