Unsupervised Quality Estimation for Neural Machine Translation

Marina Fomicheva,[1] Shuo Sun,[2] Lisa Yankovskaya,[3] Frédéric Blain,[1]
Francisco Guzmán,[5] Mark Fishel,[3] Nikolaos Aletras,[1]
Vishrav Chaudhary,[5] Lucia Specia[1,4]
[1]University of Sheffield [2]Johns Hopkins University [3]University of Tartu
[4]Imperial College London [5]Facebook AI
[1]{m.fomicheva,f.blain,n.aletras,l.specia}@sheffield.ac.uk
[2]{ssun32}@jhu.edu [3]{lisa.yankovskaya,fishel}@ut.ee
[5]{fguzman,vishrav}@fb.com

Abstract

Quality Estimation (QE) is an important component in making Machine Translation (MT) useful in real-world applications, as it is aimed to inform the user on the quality of the MT output at test time. Existing approaches require large amounts of expert annotated data, computation, and time for training. As an alternative, we devise an unsupervised approach to QE where no training or access to additional resources besides the MT system itself is required. Different from most of the current work that treats the MT system as a black box, we explore useful information that can be extracted from the MT system as a by-product of translation. By utilizing methods for uncertainty quantification, we achieve very good correlation with human judgments of quality, rivaling state-of-the-art supervised QE models. To evaluate our approach we collect the first dataset that enables work on both black-box and glass-box approaches to QE.

1 Introduction

With the advent of neural models, Machine Translation (MT) systems have made substantial progress, reportedly achieving near-human quality for high-resource language pairs (Hassan et al., 2018; Barrault et al., 2019). However, translation quality is not consistent across language pairs, domains, and datasets. This is problematic for low-resource scenarios, where there is not enough training data and translation quality significantly lags behind. Additionally, neural MT (NMT) systems can be deceptive to the end user as they can generate fluent translations that differ in meaning from the original (Bentivogli et al., 2016; Castilho et al., 2017). Thus, it is crucial to have a feedback mechanism to inform users about the trustworthiness of a given MT output.

Quality estimation (QE) aims to predict the quality of the output provided by an MT system at test time when no gold-standard human translation is available. State-of-the-art (SOTA) QE models require large amounts of parallel data for pretraining and in-domain translations annotated with quality labels for training (Kim et al., 2017a; Fonseca et al., 2019). However, such large collections of data are only available for a small set of languages in limited domains.

Current work on QE typically treats the MT system as a black box. In this paper we propose an alternative glass-box approach to QE that allows us to address the task as an **unsupervised problem**. We posit that encoder-decoder NMT models (Sutskever et al., 2014; Bahdanau et al., 2015; Vaswani et al., 2017) offer a rich source of information for directly estimating translation quality: (a) the *output probability distribution* from the NMT system (i.e., the probabilities obtained by applying the softmax function over the entire vocabulary of the target language); and (b) the *attention mechanism* used during decoding. Our assumption is that the more confident the decoder is, the higher the quality of the translation.

While sequence-level probabilities of the top MT hypothesis have been used for confidence estimation in statistical MT (Specia et al., 2013; Blatz et al., 2004), the output probabilities from deep Neural Networks (NNs) are generally not well calibrated, that is, not representative of the true likelihood of the predictions (Nguyen and O'Connor, 2015; Guo et al., 2017; Lakshminarayanan et al., 2017). Moreover, softmax output probabilities tend to be *overconfident* and can assign a large

Transactions of the Association for Computational Linguistics, vol. 8, pp. 539–555, 2020. https://doi.org/10.1162/tacl_a_00330
Action Editor: Stefan Riezler. Submission batch: 1/2020; Revision batch: 4/2020; Published 9/2020.
© 2020 Association for Computational Linguistics. Distributed under a CC-BY 4.0 license.

