Proceedings of Machine Learning Research 204:1–17, 2023 Conformal and Probabilistic Prediction with Applications

Evaluating Machine Translation Quality with Conformal Predictive Distributions

Patrizio Giovannotti PATRIZIO.GIOVANNOTTI.2019@LIVE.RHUL.AC.UK
Royal Holloway, University of London, Egham, Surrey, UK
Centrica plc, UK

Editor: Harris Papadopoulos, Khuong An Nguyen, Henrik Boström and Lars Carlsson

Abstract

This paper presents a new approach for assessing uncertainty in machine translation by simultaneously evaluating translation quality and providing a reliable confidence score. Our approach utilizes conformal predictive distributions to produce prediction intervals with guaranteed coverage, meaning that for any given significance level ϵ, we can expect the true quality score of a translation to fall out of the interval at a rate of $1 - \epsilon$. In this paper, we demonstrate how our method outperforms a simple, but effective baseline on six different language pairs in terms of coverage and sharpness. Furthermore, we validate that our approach requires the data exchangeability assumption to hold for optimal performance.

Keywords: Conformal prediction, machine translation, natural language processing, uncertainty estimation.

1. Introduction

Machine translation (MT) is the task of automatically translating text from one language to another using a computer program. With the growing globalization of businesses and the increasing availability of multilingual content on the internet, machine translation has become an essential technology for communication across language barriers. However, machine translation is still far from perfect and often produces translations that are inaccurate or of poor quality. Evaluating the quality of machine-translated text is therefore crucial for ensuring the usability and effectiveness of translation systems. Among the several ways to evaluate MT outputs, quality estimation (QE) is the task of determining a quality score for machine-translated text without the help of reference translations. Good quality estimation is essential whenever a real-time decision must be made about a translation output, like whether to publish a translated text or informing the user about the confidence in a translation. QE can also be useful for other purposes such as selecting the best translation among multiple candidates or providing feedback to MT developers.

Since QE was introduced as a shared task in the Conference of Machine Translation (Callison-Burch et al., 2012), increasingly more systems have been able to improve on the baselines and redefine the state of the art. However, few efforts have been made towards refining how these models quantify the uncertainty of their predictions. In this paper, we propose a novel approach to express the uncertainty of quality estimates with the help of conformal predictive distributions (CPD – Vovk et al., 2017). Under the sole assumption of the data being IID, our method produces prediction intervals with *guaranteed coverage*, that is, for any chosen confidence level $1 - \epsilon$, prediction intervals will fail to include the correct

label at a rate ϵ. The ability to specify arbitrary intervals can be useful in many ways. For example, we can set a quality threshold q_{err} and compute the probability of a proposed translation to have quality smaller than q_{err}, i.e. to be "not good enough". CPD ensures that such a probability reflects the long-term relative frequency of correct predictions. For example, let the random variable Q model the quality score of any sentence pair; let $\mathcal{S}$ be the set of examples where our model predicts $P(Q \leq q_{\text{err}}) = 0.8$. Then, the number of examples in $\mathcal{S}$ for which the true label is smaller than q_{err} is $|\mathcal{S}| \cdot 0.8$.

Our contributions are the following: we demonstrate how to apply CPD to state-of-the-art MT evaluation models to equip them with uncertainty estimation capabilities; we verify that we can build prediction intervals with good coverage, given that the IID assumption holds, for several language pairs; we report results on the ability of these models to predict translation failures, that is where their quality score is too low to be accepted.

2. Related Work

In this section, we provide some context on the quality estimation problem and prior studies that have focused on estimating its uncertainty. Moreover, we discuss the application of conformal methods to various areas of natural language processing (NLP).

2.1. Background

Following Glushkova et al. (2021), we can formalise our problem as follows: let s be a sentence in the source language, t the same sentence translated by an MT system, and $\mathcal{R} = \{r_1, \ldots, r_{|\mathcal{R}|}\}$ a set of reference translations. An MT evaluator is a function that accepts as input a tuple $\langle s, t, \mathcal{R} \rangle$ and outputs a quality score $\hat{q} \in \mathbb{R}$.

The simplest evaluator would be a metric that could quantify the amount of lexical overlap between t and r_i. Examples of this approach are the popular BLEU score (Papineni et al., 2002) and Meteor (Lavie and Denkowski, 2009). The success of neural machine translation techniques inspired new metrics such as BERTScore (Zhang et al., 2020), where quality scores depends on the pairwise cosine similarity between words in t and r_i, once each word is encoded via a BERT model (Devlin et al., 2019). Despite BERT-based metrics showed to be robust under distribution shifts (Vu et al., 2022) it is still unclear how much these metrics are suitable to adequately reflect a model's performance (Blagec et al., 2022).

