Uncertainty in Structured Prediction

Andrey Malinin 1  Mark Gales 2

Abstract

Uncertainty estimation is important for ensuring safety and robustness of AI systems, especially for high-risk applications. While much progress has recently been made in this area, most research has focused on unstructured prediction, such as image classification and regression tasks. However, while task-specific forms of confidence score estimation have been investigated by the speech and machine translation communities, limited work has investigated general uncertainty estimation approaches for structured prediction. Thus, this work aims to investigate uncertainty estimation for structured prediction tasks within a single unified and interpretable probabilistic ensemble-based framework. We consider uncertainty estimation for sequence data at the token-level and complete sequence-level, provide interpretations for, and applications of, various measures of uncertainty and discuss the challenges associated with obtaining them. This work also explores the practical challenges associated with obtaining uncertainty estimates for structured prediction tasks and provides baselines for token-level error detection, sequence-level prediction rejection, and sequence-level out-of-domain input detection using ensembles of auto-regressive transformer models trained on the WMT’14 English-French and WMT’17 English-German translation and LibriSpeech speech recognition datasets.

1. Introduction

Neural Networks (NNs) have become the dominant approach to addressing computer vision (CV) (Girshick 2015 Simonyan & Zisserman 2015 Villegas et al. 2017) natural language processing (NLP) (Mikolov et al. 2013b,a 2010), machine translation (Bahdanau et al. 2015 Vaswani et al. 2017) and speech recognition (ASR) (Hinton et al. 2012 Hannun et al. 2014) tasks. However, estimating the uncertainty in a NN’s prediction, which enables improved safety in automated decision making (Amodei et al. 2016), is still an area of on-going and active research. Notable progress has recently been made on predictive uncertainty estimation for Deep Learning through the definition of baselines, tasks and metrics (Hendrycks & Gimpel 2016), and the development of Bayesian ensemble-based methods for estimating uncertainty, such as Monte-Carlo Dropout (Gal & Ghahramani 2016), Deep Ensembles (Lakshminarayanan et al. 2017), and Stochastic Weight Averaging Gaussian (Maddox et al. 2019). Uncertainty estimates derived using ensembles have been successfully applied to detecting misclassifications, out-of-distribution inputs and adversarial attacks (Carlini & Wagner 2017 Smith & Gal 2018 Malinin & Gales 2019), and to active learning (Kirsch et al. 2019). Crucially, ensemble approaches allow uncertainty to be decomposed into data uncertainty, which is the intrinsic uncertainty associated with the task, and knowledge uncertainty, which is the model’s uncertainty in the prediction due to a lack of understanding of the data Knowledge uncertainty is particularly useful for tasks such as active learning (Kirsch et al. 2019), adversarial attack detection (Smith & Gal 2018 Malinin & Gales 2019) and out-of-domain input detection (Malinin 2019), where it is necessary to know that the model is exposed to inputs it does not understand.

Despite recent advances, most work on uncertainty estimation using ensembles has focused on unstructured prediction tasks, such as image classification. Meanwhile, uncertainty estimation within a general, probabilistically interpretable ensemble-based framework for structured prediction tasks, such as language modelling, machine translation (MT) and speech recognition (ASR), where models yield a variable-length sequence of related targets \( \{y_1, \ldots, y_L\} \), has received little attention. Previous work has examined and developed bespoke confidence score estimation techniques for each task separately (Evermann & Woodland 2000 Liao & Gales 2007 Ragni et al. 2018 Chen et al. 2017 Koehn 2009 Kumar & Sarawagi 2019). Recently, however, initial investigations into general uncertainty estimation for structured prediction have appeared. The nature of data

---

1Yandex Research, Moscow, Russia 2Department of Engineering, University of Cambridge, UK. Correspondence to: Andrey Malinin <am969@yandex-team.ru>, Mark Gales <mjfg@eng.cam.ac.uk>.

Pre-print - work in progress. Copyright 2020 by the author(s).
Uncertainty in Structured Prediction

Uncertainty for translation tasks was examined in [Ott et al., 2018a]. Estimation of sequence and word-level uncertainty estimates via Monte-Carlo Dropout ensembles has been investigated for machine translation [Xiao et al., 2019; Wang et al., 2019]. However, these works mostly focus on machine translation, consider only a small range of uncertainty measures, provide limited theoretical analysis of their properties and do not make explicit their limitations. Furthermore, to our knowledge, no work has examined confidence score estimation for auto-regressive sequence-to-sequence speech recognition models. This work seeks to address these issues.

The contributions of this work are as follows. Firstly, this work examines uncertainty estimation for structured prediction tasks within a general, probabilistically interpretable ensemble-based framework. We derive information-theoretic measures of both data uncertainty and knowledge uncertainty at both the token level and the sequence level, make explicit the challenges involved and state any assumptions made. Secondly, this work explores the practical challenges associated with obtaining uncertainty estimates for structured predictions tasks and provides performance baselines for token-level and sequence-level error detection, and out-of-domain (OOD) input detection on the WMT’14 English-French and WMT’17 English-German translation datasets and the LibriSpeech ASR dataset.

2. Uncertainty Estimation via Ensembles

In this work we take a Bayesian viewpoint on ensembles, as it yields an elegant probabilistic framework within which interpretable uncertainty estimates can be obtained. The core of the Bayesian approach is to treat the model parameters \( \theta \) as random variables and place a prior \( p(\theta) \) over them to compute a posterior \( p(\theta|D) \) via Bayes’ rule:

\[
p(\theta|D) = \frac{p(D|\theta)p(\theta)}{p(D)}
\]  

However, exact Bayesian inference is intractable for neural networks. It is therefore necessary to consider an explicit or implicit approximation \( q(\theta) \) to the true posterior \( p(\theta|D) \) to generate an ensemble of models. A number of different approaches to generating an ensemble of models have been developed, such as Monte-Carlo Dropout [Gal & Ghahramani, 2016], DeepEnsembles [Lakshminarayanan et al., 2017], and Stochastic Weight Averaging Gaussian (SWAG) [Maddox et al., 2019]. A full overview is available in [Ashukha et al., 2020; Ovadia et al., 2019].

Consider an ensemble of models \( \{P(y|x; \theta^{(m)})\}_{m=1}^{M} \) sampled from an approximate posterior \( q(\theta) \), where each model captures an unstructured mapping \( x \to y \), where \( x \in \mathcal{X} \) and \( y \in \mathcal{Y} \). Each of the models \( P(y|x; \theta^{(m)}) \) yields a different estimate of data uncertainty. Uncertainty in predictions due to knowledge uncertainty is expressed as the level of spread, or ‘disagreement’, of models in the ensemble [Malinin, 2019]. The predictive posterior of the ensemble is obtained by taking the expectation with respect to the models in the ensemble:

\[
P(y|x, D) = \mathbb{E}_{p(\theta|D)}[P(y|x; \theta)]
\]

\[
\approx \frac{1}{M} \sum_{m=1}^{M} P(y|x; \theta^{(m)}), \theta^{(m)} \sim q(\theta)
\]  

The entropy of the predictive posterior will be an estimate of total uncertainty in predictions:

\[
\mathcal{H}[P(y|x, D)] = \mathbb{E}_{p(y|x, D)}[-\ln P(y|x, D)]
\]  

Total uncertainty in the prediction is due to both data uncertainty and knowledge uncertainty. However, for tasks such as misclassification detection considering the log-likelihood score assigned by the predictive posterior to a particular class can yield superior performance as it is more sensitive to the prediction made [Malinin, 2019]:

\[
SCR = \ln P(y = \hat{y}|x, D)
\]  

In certain situations, such as active learning [Kirsch et al., 2019] and out-of-distribution input detection, it is desirable to evaluate uncertainty in predictions due to knowledge uncertainty. The sources of uncertainty can be decomposed by considering the mutual information between the model parameters \( \theta \) and the prediction \( y \) [Depeweg et al., 2017]:

\[
\mathcal{I}[y, \theta|x, D] = \mathcal{H}[P(y|x, \theta)] - \mathbb{E}_{q(\theta)}[\mathcal{H}[P(y|x; \theta)]]
\]

\[
= \mathcal{I}[y, \theta|x, D] - \mathbb{E}_{q(\theta)}[\mathcal{H}[P(y|x; \theta)]]
\]

This is expressed as the difference between the entropy of the predictive posterior and the expected entropy of each model in the ensemble. The former is a measure of total uncertainty, while the latter is a measure of expected data uncertainty. Their difference will be measure of spread of the ensemble and an estimate of knowledge uncertainty. It is also possible to consider the expected pairwise KL-divergence between models in an ensemble as an alternative measure of ensemble diversity:

\[
\mathcal{K}[y, \theta, \tilde{\theta}] = \mathbb{E}_{q(\theta)q(\tilde{\theta})}[KL[P(y|x; \theta)||P(y|x, \tilde{\theta})]]
\]

\[
= \sum_{y} \mathbb{E}_{q(\theta)}[P(y|x; \theta)]\mathbb{E}_{q(\tilde{\theta})}[-\ln P(y|x, \tilde{\theta})]
\]

\[
- \mathbb{E}_{q(\theta)}[\mathcal{H}[P(y|x; \theta)]]
\]

\[
= \mathbb{E}_{q(\theta)}[\mathcal{H}[P(y|x; \theta)]]
\]

where \( q(\theta) = q(\tilde{\theta}) \). This measure is an upper bound on the mutual information and also allows total uncertainty
Uncertainty in Structured Prediction