Downloaded from http://www.mitpressjournals.org/doi/pdf/10.1162/tacl_a_00330?cookieSet=1 by guest on 28 March 2024

probability mass to predictions that are far from the training data (Gal and Ghahramani, 2016). To overcome such deficiencies, we propose ways to exploit output distributions beyond the top-1 prediction by exploring *uncertainty quantification* methods for better probability estimates (Gal and Ghahramani, 2016; Lakshminarayanan et al., 2017). In our experiments, we account for different factors that can affect the reliability of model probability estimates in NNs, such as model architecture, training, and search (Guo et al., 2017).

In addition, we study attention mechanism as another source of information on NMT quality. Attention can be interpreted as a soft alignment, providing an indication of the strength of relationship between source and target words (Bahdanau et al., 2015). Although this interpretation is straightforward for NMT based on Recurrent Neural Networks (RNN) (Rikters and Fishel, 2017), its application to current SOTA Transformer models with multihead attention (Vaswani et al., 2017) is challenging. We analyze to what extent meaningful information on translation quality can be extracted from multihead attention.

To evaluate our approach in challenging settings, we collect a new dataset for QE with 6 language pairs representing NMT training in high, medium, and low-resource scenarios. To reduce the chance of overfitting to particular domains, our dataset is constructed from Wikipedia documents. We annotate 10K segments per language pair. By contrast to the vast majority of work on QE that uses semi-automatic metrics based on post-editing distance as gold standard, we perform quality labeling based on the Direct Assessment (DA) methodology (Graham et al., 2015b), which has been widely used for popular MT evaluation campaigns in the recent years. At the same time, the collected data differs from the existing datasets annotated with DA judgments for the well known WMT Metrics task[1] in two important ways: We provide enough data to train supervised QE models and access to the NMT systems used to generate the translations, thus allowing for further exploration of the glass-box unsupervised approach to QE for NMT introduced in this paper.

Our **main contributions** can be summarized as follows: (i) A new, large-scale dataset for sentence-level[2] QE annotated with DA rather than post-edit ing metrics (§4); (ii) A set of unsupervised quality indicators that can be produced as a by-product of NMT decoding and a thorough evaluation of how they correlate with human judgments of translation quality (§3 and §5); (iii) The first attempt at analysing the attention distribution for the purposes of unsupervised QE in Transformer models (§3 and §5); and (iv) The analysis on how model confidence relates to translation quality for different NMT systems (§6). Our experiments show that unsupervised QE indicators obtained from well-calibrated NMT model probabilities rival strong supervised SOTA models in terms of correlation with human judgments.

2 Related Work

QE QE is typically addressed as a supervised machine learning task where the goal is to predict MT quality in the absence of reference translation. Traditional feature-based approaches relied on manually designed features, extracted from the MT system (*glass-box* features) or obtained from the source and translated sentences, as well as external resources, such as monolingual or parallel corpora (*black-box* features) (Specia et al., 2009).

Currently, the best performing approaches to QE use NNs to learn useful representations for source and target sentences (Kim et al., 2017b; Wang et al., 2018; Kepler et al., 2019a). A notable example is the Predictor-Estimator (PredEst) model (Kim et al., 2017b), which consists of an encoder-decoder RNN (*predictor*) trained on parallel data for a word prediction task and a unidirectional RNN (*estimator*) that produces quality estimates leveraging the context representations generated by the predictor. Despite achieving strong performances, neural-based approaches are resource-heavy and require a significant amount of in-domain labeled data for training. They do not use any internal information from the MT system.

Existing work on glass-box QE is limited to features extracted from statistical MT, such as language model probabilities or number of hypotheses in the n-best list (Blatz et al., 2004; Specia et al., 2013). The few approaches for unsupervised QE are also inspired by the work on statistical MT

[1] http://www.statmt.org/wmt19/metrics-task.html.

[2] While the paper covers QE at sentence level, the extension of our unsupervised metrics to word-level QE would be straightforward and we leave it for future work.

Downloaded from http://www.mitpressjournals.org/doi/pdf/10.1162/tacl_a_00330?cookieSet=1 by guest on 28 March 2024

and perform significantly worse than supervised approaches (Popović, 2012; Moreau and Vogel, 2012; Etchegoyhen et al., 2018). For example, Etchegoyhen et al. (2018) use lexical translation probabilities from word alignment models and language model probabilities. Their unsupervised approach averages these features to produce the final score. However, it is largely outperformed by the neural-based supervised QE systems (Specia et al., 2018).