The evaluation metrics mentioned above require a set of reference translations. The field of *quality estimation* is focused on the case $\mathcal{R} = \emptyset$, i.e. estimating translation quality in the absence of reference translations.

2.2. Quality estimation for MT

In many real-world situations, reference translations are not available, and the only score returned by modern transformer-based models is the sum of the log-probabilities of each token in the generated sentence. This score, denoted by a real number $c \in (-\infty, 0]$, can be used to rank different translation candidates, however it does not generally correlate with human judgement concerning quality of translation.

To help in this scenario, known as quality estimation, several datasets with human-annotated quality scores have been created. Such scores include direct assessments (DA, Graham et al., 2013) and human translation error rates (HTER, Snover et al., 2006). MT evaluation systems are asked to generate quality scores $\hat{q}$ that correlate with ground truth scores q^* as much as possible, in what is in essence a regression task.

Bleurt (Sellam et al., 2020) and Comet (Rei et al., 2020) are two examples of quality estimators, with the latter relying on a multilingual RoBERTa pre-trained model (Liu et al., 2019).

2.3. Uncertainty quantification in MT

Although many QE efforts have achieved good predictive performance, a significant limitation of the proposed models is that they are often unable to convey the uncertainty associated with their predictions. The entire topic of uncertainty quantification for MT has not been explored enough, with Beck et al. (2016) being among the very few to have addressed the issue in several years.

More recently, Glushkova et al. (2021) presented a modified version of Comet that outputs quality scores *intervals* of variable width, depending on the confidence associated to the prediction. They propose to treat translation quality as a random variable Q and predict a distribution $\hat{P}_Q(q)$, rather than a point estimate $\hat{q}$. They choose two parametric approaches: *MC dropout*, where h is run N times, with different units dropped out each time, and *deep ensembles*, where N separate h instances are randomly initialised and used to obtain scores. Both methods provide a set of quality scores $\mathcal{Q} = \{\hat{q}_1, \ldots, \hat{q}_N\}$. The authors treat $\mathcal{Q}$ as a sample drawn from a Gaussian distribution, hence they estimate mean $\hat{\mu}$ and variance $\hat{\sigma}^2$ to use in the calculation of confidence intervals $I[q_{min}(\epsilon), q_{max}(\epsilon)]$ for $\epsilon \in [0, 1]$. The aim is to provide a quality score interval that includes q^* as much as possible, while being as tight as possible. Some limitations of this approach are the need of training several models or predict several times, the need for a post-calibration step and, more importantly, the strong assumption about the shape of $\hat{P}_Q(q)$. In contrast, our approach requires only one model, is well calibrated out of the box and makes no assumption about the distribution of $\hat{P}_Q(q)$.

2.4. Conformal methods for NLP

Conformal prediction has been applied in several forms to numerous NLP tasks. Paisios et al. (2019) first explored the use of traditional CP for text classification, while Maltoudoglou et al. (2020) and Giovannotti and Gammerman (2021) extended it to sentiment analysis and paraphrase detection by building on a BERT pre-trained model and experimenting with new nonconformity measures; Giovannotti (2022) studied the use of Venn–ABERS predictors (Vovk and Petej, 2014) in the context of binary classification for natural language understanding; Dey et al. (2022) applied CP to part-of-speech tagging and the important task of text infilling (or masked language modelling). Very recently, Robinson et al. (2023) examined the use of conformal prediction for question answering in the context of large language models, while Ravfogel et al. (2023) used a conformal approach to calibrate the parameter p in top-p sampling for language generation.

Conformal methods have also been used to optimize transformers, the current state-of-the-art of NLP architectures: Schuster et al. (2022) proposed a method to accelerate text generation through an early-exiting mechanism enabled by conformal prediction.

3. Methodology

Our methods are rooted in recent advances in conformal prediction (CP), a machine learning framework introduced by Gammerman et al. (1998) and fully developed by Vovk et al. (2005).

3.1. Conformal predictive distributions

Introduced in Vovk et al. (2017), conformal predictive distributions (CPDs) are a novel approach to estimating the probability distribution of a continuous variable that depends on a number of features. Under minimal assumptions, i.e. the data being generated independently by an unknown fixed distribution, CPDs provide probabilities that correspond to long-term frequencies. CPDs make no assumption on the particular distribution of the data, hence no prior is required either.

In this work, we will use a computationally efficient version of CPD, namely split conformal predictive distributions (Vovk et al., 2020). In the split CPDs framework, we require our original training sequence $z_1, \dots, z_n$ to be divided into a proper training sequence $z_1, \dots, z_m$ and a calibration sequence $z_{m+1}, \dots, z_n$. Here each observation is a pair $z = (x, y)$ of an object $x \in \mathbf{X}$ and its label $y \in \mathbb{R}$, where $\mathbf{X}$ is any nonempty measurable space.