3. Uncertainty for Structured Prediction

In this section the probabilistic framework for uncertainty estimation, discussed in the previous section, is applied to structured prediction. Let’s consider models that capture a mapping \( \{x_1, \ldots, x_T\} \to \{y_1, \ldots, y_L\} \) between a \( T \)-length sequence of inputs \( \{x_1, \ldots, x_T\} = X \) and an \( L \)-length sequence of targets \( \{y_1, \ldots, y_L\} = y \). In this work we consider models where the targets \( y \) are auto-regressive:

\[
P(y|X, \theta) = P(y_1|X; \theta) \prod_{l=2}^{L} P(y_l|y_{<l}, X, \theta) \quad (7)
\]

In the model above the distribution over each \( y_l \) is conditioned on all the previous \( y_{<l} = \{y_1, \ldots, y_{l-1}\} \), which we shall refer to as the context. This formulation describes all neural machine translation (Bahdanau et al., 2015; Vaswani et al., 2017), end-to-end speech recognition (Chan et al., 2015) and other related tasks. For these models we will examine uncertainty estimation at two levels - the token level, which considers uncertainty in the prediction of a single \( y_l \) given a context \( y_{<l} \) and input \( X \), and the sequence level, which considers the uncertainty of predicting the entire sequence \( y \) given \( X \). For the remainder of this work we omit explicitly conditioning on \( X \) for conciseness of notation.

3.1. Token-level uncertainty estimates

Token level uncertainty estimates for models described by equation [7] are a direct extension of uncertainty measures for unstructured predictions models discussed in section [2]. By considering the mutual information between \( y_l \) and \( \theta \) we can obtain measures of total uncertainty, knowledge uncertainty and expected data uncertainty:

\[
\begin{align*}
\mathbb{I} \left[ y_l, \theta | y_{<l}, D \right] &= \mathcal{H} \left[ P(y_l | y_{<l}, D) \right] - \mathbb{E}_{q(\theta)} \left[ \mathcal{H} \left[ P(y_l | y_{<l}; \theta) \right] \right] \\
&= \text{Total Uncertainty} - \text{Expected Data Uncertainty}
\end{align*}
\]

The expected pair-wise KL-divergence between models in the ensemble at the token level can also be considered:

\[
\mathcal{K}[y_l, \theta | y_{<l}] = \mathbb{E}_{q(\theta) q(\bar{\theta})} \left[ \text{KL}[P(y_l | y_{<l}; \theta) || P(y_l | y_{<l}; \bar{\theta})] \right] \\
(9)
\]

where \( q(\theta) = q(\bar{\theta}) \). As before, all expectation are taken via Monte-Carlo approximations. For auto-regressive models, these measures represent uncertainty in the prediction given a specific combination of input \( X \) and context \( y_{<l} \). However, at the beginning of a sequence token-level measures of uncertainty are more sensitive to the input \( X \) and at the end of a sequence become more sensitive to the context \( y_{<l} \).

3.2. Sequence-level uncertainty estimates

Having discussed token-level measures of uncertainty, we now examine sequence-level measures of uncertainty obtained via ensemble approaches. Consider that while there is only one way to express the predictive posterior \( P(y|y_{<l}, D) \) at the token-level, the sequence-level predictive posterior \( P(y|D) \) for an auto-regressive model can be expressed in two ways. The first is as an expectation-of-products:

\[
P(y|D) = \mathbb{E}_{q(\theta)} \left[ P(y|\theta) \right] \\
= \mathbb{E}_{q(\theta)} \left[ P(y_1|\theta) \prod_{l=2}^{L} P(y_l|y_{<l}, \theta) \right] \\
(10)
\]

The other is as a product-of-expectations:

\[
P(y|D) = P(y_1|D) \prod_{l=2}^{L} P(y_l|y_{<l}, D) \\
= \mathbb{E}_{q(\theta)} \left[ P(y_1|\theta) \prod_{l=2}^{L} \mathbb{E}_{q(\theta)} \left[ P(y_l|y_{<l}, \theta) \right] \right] \\
(11)
\]

Both are valid ways to do Bayesian model combination.\(^2\)

However, the optimal \( q(\theta) \) will depend on the choice of combination approach. Consider the \( q(\theta) \) which maximizes the log-likelihood of a dataset \( D = \{y^{(i)}, X^{(i)}\}_{i=1}^{N} \) under \( P(y|D) \) in the case of both decompositions:

\[
\begin{align*}
q(\theta) &= \arg \max_{q(\theta)} E_D \left[ \ln \mathbb{E}_{q(\theta)} \left[ \prod_{l=1}^{L} P(y_l|y_{<l}, \theta) \right] \right] \\
q(\theta) &= \arg \max_{q(\theta)} E_D \left[ \ln \prod_{l=1}^{L} \mathbb{E}_{q(\theta)} \left[ P(y_l|y_{<l}, \theta) \right] \right] \\
(12)
\end{align*}
\]

In the first case, the optimal \( q(\theta) \) will maximize the joint log-probability of all tokens in all sequences in the dataset \( D \), but not necessarily the conditional log-probabilities of all tokens. In the latter case, the optimal \( q(\theta) \) will maximize both the joint and conditional probabilities of all tokens. Thus, \( q(\theta) \) is not in general equal to \( q(\theta) \). This has two implications. Firstly, given a set of samples of model parameters \( \theta^{(m)} \) from a particular \( q(\theta) \), it is not immediately obvious what is the best way to combine models within the ensemble. Secondly, the choice of ensemble combination will affect the severity of any approximations involved in obtaining the measures of uncertainty.

\(^2\)In the current Fairseq (Ott et al., 2019) implementation ensembles are combined as a product-of-expectations.
Having discussed the subtleties of ensemble combination for auto-regressive models, we can now consider how to obtain ensemble-based measures of total uncertainty, data uncertainty and knowledge uncertainty at the sequence level. Let’s begin by examining the mutual information between the model parameters $\theta$ and the sequence $y$:

$$\mathcal{I}[y, \theta | D] = \mathcal{H}[P(y | D)] - \mathbb{E}_{q(\theta)} [\mathcal{H}[P(y | \theta)]]$$ (13)

Evaluating this expression is more difficult than at the token level. First, let’s consider the expression for expected data uncertainty, which is simply the average entropy of each model in the ensemble:

$$\mathcal{H}[P(y | \theta)] = - \sum_y P(y | \theta) \ln P(y | \theta)$$

$$= - \sum_{y_1} \cdots \sum_{y_L} P(y | \theta) \ln P(y | \theta)$$ (14)

However, evaluating the entropy of an auto-regressive model is intractable due to a combinatorial explosion of the hypothesis space - there are a total of $|K|^L$ possible L-length sequences, where $K$ is the vocabulary size, and it is now necessary to do a forward-pass through the network for each hypothesis, for all $M$ models. Clearly, it is necessary to use some kind of approximation to make this tractable.

One simple, but crude, approximation is to approximate sequence-level entropy as a sum of token-level entropies:

$$\mathcal{H}[P(y | \theta)] \approx \sum_{l=1}^L \mathcal{H}[P(y_l | y_{<l}, \theta)]$$ (15)

Thus, expected data uncertainty can be approximated as:

$$\mathbb{E}_{q(\theta)} [\mathcal{H}[P(y | \theta)]] \approx \sum_{l=1}^L \mathbb{E}_{q(\theta)} [\mathcal{H}[P(y_l | y_{<l}, \theta)]]$$ (16)

Now let’s consider how to obtain an estimate of total uncertainty. As discussed at the beginning of this section, the predictive posterior $P(y | D)$ can be expressed either as an expectation-of-products or a product-of-expectations. In either case, evaluating the entropy is intractable, as that would require $M |K|^L$ model evaluations. However, if we consider the product-of-expectations ensemble combination, then total uncertainty can also be approximated as a sum of token-level total-uncertainty via in equation 15:

$$\mathcal{H}[P(y | D)] \approx \sum_{l=1}^L \mathcal{H}[\mathbb{E}_{q(\theta)} [P(y_l | y_{<l}, \theta)]]$$ (17)

Thus, sequence-level mutual information can be approximated as a sum of token-level mutual information:

$$\mathcal{I}[y, \theta | D] \approx \sum_{l=1}^L \mathcal{I}[y_l, \theta | y_{<l}, D]$$ (18)

Similarly, sequence-level EPKL can also be expressed as a sum of token-level EPKL:

$$\mathcal{K}[y, \theta] \approx \sum_{l=1}^L \mathcal{K}[y_l, \theta | y_{<l}]$$ (19)

It is important to state that all of these approximations will in fact be exact under two conditions. Firstly, the distribution $q(\theta)$ should be such that the sequence-level predictive posterior $P(y | D)$ is optimally combined as a product-of-expectations. Secondly, the posterior distributions over each token are independent of the context $y_{<l}$ given the input $X$:

$$P(y_l | y_{<l}, \theta) = P(y_l | \theta)$$ (20)

The first condition only affects the approximations for total uncertainty and mutual information, as they are defined in terms of $P(y | D)$, making them sensitive to how the predictive posterior is obtained. In contrast, EPKL is more robust to the nature of $q(\theta)$, and therefore a potentially more reliable estimate of knowledge uncertainty. Due to length constraints, the derivations of results are provided in appendix D.