The only works that explore internal information from neural models as an indicator of translation quality rely on the entropy of attention weights in RNN-based NMT systems (Rikters and Fishel, 2017; Yankovskaya et al., 2018). However, attention-based indicators perform competitively only when combined with other QE features in a supervised framework. Furthermore, this approach is not directly applicable to the SOTA Transformer model that uses multihead attention mechanism. Recent work on attention interpretability showed that attention weights in Transformer networks might not be readily interpretable (Vashishth et al., 2019; Vig and Belinkov, 2019). Voita et al. (2019) show that different attention heads of Transformer have different functions and some of them are more important than others. This makes it challenging to extract information from attention weights in Transformer (see §5).

To the best of our knowledge, our work is the first on glass-box unsupervised QE for NMT that performs competitively with respect to the SOTA supervised systems.

QE Datasets The performance of QE systems has been typically assessed using the semi-automatic Human-mediated Translation Edit Rate (Snover et al., 2006) metric as gold standard. However, the reliability of this metric for assessing the performance of QE systems has been shown to be questionable (Graham et al., 2016). The current practice in MT evaluation is the so-called Direct Assessment (DA) of MT quality (Graham et al., 2015b), where raters evaluate the MT on a continuous 1–100 scale. This method has been shown to improve the reproducibility of manual evaluation and to provide a more reliable gold standard for automatic evaluation metrics (Graham et al., 2015a).

DA methodology is currently used for manual evaluation of MT quality at the WMT translation tasks, as well as for assessing the performance of reference-based automatic MT evaluation metrics at the WMT Metrics Task (Bojar et al., 2016, 2017; Ma et al., 2018, 2019). Existing datasets with sentence-level DA judgments from the WMT Metrics Task could in principle be used for benchmarking QE systems. However, they contain only a few hundred segments per language pair and thus hardly allow for training supervised systems, as illustrated by the weak correlation results for QE on DA judgments based on the Metrics Task data recently reported by Fonseca et al. (2019). Furthermore, for each language pair the data contains translations from a number of MT systems often using different architectures, and these MT systems are not readily available, making it impossible for experiments on glass-box QE. Finally, the judgments are either crowd-sourced or collected from task participants and not professional translators, which may hinder the reliability of the labels. We collect a new dataset for QE that addresses these limitations (§4).

Uncertainty Quantification Uncertainty quantification in NNs is typically addressed using a Bayesian framework where the point estimates of their weights are replaced with probability distributions (MacKay, 1992; Graves, 2011; Welling and Teh, 2011; Tran et al., 2019). Various approximations have been developed to avoid high training costs of Bayesian NNs, such as Monte Carlo Dropout (Gal and Ghahramani, 2016) or model ensembling (Lakshminarayanan et al., 2017). The performance of uncertainty quantification methods is commonly evaluated by measuring calibration, that is, the relation between predictive probabilities and the empirical frequencies of the predicted labels, or by assessing generalization of uncertainty under domain shift (see §6).

Only a few studies have analyzed calibration in NMT and they came to contradictory conclusions. Kumar and Sarawagi (2019) measure calibration error by comparing model probabilities and the percentage of times NMT output matches reference translation, and conclude that NMT probabilities are poorly calibrated. However, the calibration error metrics they use are designed for binary classification tasks and cannot be easily transferred to NMT (Kuleshov and Liang, 2015). Ott et al. (2019) analyze uncertainty in NMT by comparing predictive probability distributions with the empirical distribution observed in human translation data. They conclude that NMT models

Downloaded from http://www.mitpressjournals.org/doi/pdf/10.1162/tacl_a_00330?cookieSet=1 by guest on 28 March 2024

are well calibrated. However, this approach is limited by the fact that there are many possible correct translations for a given sentence and only one human translation is available in practice. Although the goal of this paper is to devise an unsupervised solution for the QE task, the analysis presented here provides new insights into calibration in NMT. Different from existing work, we study the relation between model probabilities and human judgments of translation correctness.