The essential component of CPDs is a (split) conformity measure, a function that should indicate how large is any label y_{m+1} compared to the m labels in the proper training set. The standard choice of conformity measure for a (test) observation (x, y) is

$$A(z_1, \dots, z_m, (x, y)) = \frac{y - \hat{y}}{\hat{\sigma}}$$

where $\hat{y}$ is a prediction for y and $\hat{\sigma}$ is an estimate of the quality of $\hat{y}$ (this is also referred as *difficulty*, see also Boström et al., 2021). We will obtain $\hat{y}$ from a pretrained transformer model that we fine-tune on $z_1, \dots, z_m$ (see Section 3.2); as for $\hat{\sigma}$, we use the sum of the distances between (x, y) and its K nearest neighbours.

Once A is defined, we are able to compute the conformity scores

$$\begin{aligned} \alpha_i &= A(z_1, \dots, z_m, (x_i, y_i)) \qquad i = m+1, \dots, n \\ \alpha^y &= A(z_1, \dots, z_m, (x, y)) \end{aligned}$$

which express how well a label conforms to the property of being large (we think of a label as nonconforming only when it is too small). The *split conformal transducer* built on A produces the values

$$Q(z_1, \dots, z_n, (x, y)) := \frac{1}{n - m + 1} |\{i = m+1, \dots, n \mid \alpha_i < \alpha^y\}| + Q_\tau$$

with

$$Q_\tau := \frac{\tau}{n - m + 1} |\{i = m+1, \dots, n \mid \alpha_i = \alpha^y\}| + \frac{\tau}{n - m + 1}, \qquad \tau \sim U(0, 1).$$

Because of the validity property of conformal transducers, values of Q are uniformly distributed on $[0, 1]$ when $z_1, \ldots, z_m, z$ are IID. Additionally, it can be proved that the function Q is monotonically increasing in y and tends asymptotically to 0 and 1 when y tends to $-\infty$ and $+\infty$, respectively. These properties are what defines *randomized predictive systems*, which are able to generate distribution functions. In other words, CPDs can be considered as p-values arranged into a distribution function (see Figure 1).

An accessible tutorial on CPDs was written by Toccaceli (2020).

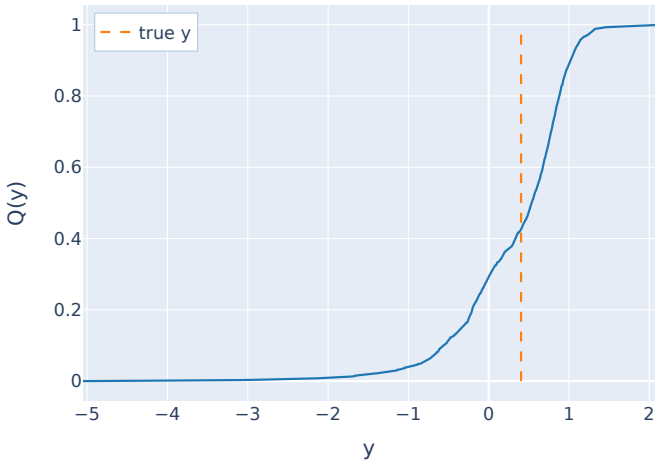

Figure 1: Conformal predictive distribution for a test example of the English→German dataset. Values for the quality label y are normalized.

In practice, for a new test object x, a split conformal predictive system can be implemented as follows:

1. Predict a label $\hat{y} \in \mathbb{R}$ and its estimated quality $\hat{\sigma}$
2. Calculate the calibration scores $C_i = \hat{y} + \frac{\hat{\sigma}}{\hat{\sigma}_i}(y_i - \hat{y}_i)$ for $i = m + 1, \ldots, n$
3. Sort the confidence scores in ascending order, obtaining the sequence $C_{(1)}, \ldots, C_{(n-m)}$, then set $C_{(0)} = -\infty$ and $C_{(n-m+1)} = \infty$
4. Return the predictive distribution

$$Q(y) = \begin{cases} \frac{i+\tau}{n-m+1} & \text{if } y \in (C_i, C_{i+1}), \text{ for } i \in \{0, 1, \ldots, n-m\} \\ \frac{i'-1+(i''-i'+2)\tau}{n-m+1} & \text{if } y = C_i, \text{ for } i \in \{1, \ldots, n-m\} \end{cases} \tag{1}$$

where $i' := \min\{j : C_j = C_i\}$ and $i'' := \max\{j : C_j = C_i\}$.