While the first requirement for these approximations to be exact can, in principle, be satisfied, the second one almost surely cannot. This begs the question of whether it is possible to obtain alternative measures of uncertainty which do not depend on such strong assumptions. An alternative which can be examined is to derive measures of uncertainty based on the joint probability $P(y = \hat{\omega} | \theta)$ of a particular sequence of tokens $\hat{\omega}$. This is one of the approaches examined in (Xiao et al., 2019). Within this approach we can examine the negative log-likelihood score of a hypothesis under the ensemble predictive posterior, and the expected negative log-likelihood score across the ensemble:

$$\text{SCR} = - \ln \text{P}(\hat{\omega} | D)$$

$$\text{ESCR} = \mathbb{E}_{q(\theta)} [ - \ln \text{P}(\hat{\omega} | \theta) ]$$ (21)

Note, that the value of SCR will depend on whether the joint predictive posterior is obtained via an product-of-expectations (SCR-EP) or an expectation-of-products (SCRP-PE). The difference between the negative log-likelihood of the predictive posterior and the negative-log-likelihood of a particular model is the Point-wise Mutual Information:

$$\text{PMI}(\theta) = - \ln \text{P}(\hat{\omega} | \theta) + \ln \text{P}(\hat{\omega} | D)$$ (22)

This is an information-theoretic measure of the correlation of this particular combination of sequence $\hat{\omega}$ and the parameters $\theta$ (Murphy, 2012). As PMI is defined in terms of $P(y | D)$, it’s value will depend on whether $P(y | D)$ is obtained via an product-of-expectations (PMI-EP) or an expectation-of-products (PMI-PE). Note, that mutual information is the expectation of PMI across all possible hypotheses and model parameters. However, as we do not
have access to all possible hypotheses, the best which can be done is to take the expectation of PMI across all models, which yields a strictly non-positive measure EPMI:

\[
EPMI = -\ln P(\hat{\omega}|D) - \mathbb{E}_q(\theta)\left[-\ln P(\hat{\omega}|\theta)\right] \leq 0 \quad (23)
\]

EPMI will be zero if all the models yield identical joint log-likelihoods of a particular sequence, and will be negative if the models disagree. This way EPMI is a measure of ensemble diversity and, therefore, knowledge uncertainty. However, while the measures of uncertainty considered above do not make any conditional-independence assumptions, they only examine the probabilities assigned to a particular output sequence \(\hat{\omega}\) by models in the ensemble. Thus, their main advantage is that they are only sensitive to how well models in the ensemble agree on the particular hypothesis \(\hat{\omega}\). Their limitation is that they yield less information about the overall uncertainty in a prediction for an input \(X\).

Finally, it is important to point out that all sequence-level measures of uncertainty considered must be normalized by sequence length in order to be able to compare uncertainties of sequences of different lengths.

### 4. Experimental Evaluation

Having discussed ensemble-based measures of uncertainty for structured-prediction tasks, we now provide performance baselines on three applications of these uncertainty estimates: sequence-level and token-level error detection, and out-of-distribution input detection. In this work uncertainty estimates are obtained for ensembles of auto-regressive sequence-to-sequence neural machine translation (NMT) and speech recognition (ASR) models. Ensembles of 6 transformer-big [Vaswani et al., 2017] models were trained on the WMT ’17 English-to-German and WMT ’14 English-to-French translation tasks in both EN-DE/DE-EN and EN-FR/FR-EN directions, and evaluated on the newstest14 dataset. All models were trained using the same baseline configuration described in [Ott et al., 2018b]. For ASR, ensembles of 4 VGG-Transformer [Mohamed et al., 2019] models were trained on the LibriSpeech [Panayotov et al., 2015] dataset. Only ensembles of models generated by training multiple (4/6) identical models starting from different random initialization [Lakshminarayanan et al., 2017] are considered, as it was shown in [Ashukha et al., 2020]

| Model | Libr-DC | Libr-DO | Libr-TC | Libr-TO |
|-------|---------|---------|---------|---------|
| Single | 5.4 ±0.3 | 14.6 ±0.2 | 5.6 ±0.2 | 14.7 ±0.5 |
| PrEx   | 4.2     | 11.0    | 4.3     | 11.3    |
| ExPr   | 4.5     | 12.3    | 4.5     | 12.5    |

Ovadia et al., 2019; Fort et al., 2019) that this approach consistently outperforms other ensemble generation techniques using exponentially fewer ensemble-members. All models are implemented using the publicly available Fairseq [Ott et al., 2019] toolkit using standard configurations with no modification. Further details of model configurations are available in appendix A. Finally, in this work no comparison is made to bespoke uncertainty estimation techniques for each area, such as those described in [Liao & Gales, 2007; Koehn, 2009], because, to our knowledge, they have not been applied to auto-regressive sequence-to-sequence models. While a comparison with these techniques is interesting, adapting them to the models considered here is beyond the scope of this work.

### 4.1. Choice of Ensemble Combination

| ENSM | Libr-DC | Libr-DO | Libr-TC | Libr-TO |
|------|---------|---------|---------|---------|
| PrEx | 0.172   | 0.480   | 0.209   | 0.501   |
| ExPr | 0.194   | 0.576   | 0.236   | 0.606   |

As discussed at the start of section 5 ensembles can be combined as an expectation-of-products (ExPr) or as a product-of-expectations (PrEx), described in equations 10 and 11 respectively. Thus, it is first necessary to establish which combination strategy is better given the ensemble generation method considered here. Tables 1 and 2 show that a product-of-expectations combination yields consistently higher translation BLEU [Post, 2018] and lower ASR word-error-rate (WER) in beam-search decoding for all tasks. Here, NMT models use a beam-width of 5 and ASR models a beam-width of 20. Similarly, tables 3 and 4 show that the product-of-expectations yields lower average length-normalized negative-log-likelihood on reference data (teacher-forcing) for all translation and speech recognition tasks considered. This means that the reference sentences
are more likely under a product-of-expectations ensemble combination.

These results suggest that the ensembles of NMT and ASR models considered in this work are best combined as a product-of-expectations. We speculate that this is because product-of-expectations combination is consistent with how a single auto-regressive NMT/ASR model is trained via teacher-forcing, where the probability of the next reference token $y_j$ is maximized given the reference context $y_{<j}$ and input $X$. Thus, the ‘atomic’ model is in fact a token-level prediction conditioned on the context and input, rather than a sequence-level prediction given the input. The product-of-expectations ensemble combination is consistent with combining these ‘atomic’ models, while the expectation-of-products combination is not. Thus, all further NMT/ASR experiments in this work will use product-of-expectations ensemble combination both for Beam-Search decoding and for calculation of uncertainty measures.

### 4.2. Sequence-level Error Detection

In this section we investigate whether the measures of uncertainty derived in section 3 can be used to detect sentences which are challenging to translate or transcribe using an NMT or ASR model, respectively. In the following experiment a model’s hypotheses are sorted in order of decreasing uncertainty and incrementally replaced by the references. The mean sentence-BLEU or sentence-WER is plotted against the fraction of data replaced on a rejection curve. If the uncertainty estimates are informative, then the increase in BLEU or decrease in WER should be greater than random (linear). Rejection curves are summarised using the Prediction Rejection Ratio (PRR) (Malinin, 2019; Malinin et al., 2020), which is 100% if the measures of uncertainty perfectly correlate with sentence level measure of ‘error’ (BLEU/WER), and 0% if they are completely uninformative. If PRR is negative, then the measures of uncertainty are inversely related to sentence-level ‘error’.

The results provided in table 5 and the associated rejection curves are available in appendix B. They show that challenging sentences can be successfully rejected early, and that measures of total uncertainty yield the best rejection performance on all tasks. This is because it does not matter why a sentence in difficult to translate/transcribe, only that it is. On LibriSpeech test-clean (Libr-TC) and test-other (Libr-TO) the entropy-based (TU) and score-based (SCR-PE) estimates of total uncertainty, given in equations 17 and 21, yield comparable rejection performance. However, on NMT tasks score-based measures do better. This is because compared to ASR, NMT is a task with higher intrinsic uncertainty. In higher uncertainty conditions entropy-based measures are affected by the probability-mass allocated to non-generated tokens, while score-based measures of total uncertainty are not and yield better performance. This is consistent with results on misclassification detection and prediction rejection for unstructed-prediction (Malinin, 2019).

Additionally, the results show that uncertainty-based rejection works better for ASR models than for NMT models, as they achieve a higher PRR. This indicates that these uncertainty estimates are more strongly correlated with sentence-WER and than with sentence-BLEU. However, the issue is that while sentence-WER is an objective measure of transcription quality, sentence-BLEU is only a proxy measure of translation quality. Specifically, while a high sentence BLEU indicates a good translation, and low BLEU does not necessarily indicate a poor one. Thus, a model may yield a low uncertainty, high-quality translation which has little word-overlap with the reference and low sentence-BLEU. This will negatively impact the PRR. However, the issue lies in the nature of machine translation itself - it is inherently difficult to objectively say what is a bad translation\footnote{Provided that the model is high-performing in general.}. A better, but more expensive, approach to assess the quality of uncertainty estimates in machine translation is whether they correlate well with human assessment of translation quality.