Uncertainty quantification methods have been successfully applied to various practical tasks, for example, neural semantic parsing (Dong et al., 2018), hate speech classification (Miok et al., 2019), or back-translation for NMT (Wang et al., 2019). Wang et al. (2019), whose work is the closest to our work, explore a small set of uncertainty-based metrics to minimize the weight of erroneous synthetic sentence pairs for back translation in NMT. However, improved NMT training with weighted synthetic data does not necessarily imply better prediction of MT quality. In fact, metrics that Wang et al. (2019) report to perform the best for back-translation do not perform well for QE (see §3.2).

3 Unsupervised QE for NMT

We assume a sequence-to-sequence NMT architecture consisting of encoder-decoder networks using attention (Bahdanau et al., 2015). The encoder maps the input sequence $\vec{x} = x_1, \ldots, x_I$ into a sequence of hidden states, which is summarized into a single vector using attention mechanism (Bahdanau et al., 2015; Vaswani et al., 2017). Given this representation the decoder generates an output sequence $\vec{y} = y_1, \ldots, y_T$ of length T. The probability of generating $\vec{y}$ is factorized as:

$$p(\vec{y}|\vec{x}, \theta) = \prod_{t=1}^{T} p(y_t|\vec{y}_{<t}, \vec{x}, \theta)$$

where θ represents model parameters. The decoder produces the probability distribution $p(y_t|\vec{y}_{<t}, \vec{x}, \theta)$ over the system vocabulary at each time step using the *softmax function*. The model is trained to minimize cross-entropy loss. We use SOTA Transformers (Vaswani et al., 2017) for the encoder and decoder in our experiments.

In what follows, we propose unsupervised quality indicators based on: (i) output probability distribution obtained either from a standard deterministic NMT (§3.1) or (ii) using uncertainty quantification (§3.2), and (iii) attention weights (§3.3).

3.1 Exploiting the Softmax Distribution

We start by defining a simple QE measure based on sequence-level translation probability normalized by length:

$$\text{TP} = \frac{1}{T} \sum_{t=1}^{T} \log p(y_t|\vec{y}_{<t}, \vec{x}, \theta)$$

However, 1-best probability estimates from the softmax output distribution may tend towards overconfidence, which would result in high probability for unreliable MT outputs. We propose two metrics that exploit output probability distribution beyond the average of top-1 predictions. First, we compute the entropy of softmax output distribution over target vocabulary of size V at each decoding step and take an average to obtain a sentence-level measure:

$$\text{Softmax-Ent} = -\frac{1}{T} \sum_{t=1}^{T} \sum_{v=1}^{V} p(y_t^v) \log p(y_t^v)$$

where $p(y_t)$ represents the conditional distribution $p(y_t|\vec{x}, \vec{y}_{<t}, \theta)$.

If most of the probability mass is concentrated on a few vocabulary words, the generated target word is likely to be correct. By contrast, if softmax probabilities approach a uniform distribution picking any word from the vocabulary is equally likely and the quality of the resulting translation is expected to be low.

Second, we hypothesize that the dispersion of probabilities of individual words might provide useful information that is inevitably lost when taking an average. Consider, as an illustration, that the sequences of word probabilities [0.1, 0.9] and [0.5, 0.5] have the same mean, but might indicate very different behavior of the NMT system, and consequently, different output quality. To formalize this intuition we compute the standard deviation of word-level log-probabilities,

$$\text{Sent-Std} = \sqrt{\mathbb{E}[\text{P}^2] - (\mathbb{E}[\text{P}])^2}$$

where $\text{P} = p(y_1), \ldots, p(y_T)$ represents word-level log-probabilities for a given sentence.