3.2. Underlying algorithm

Our algorithm of choice for predicting quality scores is XLM-RoBERTa (Conneau et al., 2020), a multilingual masked language model that was specifically trained to compute dense representations of pairs of sentences written in different languages.

As a variant of the RoBERTa model (Liu et al., 2019) pre-trained on multilingual corpora, XLM-RoBERTa is a transformer-based architecture that learns representations of text by processing it in a series of layers, each of which applies self-attention to the input tokens. This allows the model to capture long-range dependencies and contextual information in the text. The RoBERTa model was pre-trained on a large corpus of diverse text in multiple languages using the masked language modelling (MLM) and the next sentence prediction (NSP) objectives. This pre-training has been shown to improve the model's performance on downstream NLP tasks.

XLM-RoBERTa further improves upon the RoBERTa model by using cross-lingual language modeling (XLM) pre-training. This involves training the model on multiple languages simultaneously and encouraging it to learn shared representations of language. This enables the model to transfer knowledge between languages, which can be useful for MT quality estimation, as the model can learn to recognize patterns in one language that are indicative of high quality translations in another language.

In our experiments, we fine-tune the pre-trained `xlm-roberta-base` model – equipped with a single-unit classification head, for regression tasks – on the MT quality estimation dataset and use its predicted quality scores as a feature to train a second, lighter KNN model. Our KNN model takes as input these RoBERTa predictions and the quality scores predicted by another baseline model, which are included with in WMT 20 dataset. Training the KNN model over these two features resulted in improved Pearson's correlation with the true labels.

3.3. Baseline

We compare our approach with the same baseline method proposed by Glushkova et al. (2021), which proved to perform reasonably well despite its simplicity. As a baseline, we map the original quality scores $\hat{q}$ computed by our RoBERTa model to a Gaussian distribution $\mathcal{N}(q; \hat{\mu}, \hat{\sigma}^2)$ with $\hat{\mu} := \hat{q}$ and $\hat{\sigma}^2 := \sigma_{\text{fixed}}$, the average of the squared residuals $(\hat{q} - \hat{\mu})^2$ over the validation set. One major limitation of our baseline is its inability to adapt the quality intervals' width to the uncertainty associated to each example.

In order to generate the prediction interval for a confidence level α, we compute

$$[q_{\min}(\alpha), q_{\max}(\alpha)] = \hat{\mu} \pm \hat{\sigma} \cdot \text{probit}\left(\frac{1+\alpha}{2}\right)$$

where we used the quantile function: $\text{probit}(p) = \sqrt{2}\text{erf}^{-1}(2p-1)$, where erf is the error function.

3.4. Evaluation Metrics

We evaluated the performance of our proposed method, CPD, against our baseline approach across three metrics: Expected Calibration Error (ECE), Sharpness, and AUROC.

Expected Calibration Error (ECE) ECE measures the difference between the expected accuracy of a set of predictions and their actual accuracy. It was first introduced by Naeini et al. (2015). We will use the version described by Kuleshov et al. (2018) for regression:

$$\text{ECE} = \frac{1}{|\mathcal{E}|} \sum_{\epsilon \in \mathcal{E}} |\text{err}(\epsilon) - \epsilon|,$$

where $\mathcal{E}$ is a set of significance levels $\epsilon \in [0, 1]$ and $\text{err}(\epsilon)$ is the error rate at a given significance level, namely the proportion of prediction intervals that do not include the true label (for a perfectly calibrated system, ECE $= 0$). This type of error depends on the number of significance levels we consider: in our work, we chose $|\mathcal{E}| = 50$, that is $\mathcal{E} = \{0.00, 0.02, 0.04, \ldots, 1.00\}$.

Sharpness Sharpness measures the degree of concentration of the predicted scores around the actual scores. A sharper distribution implies that the model is more confident in its predictions. In our work, sharpness will be measured by the average prediction interval width at a certain confidence level. We report sharpness values for prediction intervals at 90% confidence, since we are more interested in high-confidence predictions, whereas higher-confidence intervals are likely to be too wide to be useful in a real scenario.

AUROC The Area Under the Receiver Operating Characteristic Curve measures the ability of a model to distinguish between positive and negative samples, and is particularly useful when the classes are imbalanced. AUROC measures the area under the curve obtained by plotting a model's true positive rate against its false positive rate at different classification thresholds. The AUROC score ranges from 0.5 to 1, with a score of 0.5 indicating random guessing and a score of 1 indicating perfect classification. We use AUROC to assess the ability of our models to detect critically wrong translations, i.e. with quality q^* in the bottom decile of the test set.