### 4.3. Token-level Error Detection

Having discussed sequence-level error detection, let’s now consider token-level error detection. Here, the goal is to use measures of uncertainty to detect errors at the level of BPE-
Uncertainty in Structured Prediction

| Task        | Test Data | Total Uncertainty | Data Uncertainty | Knowledge Uncertainty | % Error |
|-------------|-----------|-------------------|------------------|-----------------------|---------|
| LibriSpeech | Libr-TC   | 36.8              | 37.2             | MI: 33.9              | 3.9     |
|            | Libr-TO   | **44.1**          | **43.5**         | EPKL: 37.8            | 10.3    |

Table 6. Token-level Error Detection %AUPR for LibriSpeech in Beam-Search Decoding regime.

tokens (Sennrich et al., 2015) in the models’ hypotheses. Ground-truth error-labels are obtained by aligning the hypotheses to the references using the SCLITE NIST scoring tool and marking insertions and substitutions. The goal is to detect whether a BPE token is incorrect based on token-level estimates of uncertainty. Performance is assessed via area-under a Precision-Recall curve. Random performance corresponds to the baseline recall, which is equal to the total number of errors present in the dataset. It is important to stress that token-level error labelling by aligning to a reference is appropriate only for ASR, not machine translation. In machine translation many correct tokens will be incorrectly labelled as errors due various valid re-arrangements and word substitutions which can occur. A more meaningful approach would be to use manual human labelling. Due to a lack of human-provided error labels we do not investigate token-level error detection for NMT.

Results in table 6 show that measures of token-level uncertainty can be successfully used for error detection. Crucially, measures of total uncertainty are more informative than measures of data uncertainty or knowledge uncertainty. This is expected, as it is irrelevant why an error was made, as long it as it was made. Score-based and entropy-based measures yield comparable performance on the Librispeech test-clean and test-other datasets, as the test datasets are matched to the training data and the model operates in a low-uncertainty regime. These results are consistent with the previous section.

4.4. Out-of-Distribution input Detection

Having examined the tasks of sequence-level and token-level error detection, let’s now consider OOD input detection. The goal is to use measures of uncertainty to discriminate between in-domain test data and an out-of-distribution dataset. Performance is assessed via area under a ROC-curve (ROC-AUC), where 100% is ideal performance and 50% is random. Anything below 50% indicates the perversive situation where the model yields lower uncertainty for the OOD data than for in-domain data and above 50% performance can be achieved by negating the uncertainty measure. Results are presented in table 7.

First, let’s examine OOD detection for speech recognition. The results show that it is possible to discriminate between test-clean (WER 4.3%) and test-other (WER 11.3%) LibriSpeech datasets with a ROC-AUC of 77.4%. This indicates that is is possible to use measures of uncertainty to detect sentences which are more difficult to successfully transcribe, which is consistent with results in subsections 4.2 and 4.3.

Secondly, it is possible to discriminate between the LibriSpeech datasets, which consist of books being read, and the AMI meeting transcription dataset (Kraaij et al., 2005), which is from a different domain and mismatched to LibriSpeech, with an ROC-AUC of 92.4-97.5%. Clearly, the proposed measures of uncertainty can easily detect when the speech has a very different domain. Finally, we show that it is possible to near-perfectly discriminate between LibriSpeech data and speech spoken in French or Russian taken from the Common Voice Project (Ardila et al., 2019).

Now let’s consider OOD detection for WMT’17 English-German machine translation. Results for En-Fr, De-En and Fr-En tell the same story and are available in appendix C. Firstly, it is difficult to discriminate between newestest14 and combined development and test sets of the khresmoi-summary (MED) dataset which was part of the WMT’14 medical translation task. This suggests that the models interpret text from the medical domain to be in-domain. A likely explanation is the presence of medical-domain text in the WMT’17 English-German training data. Secondly, it is clearly possible to discriminate between the news domain and LibriSpeech reference transcriptions, which are OOD both in terms of structure, as spoken English has different structure relative to written English, as well as domain. Thirdly, it is simple to detect OOD sentences where the original text has been corrupted by randomly permuting the order of BPE tokens. However, detection of text from other languages is particularly difficult. The ensemble displays a pathological behaviour where the input BPE tokens are copied to the output with very high confidence. As a result, estimates of both total uncertainty and data uncertainty are lower for the copied-through OOD data than for in-domain. However, estimates of knowledge uncertainty, especially EPKL, suffer less from this effect, and are able to discriminate between the in-domain and OOD data. On the task of OOD detection entropy-based measures of uncertainty, especially knowledge uncertainty, do better than score-based measures. It is likely that for OOD detection measures of overall uncertainty in the prediction are more
informative than measures of uncertainty in predicting a
particular hypothesis.
An interesting trend in the results is that the best OOD
detection performance for ASR is obtained using measures
of total uncertainty or data uncertainty, but for NMT is
obtained using measures of knowledge uncertainty. This
suggests that speech recognition is, in general, a lower
data uncertainty task than speech recognition, as separation
of measures of uncertainty does not bring benefits. This makes
sense, as the ASR model receives an informationally rich
input, and can plausibly detect differences not only in the
nature of the speech, such as the domain or the language, but
also in the background noise, recording conditions and so on.
In contrast, the only information NMT models have access
to the sequence of BPE tokens fed to the encoder. Clearly,
it is far more difficult to assess whether a certain sequence
of input tokens is in-domain or out-of-domain than whether
a rich acoustic signal is in-domain or out-of-domain. This
highlights the value of measures of knowledge uncertainty.

5. Conclusion
Previously, uncertainty estimates obtained via a general,
probabilistically interpretable ensemble-based framework
were examined for unstructured prediction tasks, like image
classification. This work investigated applying the
same ensemble-based uncertainty estimation framework
to auto-regressive structured-prediction tasks. A range of
information-theoretic measures of uncertainty both at the
token level and sequence level was discussed in section[3].
It was shown that while token-level measures of uncertainty
can be derived in the same way as for un-structured pre-
dictions, obtaining exact estimates of sequence-level uncer-
tainty is challenging. Two types of approximations were
proposed. The first approximates sequence-level measures
of uncertainty as sums of token-level measures. The other
considers log-scores of a particular hypothesis to derive
measures like mean point-wise mutual information. In section[4] this work explored the practical challenges associated
with obtaining uncertainty estimates for structured predic-
tions tasks and provided performance baselines for token-
level and sequence-level error detection, and out-of-domain
(OOD) input detection using ensembles of auto-regressive
transformer models trained on the WMT'14 English-French
and WMT'17 English-German translation datasets and the
LibriSpeech ASR dataset. These initial results are promis-
ing and show that ensemble-based measures of uncertainty
are useful for all of the applications considered. Estimates of
knowledge uncertainty were especially valuable for transla-
tion OOD detection. Additionally, many properties observed
on unstructured prediction tasks carry over to structured pre-
dictions. However, additional properties were also observed.
For example, while ensembles of ASR models are robust
to OOD inputs, NMT models can display pathological be-
haviours, such as confident copy-though with low ensemble
diversity on text from a different language, adversely af-
flecting the quality of uncertainty estimates. Furthermore,
assessing the quality of uncertainty estimates for NMT via
sequence and token level error detection is challenging due
to the nature of the BLEU metric. Future work should in-
vestigate whether uncertainty estimates correlate well with
human judgements of translation quality, explore better ap-
proximations for sequence-level uncertainty and consider
alternative ensemble generation techniques. Furthermore,
ensemble-based uncertainty estimates should be compared
to the task-specific confidence estimates previously explored
for ASR and NMT. Finally, the practical value of ensemble-
based uncertainty estimates should be evaluated on large-
scale ASR and NMT models used in production.
References

Amodei, D., Olah, C., Steinhardt, J., Christiano, P. F., Schulman, J., and Mané, D. Concrete problems in AI safety. http://arxiv.org/abs/1606.06565, 2016. arXiv: 1606.06565.

Ardila, R., Branson, M., Davis, K., Henretty, M., Kohler, M., Meyer, J., Morais, R., Saunders, L., Tyers, F. M., and Weber, G. Common voice: A massively-multilingual speech corpus. arXiv preprint arXiv:1912.06670, 2019.

Ashukha, A., Lyzhov, A., Molchanov, D., and Vetrov, D. Pitfalls of in-domain uncertainty estimation and ensembling in deep learning. In International Conference on Learning Representations, 2020. URL https://openreview.net/forum?id=BJxI5gHKDr

Bahdanau, D., Cho, K., and Bengio, Y. Neural machine translation by jointly learning to align and translate. In Proc. International Conference on Learning Representations (ICLR), 2015.

Carlini, N. and Wagner, D. A. Adversarial examples are not easily detected: Bypassing ten detection methods. CorRR, 2017. URL http://arxiv.org/abs/1705.07263

Chan, W., Jaitly, N., Le, Q. V., and Vinyals, O. Listen, Attend and Spell. http://arxiv.org/abs/1508.01211 2015. arXiv:1508.01211.

Chen, Z., Zhuang, Y., and Yu, K. Confidence measures for ctc-based phone synchronous decoding. In 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4850–4854. IEEE, 2017.

Depeweg, S., Hernández-Lobato, J. M., Doshi-Velez, F., and Udluft, S. Decomposition of uncertainty for active learning and reliable reinforcement learning in stochastic systems. stat, 1050:11, 2017.

Evermann, G. and Woodland, P. Large vocabulary decoding and confidence estimation using word posterior probabilities. In Proc. of IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP), 2000.

Fort, S., Hu, H., and Lakshminarayanan, B. Deep ensembles: A loss landscape perspective. arXiv preprint arXiv:1912.02757, 2019.