Downloaded from http://www.mitpressjournals.org/doi/pdf/10.1162/tacl_a_00330?cookieSet=1 by guest on 28 March 2024

3.2 Quantifying Uncertainty

It has been argued in recent work that deep neural networks do not properly represent model uncertainty (Gal and Ghahramani, 2016; Lakshminarayanan et al., 2017). Uncertainty quantification in deep learning typically relies on the Bayesian formalism (MacKay, 1992; Graves, 2011; Welling and Teh, 2011; Gal and Ghahramani, 2016; Tran et al., 2019). Bayesian NNs learn a posterior distribution over parameters that quantifies *model* or *epistemic* uncertainty, i.e., our lack of knowledge as to which model generated the training data.[3] Bayesian NNs usually come with prohibitive computational costs and various approximations have been developed to alleviate this. In this paper we explore the **Monte Carlo (MC) dropout** (Gal and Ghahramani, 2016).

Dropout is a method introduced by Srivastava et al. (2014) to reduce overfitting when training neural models. It consists in randomly masking neurons to zero based on a Bernoulli distribution. Gal and Ghahramani (2016) use dropout at test time before every weight layer. They perform several forward passes through the network and collect posterior probabilities generated by the model with parameters perturbed by dropout. Mean and variance of the resulting distribution can then be used to represent model uncertainty.

We propose two flavors of MC dropout-based measures for unsupervised QE. **First**, we compute the expectation and variance for the set of sentence-level probability estimates obtained by running N stochastic forward passes through the MT model with model parameters $\hat{\theta}$ perturbed by dropout:

$$\text{D-TP} = \frac{1}{N}\sum_{n=1}^{N} \text{TP}_{\hat{\theta}^n}$$

$$\text{D-Var} = \mathbb{E}[\text{TP}_{\hat{\theta}}^2] - (\mathbb{E}[\text{TP}_{\hat{\theta}}])^2$$

where TP is sentence-level probability as defined in §3.1. We also look at a combination of the two:

$$\text{D-Combo} = \left(1 - \frac{\text{D-TP}}{\text{D-Var}}\right)$$

We note that these metrics have also been used by Wang et al. (2019), but with the purpose of minimizing the effect of low-quality outputs on NMT training with back translations.

Second, we measure lexical variation between the MT outputs generated for the same source segment when running inference with dropout. We posit that differences between likely MT hypotheses may also capture uncertainty and potential ambiguity and complexity of the original sentence. We compute an average similarity score (sim) between the set $\mathbb{H}$ of translation hypotheses:

$$\text{D-Lex-Sim} = \frac{1}{C}\sum_{i=1}^{|\mathbb{H}|}\sum_{j=1}^{|\mathbb{H}|} sim(h_i, h_j)$$

where $h_i, h_j \in \mathbb{H}$, $i \neq j$ and $C = 2^{-1}|\mathbb{H}|(|\mathbb{H}|-1)$ is the number of pairwise comparisons for $|\mathbb{H}|$ hypotheses. We use Meteor (Denkowski and Lavie, 2014) to compute similarity scores.

3.3 Attention

Attention weights represent the strength of connection between source and target tokens, which may be indicative of translation quality (Rikters and Fishel, 2017). One way to measure it is to compute the entropy of the attention distribution:

$$\text{Att-Ent} = -\frac{1}{I}\sum_{i=1}^{I}\sum_{j=1}^{J} \alpha_{ji} \log \alpha_{ji}$$

where α represents attention weights, I is the number of target tokens and J is the number of source tokens.

This mechanism can be applied to any NMT model with encoder-decoder attention. We focus on attention in Transformer models, as it is currently the most widely used NMT architecture. Transformers rely on various types of attention, multiple attention heads, and multiple encoder and decoder layers. Encoder-decoder attention weights are computed for each head (H) and for each layer (L) of the decoder, as a result we get $[H \times L]$ matrices with attention weights. It is not clear which combination would give the best results for QE. To summarize the information from different heads and layers, we propose to compute the entropy scores for each possible head/layer combination and then choose the minimum value or compute the average:

$$\text{AW:Ent-Min} = min_{\{hl\}}(\text{Att-Ent}_{\text{hl}})$$

$$\text{AW:Ent-Avg} = \frac{1}{H \times L}\sum_{h=1}^{H}\sum_{l=1}^{L} \text{Att-Ent}_{\text{hl}}$$

[3] A distinction is typically made between epistemic and aleatoric uncertainty, where the latter captures the noise inherent to the observations (Kendall and Gal, 2017). We leave modeling this distinction in NMT for future work.