4. Experiments

Our experiments made extensive use of the two main Hugging Face libraries `transformers` (Wolf et al., 2020) and `datasets` (Lhoest et al., 2021). The conformal predictive distribution implementation used is `crepes` (Boström, 2022).

4.1. Datasets and experimental setup

Our dataset was released for Task 1 of the WMT 2020 conference (Specia et al., 2020). It consists of labelled sentence pairs (s, t) for 6 language pairs: two high-resource English→German (En-De) and English→Chinese (En-Zh) pairs; two medium-resource Romanian→English (Ro-En) and Estonian→English (Et-En) pairs; and two low-resource Sinhala→English (Si-En) and Nepali→English (Ne-En) pairs. Most of the sentences are extracted from Wikipedia and translated with a transformer-based NMT model trained using publicly available data.

Each sentence is manually labelled with a score from 0 to 100 by a group of 3 independent annotators. The score is assigned on the basis of the following guidelines: the 0-10 range

represents an incorrect translation; 11-29, a translation with few correct keywords, but the overall meaning is different from the source; 30-50, a translation with major mistakes; 51-69, a translation which is understandable and conveys the overall meaning of the source but contains typos or grammatical errors; 70-90, a translation that closely preserves the semantics of the source sentence; and 91-100, a perfect translation. These guidelines were introduced in FLORES (Guzmán et al., 2019).

All quality scores are then transformed into their z-normalized values. The label to predict is then $y = \frac{q^* - \mu_{\mathcal{D}}}{\sigma_{\mathcal{D}}}$ where $\mu_{\mathcal{D}}, \sigma_{\mathcal{D}}$ are respectively the mean and the standard deviation of the quality scores q^* in the dataset $\mathcal{D}$. For more information about the dataset creation and composition refer to Specia et al. (2020).

All datasets were randomly shuffled to ensure the independent and identically distributed (IID) assumption, which is crucial for conformal prediction as a framework in general. In Figure 2(a), we plot the distribution of label values for training and test sets of the Estonian→English dataset, before shuffling. Figure 2(b) shows the same quantities after randomly shuffling the original dataset. As can be seen, the shuffled version exhibits a higher similarity in label distribution between training and test set, indicating that the shuffling leads to a better IID property of the data.

Our experiments confirm the impact of shuffling on the performance of our models. The results show that our method consistently performs better on shuffled datasets, compared to datasets that were not shuffled. More details are included in Appendix A.

We shuffle our datasets three times, each with a different random seed, and report the average performance of the models over the three versions of the dataset.

Our fine-tuning process is a three-epochs training of the pre-trained XLM-RoBERTa model over each dataset; we keep the model which scored the best Pearson's correlation on the validation set among the three epochs.

4.2. Results and analysis

In this section we demonstrate and analyse our results. Table 1 summarises the performance of our models against the metrics described in section 3.4.

Coverage ECE measures how often the prediction intervals fail to include the true label, averaged over several significance levels $\epsilon \in \mathcal{E}$. For instance, an ECE of 1% corresponds to a 1% rate of the true label missing in the prediction interval, on average over $|\mathcal{E}|$ significance levels.

We note that the baseline model performs relatively well, apart from the English→German case. However, our model based on CPD consistently improves over the baseline by a factor of at least 2× and does not suffer as much on the English→German dataset. The calibration error is always less or equal to 2%, a good coverage result that may be improved upon with the use of more complex underlying algorithms or additional textual features.

It is important to highlight that these results are conditional to the datasets being randomly shuffled; in Appendix A we show how CPD's performance is poorer whenever the IID assumption is less likely to hold.

Sharpness Again we verify a consistent improvement over the baseline. Differences are not as marked as they are in the coverage case, but in general we can say that at 90%

(a) (b)

Figure 2: Distribution of label values (a) before and (b) after shuffling the Estonian→English dataset splits. Before shuffling, there is an evident mismatch between train and test distributions, hence the IID property may not hold.

		%ECE	Sha@90%	AUC@10%
En-De	baseline	13.88	2.81	0.63
	CPD	2.06	2.29	0.62
En-Zh	baseline	3.19	2.51	0.78
	CPD	1.06	2.48	0.78
Ro-En	baseline	3.96	1.92	0.92
	CPD	1.88	1.77	0.92
Et-En	baseline	4.20	2.45	0.83
	CPD	1.69	2.33	0.83
Si-En	baseline	2.82	2.61	0.80
	CPD	1.34	2.53	0.81
Ne-En	baseline	3.11	2.40	0.79
	CPD	1.62	2.39	0.80

Table 1: Performance averaged over three runs with different random train/validation/test splits. Our CPD-based model consistently outperforms the baseline in terms of expected calibration error and sharpness.

confidence ($\epsilon = 0.1$), CPD produces tighter prediction intervals with better coverage. In Figure 3 we have a detailed view of how prediction interval size varies with significance level, for both models. On the English→German dataset, we note that CPD widens its intervals rapidly when approaching high-confidence levels ($\epsilon < 0.1$). At the upper limit of the significance range, baseline models exhibit greater sharpness than CPD. However, narrow intervals can be misleading if not paired to high-coverage predictions. We found this behaviour to be less prominent in the case of the other language pairs.