Gal, Y. and Ghahramani, Z. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In Proc. 33rd International Conference on Machine Learning (ICML-16), 2016.

Girshick, R. Fast R-CNN. In Proc. 2015 IEEE International Conference on Computer Vision (ICCV), pp. 1440–1448, 2015.

Hannun, A. Y., Case, C., Casper, J., Catanzaro, B., Diamos, G., Elsen, E., Prenger, R., Satheesh, S., Sengupta, S., Coates, A., and Ng, A. Y. Deep speech: Scaling up end-to-end speech recognition, 2014. URL http://arxiv.org/abs/1412.5567 arXiv:1412.5567.

Hendrycks, D. and Gimpel, K. A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks. http://arxiv.org/abs/1610.02136, 2016. arXiv:1610.02136.

Hinton, G., Deng, L., Yu, D., Dahl, G., rahman Mohamed, A., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T., and Kingsbury, B. Deep neural networks for acoustic modeling in speech recognition. Signal Processing Magazine, 2012.

Kirsch, A., van Amersfoort, J., and Gal, Y. Batchbald: Efficient and diverse batch acquisition for deep bayesian active learning, 2019.

Koehn, P. Statistical machine translation. Cambridge University Press, 2009.

Kraaij, W., Hain, T., Lincoln, M., and Post, W. The ami meeting corpus. 2005.

Kumar, A. and Sarawagi, S. Calibration of encoder decoder models for neural machine translation. arXiv preprint arXiv:1903.00802, 2019.

Lakshminarayanan, B., Pritzel, A., and Blundell, C. Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles. In Proc. Conference on Neural Information Processing Systems (NIPS), 2017.

Liao, H. and Gales, M. J. Uncertainty decoding for noise robust speech recognition. In Proceedings of Interspeech, volume 37. Citeseer, 2007.

Maddox, W., Garipov, T., Izmailov, P., Vetrov, D., and Wilson, A. G. A simple baseline for bayesian uncertainty in deep learning. arXiv preprint arXiv:1902.02476, 2019.

Malinin, A. Uncertainty Estimation in Deep Learning with application to Spoken Language Assessment. PhD thesis, University of Cambridge, 2019.

Malinin, A. and Gales, M. J. Reverse kl-divergence training of prior networks: Improved uncertainty and adversarial robustness. 2019.

Malinin, A., Ragni, A., Gales, M., and Knill, K. Incorporating Uncertainty into Deep Learning for Spoken Language Assessment. In Proc. 55th Annual Meeting of the Association for Computational Linguistics (ACL), 2017.
Uncertainty in Structured Prediction

Malinin, A., Mlodozeniec, B., and Gales, M. J. Ensemble distribution distillation. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=BygSP6Vtvr

Mikolov, T., Karafiát, M., Burget, L., Cernocký, J., and Khudanpur, S. Recurrent Neural Network Based Language Model. In *Proc. INTERSPEECH*, 2010.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. Efficient Estimation of Word Representations in Vector Space. [http://arxiv.org/abs/1301.3781](http://arxiv.org/abs/1301.3781), 2013a. arXiv:1301.3781.

Mikolov, T. et al. Linguistic Regularities in Continuous Space Word Representations. In *Proc. NAACL-HLT*, 2013b.

Mohamed, A., Okhonko, D., and Zettlemoyer, L. Transformers with convolutional context for asr. *arXiv preprint arXiv:1904.11660*, 2019.

Murphy, K. P. *Machine Learning*. The MIT Press, 2012.

Ott, M., Auli, M., Grangier, D., and Ranzato, M. Analyzing uncertainty in neural machine translation. *arXiv preprint arXiv:1803.00047*, 2018a.

Ott, M., Edunov, S., Grangier, D., and Auli, M. Scaling neural machine translation. *arXiv preprint arXiv:1806.00187*, 2018b.

Ott, M., Edunov, S., Baevski, A., Fan, A., Gross, S., Ng, N., Grangier, D., and Auli, M. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*, 2019.

Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., Dillon, J. V., Lakshminarayanan, B., and Snoek, J. Can you trust your model’s uncertainty? evaluating predictive uncertainty under dataset shift. *Advances in Neural Information Processing Systems*, 2019.

Panayotov, V., Chen, G., Povey, D., and Khudanpur, S. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5206–5210. IEEE, 2015.

Post, M. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pp. 186–191, Belgium, Brussels, October 2018. Association for Computational Linguistics. URL [https://www.aclweb.org/anthology/W18-6319](https://www.aclweb.org/anthology/W18-6319).

Ragni, A., Li, Q., Gales, M. J., and Wang, Y. Confidence estimation and deletion prediction using bidirectional recurrent neural networks. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pp. 204–211. IEEE, 2018.

Sennrich, R., Haddow, B., and Birch, A. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*, 2015.

Simonyan, K. and Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *Proc. International Conference on Learning Representations (ICLR)*, 2015.

Smith, L. and Gal, Y. Understanding Measures of Uncertainty for Adversarial Example Detection. In *UAI*, 2018.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. Attention is all you need. In *Advances in Neural Information Processing Systems*, pp. 5998–6008, 2017.

Villegas, R., Yang, J., Zou, Y., Sohn, S., Lin, X., and Lee, H. Learning to Generate Long-term Future via Hierarchical Prediction. In *Proc. International Conference on Machine Learning (ICML)*, 2017.

Wang, S., Liu, Y., Wang, C., Luan, H., and Sun, M. Improving back-translation with uncertainty-based confidence estimation. *arXiv preprint arXiv:1909.00157*, 2019.

Xiao, T. Z., Gomez, A. N., and Gal, Y. Wat heb je gezegd? detecting out-of-distribution translations with variational transformers. 2019. URL [http://bayesiandeeplearning.org/2019/papers/90.pdf](http://bayesiandeeplearning.org/2019/papers/90.pdf)
A. Details of Experimental Configuration

The current section of the appendix provides both a description of the datasets and details of the models and experimental setups used in this work.

### A.1. ASR model configuration

| Dataset      | Subset       | Hours | Utterances | Words / Utterance | Domain        |
|--------------|--------------|-------|------------|-------------------|---------------|
| Librispeech  | Train        | 960   | 281.2K     | 33.4              | Story Books   |
|              | Dev-Clean    | 5.4   | 2703       | 17.8              |               |
|              | Dev-Other    | 5.3   | 2864       | 18.9              |               |
|              | Test-Clean   | 5.4   | 2620       | 20.1              |               |
|              | Test-Other   | 5.1   | 2939       | 17.8              |               |
| AMI          | Eval         | -     | 12643      | 7.1               | Meetings      |
| Common-Voice RU | Test      | -     | 6300       | 9.6               |               |
| Common-Voice FR | Test      | -     | 14760      | 9.5               | General       |
| Common-Voice DE | Test     | -     | 13511      | 8.8               |               |

In this work ensembles of the VGG-Transformer sequence-to-sequence ASR model (Mohamed et al., 2019) were considered. An ensemble of 4 models was constructed using a different seed for both initialization and mini-batch shuffling in each model. We used ensembles of only 4 VGG-Transformer models for inference, as more did not fit into the memory of an RTX 2080ti card used for inference. We used the Fairseq (Ott et al., 2019) implementation and training recipe for this model with no modifications. Specifically, models were trained at a fixed learning rate for 80 epochs, where an epoch is a full pass through the entire training set. Checkpoints over the last 30 epochs were averaged together, which proved to be crucial to ensuring good performance. Training took 8 days using 8 V100 GPUs. Models were trained on the full 960 hours of the LibriSpeech dataset (Panayotov et al., 2015) in exactly the same configuration as described in (Mohamed et al., 2019). LibriSpeech is a dataset with 1000 hours of read books encoded in 16-bit, 16kHz FLAC format. The reference transcriptions were BPE tokenized using a vocabulary of 5000 tokens, as per the standard recipe in Fairseq for the VGG-transformer (Ott et al., 2019; Mohamed et al., 2019). For OOD detection we also considered the evaluation subset of the AMI dataset (Kraaij et al., 2005), which is a dataset of meeting transcriptions, as well as the Russian and French datasets of the Common Voice Project (Ardila et al., 2019), which consist of people reading out diverse text from the internet. AMI is encoded in 16-bit, 16Khz WAV format. Common Voice data was stored as 24kHz 32-bit MP3 files which were converted into 16-bit 16kHz WAV format via the SOX tool. Finally, WER was evaluated using the NIST SCLITE scoring tool.

### A.2. NMT model configuration

| Dataset      | Subset       | LNG      | Sentences | Words / Sent. | Domain        |
|--------------|--------------|----------|-----------|---------------|---------------|
| WMT'14 EN-FR | Train        | En Fr    | 40.8M     | 29.2          | Policy, News, Web |
|              |              |          |           | 33.5          |               |
| WMT'17 EN-DE | Train        | En De    | 4.5M      | 26.2          | Policy, News, Web |
|              |              |          |           | 24.8          |               |
| Newstest14   | - En         |          | 3003      | 27.0          | News          |
|              | - Fr         |          |           | 32.1          |               |
|              | - De         |          |           | 28.2          |               |
| Khresmoi-Summary | Dev+Test | En Fr De | 1500      | 19.0          | Medical       |
|              |              |          |           | 21.8          |               |
|              |              |          |           | 17.9          |               |
This work considered ensembles of Transformer-Big (Vaswani et al. 2017) neural machine translation (NMT) models. An ensemble of 6 models was constructed using a different seed for both initialization and mini-batch shuffling in each model. NMT models were trained on the WMT'14 English-French and WMT'17 English-German datasets in both directions (En-Fr, Fr-En, En-De, De-En). All models were trained using the standard Fairseq (Ott et al. 2019) implementation and recipe, which is consistent with the baseline setup in described in (Ott et al. 2018b). The data was tokenized using a BPE vocabulary of 40,000 tokens as per the standard recipe. For each dataset and translation direction an ensemble of 6 models was trained using different random seeds. All 6 models were used during inference. Models trained on WMT'17 English-German were trained for 193,000 steps of gradient descent, which corresponds to roughly 49 epochs, while WMT'14 English-French models were trained for 800,000 steps of gradient descent, which corresponds to roughly 19 epochs. All models were trained using mixed-precision training. Models were evaluated on newstest14, which was treated as in-domain data. OOD data was constructed by considering BPE-token permuted and language-flipped versions of the newstest14 dataset. Furthermore, the khresmoi-summary medical dataset as well the reference transcriptions of the LibriSpeech test-clean and test-other datasets were also used as OOD evaluation datasets. All additional datasets used consistent tokenization using the 40K BPE vocabulary.
B. Sequence-level Error Detection

The current appendix provides a description of the Prediction Rejection Ratio metric, the rejection curves which correspond to results in section 4.2, and histograms of sentence-WER and sentence-BLEU which provide insights into the behaviour of the corresponding rejection curves.

B.1. Prediction Rejection Ratio

Here we describe the Prediction Rejection Ratio metric, proposed in (Malinin, 2019; Malinin et al., 2017), which in this work is used to assess how well measures of sequence-level uncertainty are able to identify sentences which are hard to translate/transcribe. Consider the task of identifying misclassifications - ideally we would like to detect all of the inputs which the model has misclassified based on a measure of uncertainty. Then, the model can either choose to not provide any prediction for these inputs, or they can be passed over or ‘rejected’ to an oracle (ie: human) to obtain the correct prediction (or translation/transcription). The latter process can be visualized using a rejection curve depicted in figure 1, where the predictions of the model are replaced with predictions provided by an oracle in some particular order based on estimates of uncertainty. If the estimates of uncertainty are uninformative, then, in expectation, the rejection curve would be a straight line from base error rate to the lower right corner, given the error metric is a linear function of individual errors. However, if the estimates of uncertainty are ‘perfect’ and always bigger for a misclassification than for a correct classification, then they would produce the ‘oracle’ rejection curve. The ‘oracle’ curve will go down linearly to 0% classification error at the percentage of rejected examples equal to the number of misclassifications. A rejection curve produced by estimates of uncertainty which are not perfect, but still informative, will sit between the ‘random’ and ‘oracle’ curves. The quality of the rejection curve can be assessed by considering the ratio of the area between the ‘uncertainty’ and ‘random’ curves and the area between the ‘oracle’ and ‘random’ curves $AR_{\text{orc}}$ (blue in figure 1), and the area between the ‘oracle’ and ‘random’ curves $AR_{\text{unc}}$ (orange in figure 1). This yields the prediction rejection area ratio $PRR$:

$$PRR = \frac{AR_{\text{unc}}}{AR_{\text{orc}}}$$  \hspace{1cm} (24)

A rejection area ratio of 1.0 indicates optimal rejection, a ratio of 0.0 indicates ‘random’ rejection. A negative rejection ratio indicates that the estimates of uncertainty are ‘perverse’ - they are higher for accurate predictions than for misclassifications. An important property of this performance metric is that it is independent of classification performance, unlike AUPR, and thus it is possible to compare models with different base error rates. Note, that similar approaches to assessing misclassification detection were considered in (Lakshminarayanan et al., 2017; Malinin et al., 2017; Malinin, 2019). In this work instead of considered misclassifications we assess whether measures of uncertainty correlate well with sentence-level BLEU or WER. The overall ‘error’ is then the average of sentence-level BLEU/WER over the test-set.

![Figure 1. Example Prediction Rejection Curves](image-url)
B.2. Rejection Curves

The rejection curves for all NMT models on newstest14 and the ASR model on LibriSpeech test-clean and test-other are presented in figure 2. The main difference between the NMT and ASR curves is that the ‘oracle’ rejection curve for the former is not much better than random, while the rejection curve for the latter is far better than random. This can be
explained by considering the histograms of sentence-level BLEU and sentence-level WER presented in figure 3. Notice, that the sentence-level BLEUs are varied across the spectrum, and very few sentences reach a BLEU of 100. In contrast, 55-75% of all utterances transcribed by the ASR models have a sentence-WER of 0-10%, and then there are a few utterances with a much larger WER. Thus, if the measures of uncertainty can identify the largest errors, which contribute most to the mean WER over the dataset, then a large decrease can be achieved. Hence the shape of the ‘oracle’ WER-rejection curve. In contrast, the contributions from each sentence to mean sentence-BLEU are more evenly spread. Thus, it is difficult to significantly raise the mean-sentence BLEU by rejecting just a few sentences. Hence the shape of the ‘oracle’ BLEU rejection curve for NMT.

Figure 3. Sentence BLEU and WER Histograms.
more similar to the NMT ‘oracle’ rejection curves. This clearly shows that the shape of the oracle curve is not determined by the task (ASR/NMT), but the error (BLEU/WER) distribution across a dataset.

The second trend in the results provided in section 4.2 is that score-based measures of uncertainty work better than entropy-based measures on NMT tasks, while on ASR they perform comparably. The justification provided states that NMT models yield far less confident predictions, and therefore entropy-based measures suffer due to probability mass assigned to other tokens. In contrast, ASR models yield more confident predictions, as shown in figure 5. Notably, on AMI and Common Voice datasets the ASR model also yields less confident predictions, and thus the score-based measures of uncertainty do better than entropy-based ones in the AMI rejection curve in figure 4. These results show that on tasks where it is important to determine which particular translation/transcription hypotheses are worse, score-based measures of uncertainty do as well as or better than entropy-based measures. This result is consistent with confidences being a better measure of uncertainty for misclassification detection in unstructured prediction tasks (Malinin, 2019).
C. Additional Out-of-domain detection results

In the current section additional OOD input detection results are provided for EN-FR, DE-EN and FR-EN neural machine translation models Beam-Search decoding regime and for EN-DE and EN-FR models in a teacher-forcing regime.

| Task ID | Data | Total Uncertainty | Data Uncertainty | Knowledge Uncertainty |
|---------|------|-------------------|------------------|-----------------------|
|         |      | TU    | SCR-PE | DU    | E-SCR | MI    | EPKL  | EPMI-PE |
| MED     |      | 35.3  | 33.2   | 34.7  | 34.2   | 48.4  | 49.4  | 45.0  |
| PRM     |      | 89.7  | 88.3   | 87.0  | 92.6   | 96.9  | 97.1  | 93.7  |
| WMT'17 DE-EN newsen14 | Libri-TC | 20.0  | 21.1   | 18.4  | 29.2   | 61.0  | 67.7  | 67.9  |
|         | Libri-TO | 38.3  | 32.1   | 36.1  | 40.7   | 63.0  | 67.3  | 65.3  |
|         | LNG-EN | 24.6  | 19.6   | 23.0  | 28.1   | 48.7  | 54.4  | 56.5  |
|         | LNG-FR | 40.5  | 34.4   | 38.3  | 43.2   | 63.5  | 67.5  | 65.7  |
| WMT'14 EN-FR newsen14 | MED     | 36.0  | 40.7   | 35.4  | 41.3   | 50.3  | 51.1  | 49.5  |
|         | Libri-TC | 78.8  | 76.9   | 77.5  | 78.5   | 84.3  | 84.0  | 80.5  |
|         | Libri-TO | 76.1  | 75.0   | 74.6  | 76.8   | 84.1  | 83.9  | 80.5  |
|         | PRM     | 95.6  | 92.3   | 94.5  | 95.5   | 99.1  | 99.1  | 98.3  |
|         | LNG-DE | 5.5   | 7.0    | 5.2   | 11.8   | 27.7  | 31.8  | 39.7  |
|         | LNG-FR | 43.0  | 33.8   | 37.7  | 51.3   | 84.9  | 87.1  | 84.6  |
| WMT'14 FR-EN newsen14 | MED     | 22.5  | 21.8   | 22.2  | 22.6   | 36.6  | 37.8  | 35.2  |
|         | PRM     | 96.7  | 94.9   | 96.1  | 96.2   | 98.6  | 98.2  | 95.0  |
|         | Libri-TC | 16.1  | 11.6   | 15.4  | 14.2   | 38.0  | 42.8  | 44.1  |
|         | Libri-TO | 18.8  | 13.8   | 18.0  | 17.1   | 41.3  | 45.5  | 46.8  |
|         | LNG-EN | 9.3   | 5.9    | 8.9   | 8.3    | 24.1  | 27.7  | 31.7  |
|         | LNG-DE | 3.4   | 4.2    | 3.2   | 7.9    | 15.6  | 19.0  | 29.4  |

Table 10 provides OOD detection results for EN-FR, DE-EN and FR-EN NMT models evaluated on Beam-Search derived hypotheses. The results show that detection of medical data from the khresmoi-summary dataset is challenging. However, this is a result of medical-domain data being part of the WMT-17 EN-DE and WMT-14 EN-FR datasets. Detecting permuted inputs is very easy, especially for models trained on WMT-14 EN-FR. This is likely a result of the models having a much better understanding of what corresponds to a ‘natural’ word-order due to having access to more training data. In contrast, detection of inputs in a different language leads models to go into a pathological ‘copy-through’ regime, where they very confidently copy the input to the output. Crucially, measures of knowledge uncertainty are more robust to this effect, but are also adversely affected, especially when trained on WMT-14 EN-FR. A likely explanation for copy-through is that the training data contains examples of matching source and target sentences. Since WMT-14 EN-FR is almost 10 times bigger than WMT-17 EN-DE, it stands to reason that there are more examples of matching source-target sentence pairs, leading these models to be more severely affected. Note, that since the ROC-AUC for many copy-through sentences is less than 10%, it is actually possible to detect this situation accurately using a detection rule which is sensitive to measures of uncertainty being significantly less than for an in-domain dataset.