Note that the sharpness values in Table 1 relate to the prediction intervals for the z-normalised quality scores y. For the Estonian→English dataset, for example, $q^* \in [0, 100]$ corresponds to $y \in [-2.8, 1.4]$, and a prediction interval of width 2.46 would span a large quality range ($\hat{q} \in [20, 90]$).

Detection of critical failures AUC@10% is the AUROC score achieved by each model for the task of predicting if a translation is "not good enough" (specifically, if its quality score falls in the bottom decile). For this binary classification task, both baseline and CPD models achieve essentially the same results, both suffering on the English→German dataset and doing very well on Romanian→English.

5. Conclusion and Future Work

We presented a novel approach to quality estimation for MT based on conformal predictive distributions. Rather than returning a single quality score for each sentence pair, our model generates a prediction interval which is larger the more is the uncertainty of the prediction. More importantly, the predictions have guaranteed coverage under the IID assumption.

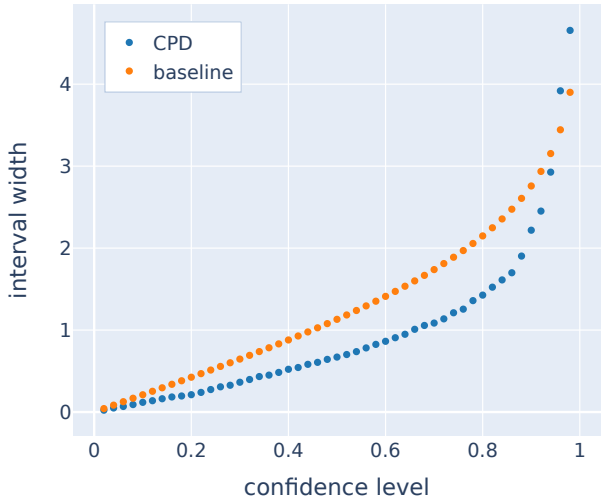

Figure 3: Sharpness performance on the English→German dataset. Each point is the prediction interval size averaged over all test examples for a particular confidence level $1 - \epsilon$. CPD intervals are tighter on average, then increase in size rapidly when approaching higher confidence levels.

The prediction intervals obtained allow for a range of useful "downstream" tasks, such as deciding whether to publish a specific translation or not, or to inform users about the confidence in the quality estimate of a certain translation, or to rank translations produced by different MT models.

Our experimental results confirm the importance of the IID assumption for the successful application of conformal methods in NLP tasks. One potential avenue for future research is to explore methods for determining whether the IID assumption has been violated during the training process and at what point this occurred. Vovk et al. (2021)'s recent research in this area is a promising step towards addressing this challenge.

Acknowledgements

Thanks to Alex Gammerman and Ilia Nouretdinov for their constant support. Thanks to Chris Watkins and Alexander Balinsky for some insightful conversations. This work was partly supported by Centrica PLC.

Appendix A. Importance of shuffling

Table 2 reports the performance over the original dataset splits. We can see that CPD does not always improve calibration over the Gaussian baseline. CPD does much better when we shuffle the datasets into new train/validation/test splits (Table 1).

		ECE	Sha@90%	AUC@10%
En-De	baseline	6.56	2.24	0.637
	CPD	3.75	1.99	0.642
En-Zh	baseline	1.31	2.21	0.728
	CPD	1.58	2.17	0.727
Ro-En	baseline	7.48	1.66	0.963
	CPD	4.04	1.54	0.963
Et-En	baseline	1.65	2.10	0.877
	CPD	2.78	2.02	0.877
Si-En	baseline	3.20	2.32	0.845
	CPD	4.03	2.31	0.850
Ne-En	baseline	5.23	2.04	0.883
	CPD	3.29	1.97	0.871

Table 2: Performance over the original train/validation/test splitting.

References

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

Kathrin Blagec, Georg Dorffner, Milad Moradi, Simon Ott, and Matthias Samwald. A global analysis of metrics used for measuring performance in natural language processing. In *Proceedings of NLP Power! The First Workshop on Efficient Benchmarking in NLP*, pages 52–63, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.nlppower-1.6. URL https://aclanthology.org/2022.nlppower-1.6.

Henrik Boström. crepes: a python package for generating conformal regressors and predictive systems. In Ulf Johansson, Henrik Boström, Khuong An Nguyen, Zhiyuan Luo, and Lars Carlsson, editors, *Proceedings of the Eleventh Symposium on Conformal and Probabilistic Prediction with Applications*, volume 179 of *Proceedings of Machine Learning Research*, pages 24–41. PMLR, 24–26 Aug 2022. URL https://proceedings.mlr.press/v179/bostrom22a.html.

Henrik Boström, Ulf Johansson, and Tuwe Löfström. Mondrian conformal predictive distributions. In Lars Carlsson, Zhiyuan Luo, Giovanni Cherubin, and Khuong An Nguyen, editors, *Proceedings of the Tenth Symposium on Conformal and Probabilistic Prediction and Applications*, volume 152 of *Proceedings of Machine Learning Research*, pages 24–38. PMLR, 08–10 Sep 2021. URL https://proceedings.mlr.press/v152/bostrom21a.html.

Chris Callison-Burch, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. Findings of the 2012 workshop on statistical machine translation. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 10–51, Montréal, Canada, June 2012. Association for Computational Linguistics. URL https://aclanthology.org/W12-3102.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.747. URL https://aclanthology.org/2020.acl-main.747.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://www.aclweb.org/anthology/N19-1423.

Neil Dey, Jing Ding, Jack Ferrell, Carolina Kapper, Maxwell Lovig, Emiliano Planchon, and Jonathan P. Williams. Conformal prediction for text infilling and part-of-speech prediction. *The New England Journal of Statistics in Data Science*, pages 1–15, 2022. ISSN 2693-7166. doi: 10.51387/22-NEJSDS8.

A. Gammerman, V. Vovk, and V. Vapnik. Learning by transduction. In *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence*, UAI'98, page 148–155, San Francisco, CA, USA, 1998. Morgan Kaufmann Publishers Inc. ISBN 155860555X.

Patrizio Giovannotti. Calibration of natural language understanding models with venn–abers predictors. In Ulf Johansson, Henrik Boström, Khuong An Nguyen, Zhiyuan Luo, and Lars Carlsson, editors, *Proceedings of the Eleventh Symposium on Conformal and Probabilistic Prediction with Applications*, volume 179 of *Proceedings of Machine Learning Research*, pages 55–71. PMLR, 24–26 Aug 2022. URL https://proceedings.mlr.press/v179/giovannotti22a.html.

Patrizio Giovannotti and Alex Gammerman. Transformer-based conformal predictors for paraphrase detection. In Lars Carlsson, Zhiyuan Luo, Giovanni Cherubin, and Khuong An Nguyen, editors, *Proceedings of the Tenth Symposium on Conformal and Probabilistic Prediction and Applications*, volume 152 of *Proceedings of Machine Learning Research*, pages 243–265. PMLR, 08–10 Sep 2021. URL https://proceedings.mlr.press/v152/giovannotti21a.html.

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

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. Continuous measurement scales in human evaluation of machine translation. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 33–41, Sofia, Bulgaria, August 2013. Association for Computational Linguistics. URL https://aclanthology.org/W13-2305.

Francisco Guzmán, Peng-Jen Chen, Myle Ott, Juan Pino, Guillaume Lample, Philipp Koehn, Vishrav Chaudhary, and Marc'Aurelio Ranzato. The FLORES evaluation datasets for low-resource machine translation: Nepali–English and Sinhala–English. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6098–6111, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1632. URL https://aclanthology.org/D19-1632.

Volodymyr Kuleshov, Nathan Fenner, and Stefano Ermon. Accurate uncertainties for deep learning using calibrated regression. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2796–2804. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/kuleshov18a.html.

Alon Lavie and Michael J Denkowski. The meteor metric for automatic evaluation of machine translation. *Machine translation*, pages 105–115, 2009.

Quentin Lhoest, Albert Villanova del Moral, Yacine Jernite, Abhishek Thakur, Patrick von Platen, Suraj Patil, Julien Chaumond, Mariama Drame, Julien Plu, Lewis Tunstall, Joe Davison, Mario Šaško, Gunjan Chhablani, Bhavitvya Malik, Simon Brandeis, Teven Le Scao, Victor Sanh, Canwen Xu, Nicolas Patry, Angelina McMillan-Major, Philipp Schmid, Sylvain Gugger, Clément Delangue, Théo Matussière, Lysandre Debut, Stas Bekman, Pierric Cistac, Thibault Goehringer, Victor Mustar, François Lagunas, Alexander Rush, and Thomas Wolf. Datasets: A community library for natural language processing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 175–184, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.emnlp-demo.21.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

Lysimachos Maltoudoglou, Andreas Paisios, and Harris Papadopoulos. Bert-based conformal predictor for sentiment analysis. volume 128 of *Proceedings of Machine Learning Research*, pages 269–284. PMLR, 09–11 Sep 2020. URL http://proceedings.mlr.press/v128/maltoudoglou20a.html.