Table 11 provides OOD detection results for EN-DE/EN-FR translation models evaluated in a teacher-forcing regime, where ‘references’ are fed into the decoder. The results are consistent with those presented in the table above and in the main paper. More insight can be obtained for detection of OOD languages. When the model is fed identical sequences to both the encoder and decoder (LNG-FRFR, LNG-ENEN, LNG-DEDE), then there is a significant copy-through effect which severely compromises measure of total uncertainty and data uncertainty, but not measures of knowledge uncertainty. Notably, entropy-based measures of knowledge uncertainty are more robust than score-based ones. However, OOD languages are easily detectable when copy-through is forcibly avoided (LNG-DEEN, LNG-FREN). Thus, the biggest obstacle to achieving reliable uncertainty estimates using ensemble approaches for NMT is copy-through. As copy-through seems to be an unavoidable, and perhaps even desirable, property of NMT systems, future work needs to investigate the derivation of measures of knowledge uncertainty which are robust to this effect.
Uncertainty in Structured Prediction

Table 11. OOD Detection % ROC-AUC in Teacher-Forcing regime

| Task ID Data | OOD Data | Total Uncertainty | Data Uncertainty | Knowledge Uncertainty |
|--------------|----------|-------------------|------------------|-----------------------|
|              |          | TU SCR-PE         | DU E-SCR         | MI EPMKL EPMI-PE      |
| MED          |          | 48.1 42.4         | 47.0 43.3        | 57.9 58.1 52.9       |
| WMT’17 EN-DE| newsest14| PRM-ENDE 99.8     | PRM-EN 98.6      | PRM-DEN 98.6         |
|              |          | 100.0 100.0       | 100.0 100.0      | 100.0 100.0          |
|              |          | 99.7 99.2         | 99.9 99.7        | 99.4 99.2            |
|              |          | 58.1              | 52.9             | 58.1                 |
|              |          | MED 32.1 33.0     | MED 31.6 33.4    | MED 43.3             |
| WMT’14 EN-FR| newsest14| PRM-ENFR 100.0    | PRM-EN 99.5      | PRM-FR 99.7          |
|              |          | 100.0 100.0       | 100.0 100.0      | 100.0 100.0          |
|              |          | 100.0 100.0       | 100.0 100.0      | 100.0 100.0          |
|              |          | 99.8 99.8         | 99.7 99.4        | 99.9 99.4            |
|              |          | 99.7 99.7         | 98.7 98.5        | 98.7 98.5            |
|              |          | 99.2 99.1         | 99.2 99.5        | 99.2 99.5            |
|              |          | MED 28.3 12.7     | MED 24.0 16.0    | MED 74.0             |
| LNG-DEDE     |          | 99.5 98.4         | 99.4 98.9        | 99.9 99.9            |
| LNG-ENEN     |          | 99.5 98.4         | 99.4 98.9        | 99.9 99.9            |
| LNG-FREN     |          | 99.5 98.4         | 99.4 98.9        | 99.9 99.9            |
| LNG-FRFR     |          | 28.3 12.7         | 24.0 16.0        | 74.0 78.0            |
| LNG-ENEN     |          | 39.1 40.8         | 35.2 42.4        | 83.3 86.0            |

D. Derivations of Uncertainty Measures

This section of the appendix provides the derivations of results from section 3. First, we provide the result that sequence-level entropy can be approximated as a sum of token-level entropy for a single auto-regressive model defined by equation 7.

Consider the following expression:

$$H[P(y|\theta)] = - \sum_{y \in Y} P(y|\theta) \ln P(y|\theta)$$

$$= - \sum_{y_1=1}^K \cdots \sum_{y_L=1}^K \left\{ \prod_{l=1}^L P(y_l|y_{<l}; \theta) \ln \prod_{l=1}^L P(y_l|y_{<l}; \theta) \right\}$$

(25)

Clearly, this expression is intractable to evaluate, as it would require $K^L$ forward passes though an auto-regressive model. However, we can show that this expression becomes a sum of token-level entropy if we make the following conditional independence assumption:

$$P(y_l|y_{<l}; \theta) \approx P(y_l|\theta)$$

(26)

Thus, expression 25 reduces down to:

$$H[P(y|\theta)] = - \sum_{y_1=1}^K \cdots \sum_{y_L=1}^K \left\{ \prod_{l=1}^L P(y_l|\theta) \ln \prod_{l=1}^L P(y_l|\theta) \right\}$$

$$\approx - \sum_{y_1=1}^K \cdots \sum_{y_L=1}^K \left\{ \prod_{l=1}^L P(y_l|\theta) \ln \prod_{l=1}^L P(y_l|\theta) \right\}$$

$$= - \sum_{y_1=1}^K \cdots \sum_{y_L=1}^K \prod_{l=1}^L P(y_l|\theta) \ln \sum_{y_1=1}^K \prod_{l=1}^L P(y_l|\theta)$$

$$= - \sum_{l=1}^L \sum_{y_l=1}^K P(y_l|\theta) \ln P(y_l|\theta)$$

$$= \sum_{l=1}^L H[P(y_l|\theta)]$$

(27)
Thus, if the token-level posteriors are content-independent, then sequence-level entropy will be exactly the sum of token-level entropies. Obviously, this is a crude approximation, as the token-level posteriors are almost surely never content-independent. Moreover, this is likely undesirable from the point of view of the task which the auto-regressive model is learning. However, as results in this work show, measures of uncertainty based on this approximation tend to work well in practice.

Now let’s show that the entropy of the predictive posterior can be similarly approximated if it is expressed as a product-of-expectations, but not if it is expressed as an expectation-of-products. The former can trivially be shown by replace $\theta$ with $\mathcal{D}$ in equations (25-27). Thus, let’s consider only the latter:

$$
\mathcal{H}[P(y|\mathcal{D})] = -\sum_{y \in \mathcal{Y}} P(y|\mathcal{D}) \ln P(y|\mathcal{D})
$$

$$
= -\sum_{y_1=1}^K \cdots \sum_{y_L=1}^K \left\{ E_{q(\theta)} \left[ \prod_{l=1}^L P(y_l|y_{<l}; \theta) \right] \ln E_{q(\theta)} \left[ \prod_{l=1}^L P(y_l|y_{<l}; \hat{\theta}) \right] \right\}
$$

$$
\approx -\sum_{y_1=1}^K \cdots \sum_{y_L=1}^K \left\{ E_{q(\theta)} \left[ \prod_{l=1}^L P(y_l|\theta) \right] \ln E_{q(\theta)} \left[ \prod_{l=1}^L P(y_l|\hat{\theta}) \right] \right\}
$$

(28)

Clearly, it’s difficult to progress any further in the expression above, as the predictive posterior does not factorize. Thus, it is possible to guarantee that sequence-level entropy of the ensemble predictive posterior can be expressed as a sum of entropy of the token-level predictive posterior only if the ensemble is combined as an expectation of products. However, as the results in section 4.1 show, combining the ensemble as a product-of-expectations is consistently better than an expectation-of-products. Thus, this condition for the approximation to be accurate does hold in practice, unlike the conditional-independence approximation considered above.

Finally, let’s show that sequence-level expected-pairwise KL-divergence (EPKL) can also be expressed as a sum of token-level EPKL and why it is robust to how the predictive-posterior is optimally expressed.

$$
\mathcal{K}[y, \theta] = -\sum_{y \in \mathcal{Y}} E_{q(\theta)} \left[ P(y|\theta) \right] E_{q(\hat{\theta})} \left[ \ln P(y|\hat{\theta}) \right] - E_{q(\theta)} \left[ \mathcal{H}[P(y|\theta)] \right]
$$

$$
= -\sum_{y_1=1}^K \cdots \sum_{y_L=1}^K \left\{ E_{q(\theta)} \left[ \prod_{l=1}^L P(y_l|y_{<l}; \theta) \right] E_{q(\hat{\theta})} \left[ \ln \prod_{l=1}^L P(y_l|y_{<l}; \hat{\theta}) \right] \right\} - E_{q(\theta)} \left[ \mathcal{H}[P(y|\theta)] \right]
$$

$$
\approx -\sum_{y_1=1}^K \cdots \sum_{y_L=1}^K \left\{ E_{q(\theta)} \left[ \prod_{l=1}^L P(y_l|\theta) \right] E_{q(\hat{\theta})} \left[ \ln \prod_{l=1}^L P(y_l|\hat{\theta}) \right] \right\} - E_{q(\theta)} \left[ \sum_{l=1}^L \mathcal{H}[P(y_l|\theta)] \right]
$$

$$
= -\sum_{l=1}^L \sum_{y_l=1}^K E_{q(\theta)} \left[ P(y_l|\theta) \right] E_{q(\hat{\theta})} \left[ \ln P(y_l|\hat{\theta}) \right] - \sum_{l=1}^L E_{q(\theta)} \left[ \mathcal{H}[P(y_l|\theta)] \right]
$$

$$
= \sum_{l=1}^L \mathcal{K}[y_l, \theta]
$$

(29)

Crucially, EPKL is explicitly defined in terms of $E_{q(\theta)} \left[ P(y|\theta) \right]$ and $E_{q(\hat{\theta})} \left[ \ln P(y|\hat{\theta}) \right]$. In contrast, mutual information is defined in terms of $\mathcal{H}[P(y|\mathcal{D})]$, and is therefore sensitive to whether a product-of-expectations is the appropriate approach to combine members of the ensemble, given the choice of ensemble generation method. Thus, EPKL is less sensitive to the nature of the ensemble than mutual information and may yield a more robust measure of uncertainty.