Mahdi Pakdaman Naeini, Gregory F Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the... AAAI Conference*

on Artificial Intelligence. AAAI Conference on Artificial Intelligence, volume 2015, page 2901. NIH Public Access, 2015.

Andreas Paisios, Ladislav Lenc, Jiří Martínek, Pavel Král, and Harris Papadopoulos. A deep neural network conformal predictor for multi-label text classification. volume 105 of *Proceedings of Machine Learning Research*, pages 228–245, Golden Sands, Bulgaria, 09–11 Sep 2019. PMLR. URL http://proceedings.mlr.press/v105/paisios19a.html.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, USA, 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL https://doi.org/10.3115/1073083.1073135.

Shauli Ravfogel, Yoav Goldberg, and Jacob Goldberger. Conformal nucleus sampling, 2023.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 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, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.213. URL https://aclanthology.org/2020.emnlp-main.213.

Joshua Robinson, Christopher Michael Rytting, and David Wingate. Leveraging large language models for multiple choice question answering, 2023.

Tal Schuster, Adam Fisch, Jai Gupta, Mostafa Dehghani, Dara Bahri, Vinh Tran, Yi Tay, and Donald Metzler. Confident adaptive language modeling. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 17456–17472. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/6fac9e316a4ae75ea244ddcef1982c71-Paper-Conference.pdf.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.704. URL https://aclanthology.org/2020.acl-main.704.

Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA, August 8-12 2006. Association for Machine Translation in the Americas. URL https://aclanthology.org/2006.amta-papers.25.

Lucia Specia, Frédéric Blain, Marina Fomicheva, Erick Fonseca, Vishrav Chaudhary, Francisco Guzmán, and André F. T. Martins. Findings of the WMT 2020 shared task on quality estimation. In *Proceedings of the Fifth Conference on Machine Translation*,

pages 743–764, Online, November 2020. Association for Computational Linguistics. URL https://aclanthology.org/2020.wmt-1.79.

Paolo Toccaceli. Tutorial on conformal predictive distributions, 2020. URL https://cml.rhul.ac.uk/people/ptocca/HomePage/COPA2020___Tutorial_on_Predictive_Distributions.pdf.

V. Vovk and Ivan Petej. Venn-abers predictors. In *UAI*, 2014. URL http://alrw.net/articles/07.pdf.

Vladimir Vovk, Alex Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*. Springer Science & Business Media, 2005. doi: https://doi.org/10.1007/b106715.

Vladimir Vovk, Jieli Shen, Valery Manokhin, and Min-ge Xie. Nonparametric predictive distributions based on conformal prediction. In Alex Gammerman, Vladimir Vovk, Zhiyuan Luo, and Harris Papadopoulos, editors, *Proceedings of the Sixth Workshop on Conformal and Probabilistic Prediction and Applications*, volume 60 of *Proceedings of Machine Learning Research*, pages 82–102. PMLR, 13–16 Jun 2017. URL https://proceedings.mlr.press/v60/vovk17a.html.

Vladimir Vovk, Ivan Petej, Ilia Nouretdinov, Valery Manokhin, and Alexander Gammerman. Computationally efficient versions of conformal predictive distributions. *Neurocomputing*, 397:292–308, 2020. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2019.10.110. URL https://www.sciencedirect.com/science/article/pii/S0925231219316042.

Vladimir Vovk, Ivan Petej, Ilia Nouretdinov, Ernst Ahlberg, Lars Carlsson, and Alex Gammerman. Retrain or not retrain: conformal test martingales for change-point detection. In Lars Carlsson, Zhiyuan Luo, Giovanni Cherubin, and Khuong An Nguyen, editors, *Proceedings of the Tenth Symposium on Conformal and Probabilistic Prediction and Applications*, volume 152 of *Proceedings of Machine Learning Research*, pages 191–210. PMLR, 08–10 Sep 2021. URL https://proceedings.mlr.press/v152/vovk21b.html.

Doan Nam Long Vu, Nafise Sadat Moosavi, and Steffen Eger. Layer or representation space: What makes BERT-based evaluation metrics robust? In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3401–3411, Gyeongju, Republic of Korea, October 2022. International Committee on Computational Linguistics. URL https://aclanthology.org/2022.coling-1.300.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.emnlp-demos.6.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=SkeHuCVFDr.