### E. Expectation of Products vs. Product-of-Expectations

The current appendix provides a comparison of score-based measures of total uncertainty and knowledge uncertainty derived using an expectation-of-products and a product-of-expected-products. Specifically, we consider the effect on sequence-level error detection and OOD input detection.
Table 12. Sequence-level Error Detection % Prediction Rejection Ratio in Beam-Search decoding regime.

| Task         | Test set | Total Uncertainty | Knowledge Uncertainty |
|--------------|----------|-------------------|-----------------------|
|              |          | SCR-PE | SCR-EP | EPMI-PE | EPMI-PE |
| LibriSpeech  | Libr-TC  | 64.3   | 63.5   | 58.9    | 57.5    |
|              | Libr-TO  | 67.1   | 65.3   | 60.5    | 54.4    |
| WMT’17 EN-DE |          | 46.7   | 46.4   | 29.1    | 22.5    |
| WMT’17 DE-EN | newstest14 | 44.7   | 44.2   | 31.0    | 26.6    |
| WMT’14 EN-FR |          | 38.9   | 38.7   | 31.5    | 26.2    |
| WMT’14 FR-EN |          | 41.4   | 41.4   | 28.8    | 24.0    |

Results in table 12 show that score-based measures of uncertainty derived using a product-of-expectation consistently yield a higher PRR for sequence-level error detection. OOD detection results in table 13 show that when models do not suffer from copy-through, then measures derived using a product-of-expectations do better, though in a few cases measures obtained via an expectation-of-products can be better (LNG-FR, LNG-DE), though it is hard to judge what is ‘better’ when all ROC-AUC numbers are below 50. Overall, these results show that a product-of-expectation ensemble combination generally yields better measures of uncertainty.

Table 13. OOD Detection % ROC-AUC in Beam-Search decoding regime

| Task         | ID OOD | Total Uncertainty | Knowledge Uncertainty |
|--------------|--------|-------------------|-----------------------|
|              | Data   | SCR-PE | SCR-EP | EPMI-PE | EPMI-EP |
| WMT’17 EN-DE | MED    | 53.3   | 53.4   | 60.7    | 59.1    |
|              | Libri-TC | 72.0   | 71.4   | 72.5    | 69.5    |
|              | Libri-TO | 69.3   | 69.0   | 72.0    | 69.0    |
|              | PRM    | 82.9   | 84.1   | 95.6    | 91.4    |
|              | LNG-DE | 29.5   | 32.4   | 76.1    | 76.8    |
|              | LNG-FR | 21.3   | 21.6   | 71.7    | 76.2    |
| WMT’17 EN-FR | MED    | 40.7   | 40.8   | 49.5    | 49.3    |
|              | Libri-TC | 76.9   | 76.8   | 80.5    | 76.7    |
|              | Libri-TO | 75.0   | 74.8   | 80.5    | 76.7    |
|              | PRM    | 92.3   | 92.5   | 98.3    | 96.4    |
|              | LNG-DE | 7.0    | 7.7    | 39.7    | 46.4    |
|              | LNG-FR | 33.8   | 37.3   | 84.6    | 85.0    |

F. Additional measures of uncertainty

In this work we considered a range of information-theoretic measures of uncertainty. However, in (Xiao et al., 2019) a range of other measures was considered, though their properties were not analyzed in detail. Firstly, (Xiao et al., 2019) considered computing the cross-bleu of an ensemble. Here, the average square of the complement of pairwise sentence-BLEU between 1-best hypotheses $\hat{y}^{(m)}$ produced by each individual model in the ensemble is used as a measure of uncertainty:

$$X\text{-BLEU} = \frac{1}{(M-1)^2} \sum_{m=1}^{M} \sum_{q \neq m} (100 - \text{BLEU}(\hat{y}^{(m)}, \hat{y}^{(q)}))^2$$

Similarly, an equivalent measure of ASR can be derived, which we will call cross WER:

$$X\text{-WER} = \frac{1}{(M-1)^2} \sum_{m=1}^{M} \sum_{q \neq m} (\text{WER}(\hat{y}^{(m)}, \hat{y}^{(q)}))^2$$
These two measure of uncertainty assess the diversity between the 1-best hypotheses of each model in an ensemble. In this way they are conceptually related to measures of knowledge uncertainty. Notably, as they are only sensitive to the 1-best hypothesis, they are ‘hard’ versions of expected point-wise mutual information, which considers the diversity in the scores assigned to the joint-ensemble 1-best hypothesis. However, it is not entirely clear why these measures are expressed as a sum-of-squares. It seems possible to consider a range of different powers, which will affect rank-ordering, though there seems to be little theoretical justification for any particular choice of power. Furthermore, these measures of uncertainty require $M$ separate decodes, unlike all the measures considered in this work, which require only a single combined ensemble decode.

Tables 14 and 15 explore the utility of these uncertainty measures and compare them to the best-performing measures of uncertainty for sequence-level error detection and OOD input detection. Results in tables 14 show that cross-BLEU and cross-WER do consistently worse than measures of total uncertainty for sequence-level error detection. This is expected, as they are measures knowledge uncertainty. Similarly, table 15 also shows that cross-BLEU and cross-WER do consistently worse, with two exceptions, than the measures of knowledge uncertainty considered in this work on the task of OOD input detection. One reason that they do worse than measures like EPMI could be that information is lost in making a hard-decision. Secondly, metrics like sentence-WER and sentence-BLEU poorly capture an appropriate ‘distance’ between hypotheses. In contrast, information theoretic measures, even approximate ones, naturally work within an interpretable probabilistic framework.

### Table 14. Sequence-level Error Detection % Prediction Rejection Ratio in Beam-Search decoding regime.

| Task   | Test set   | SCR-PE | XBLEU | XWER |
|--------|------------|--------|-------|------|
| LibriSpeech | Libr-TC   | 64.3   | 56.9  | 47.9 |
|         | Libr-TO   | 67.1   | 61.7  | 58.4 |
| WMT’17 EN-DE |           | 46.7   | 35.8  | 33.4 |
| WMT’17 DE-EN | newstest14 | 44.7   | 40.1  | 36.7 |
| WMT’14 EN-FR |           | 38.9   | 32.0  | 26.8 |
| WMT’14 FR-EN |           | 41.4   | 34.7  | 31.8 |

### Table 15. OOD Detection % ROC-AUC in Beam-Search decoding regime

| Task         | ID Data | OOD Data | Knowledge Uncertainty | Cross-Measures |
|--------------|---------|----------|-----------------------|----------------|
|              | Libr-TC | Libr-TO  | MI EPKL EPMI-PE       | XBLEU XWER     |
| LibriSpeech  |         |          | 76.7 77.1 73.9        | 74.3 71.8      |
|              | Libr-TC | AMI-EVL  | 95.7 95.4 96.0        | 95.9 95.8      |
|              | Libr-TO | AMI-EVL  | 88.0 86.7 89.0        | 89.5 90.8      |
|              |         | LNG-FR   | 99.9 99.9 99.7        | 99.5 99.7      |
|              |         | LNG-RU   | 99.9 99.9 99.7        | 99.5 99.7      |
| WMT’17 EN-DE | MED     |          | 64.9 65.2 60.7        | 58.5 57.4      |
|              |         | Libr-TC  | 77.1 76.5 72.5        | 65.1 58.2      |
|              |         | Libr-TO  | 76.2 75.9 72.0        | 64.8 56.9      |
|              |         | PERM     | 97.0 97.3 95.6        | 80.8 73.5      |
|              |         | LNG-DE   | 73.2 78.2 76.1        | 36.1 38.7      |
|              |         | LNG-FR   | 57.7 64.7 71.7        | 46.8 52.7      |
| WMT’17 EN-DE | MED     |          | 50.3 51.1 49.5        | 52.4 52.4      |
|              |         | Libr-TC  | 84.3 84.0 80.5        | 71.5 67.9      |
|              |         | Libr-TO  | 84.1 83.9 80.5        | 72.5 68.6      |
|              |         | PRM      | 99.1 99.1 98.3        | 92.3 90.2      |
|              |         | LNG-DE   | 27.7 31.8 39.7        | 26.0 26.2      |
|              |         | LNG-FR   | 84.9 87.1 84.6        | 46.3 40.5      |