Downloaded from http://www.mitpressjournals.org/doi/pdf/10.1162/tacl_a_00330?cookieSet=1 by guest on 28 March 2024

4 Multilingual Dataset for QE

The quality of NMT translations is strongly affected by the amount of training data. To study our unsupervised QE indicators under different conditions, we collected data for 6 language pairs that includes high-, medium-, and low-resource conditions. To add diversity, we varied the directions into and out-of English, when permitted by the availability of expert annotators into non-English languages. Thus our dataset is composed by the high-resource English–German (En-De) and English–Chinese (En-Zh) pairs; by the medium-resource Romanian–English (Ro-En) and Estonian–English (Et-En) pairs; and by the low-resource Sinhala–English (Si-En) and Nepali–English (Ne-En) pairs. The dataset contains sentences extracted from Wikipedia and the MT outputs manually annotated for quality.

Document and Sentence Sampling We follow the sampling process outlined in FLORES (Guzmán et al., 2019). First, we sampled documents from Wikipedia for English, Estonian, Romanian, Sinhala, and Nepali. Second, we selected the top 100 documents containing the largest number of sentences that are: (i) in the intended source language according to a language-id classifier[4] and (ii) have the length between 50 and 150 characters. In addition, we filtered out sentences that have been released as part of recent Wikipedia parallel corpora (Schwenk et al., 2019), ensuring that our dataset is not part of parallel data commonly used for NMT training.

For every language, we randomly selected 10K sentences from the sampled documents and then translated them into English using the MT models described below. For German and Chinese we selected 20K sentences from the top 100 documents in English Wikipedia. To ensure sufficient representation of high- and low-quality translations for high-resource language pairs, we selected the sentences with minimal lexical overlap with respect to the NMT training data.

NMT systems For medium- and high-resource language pairs we trained the MT models based on the standard Transformer architecture (Vaswani et al., 2017) and followed the implementation details described in Ott et al. (2018b). We used publicly available MT datasets such as Paracrawl (Esplà et al., 2019) and Europarl (Koehn, 2005). Si-En and Ne-En MT systems were trained based on Big-Transformer architecture as defined in Vaswani et al. (2017). For the low-resource language pairs, the models were trained following the FLORES semi-supervised setting (Guzmán et al., 2019),[5] which involves two iterations of backtranslation using the source and the target monolingual data. Table 1 specifies the amount of data used for training.

DA Judgments We followed the FLORES setup (Guzmán et al., 2019), which presents a form of DA (Graham et al., 2013). The annotators are asked to rate each sentence from 0–100 according to the perceived translation quality. Specifically, 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.

Each segment was evaluated independently by three professional translators from a single language service provider. To improve annotation consistency, any evaluation in which the range of scores among the raters was above 30 points was rejected, and an additional rater was requested to replace the most diverging translation rating until convergence was achieved. To further increase the reliability of the test and development partitions of the dataset, we requested an additional set of three annotations from a different group of annotators (i.e., from another language service provider) following the same annotation protocol, thus resulting in a total of six annotations per segment.

Raw human scores were converted into *z-scores*, that is, standardized according to each individual annotator's overall mean and standard deviation. The scores collected for each segment were averaged to obtain the final score. Such setting allows for the fact that annotators may genuinely disagree on some aspects of quality.

In Table 1 we show a summary of the statistics from human annotations. Besides the NMT training corpus size and the distribution of the DA scores for each language pair, we report mean

[4] https://fasttext.cc.

[5] https://bit.ly/36YaBlU.

Downloaded from http://www.mitpressjournals.org/doi/pdf/10.1162/tacl_a_00330?cookieSet=1 by guest on 28 March 2024

	Pair	size	scores				diff	
			avg	p25	median	p75	avg	std
High-resource	**En-De**	23.7M	84.8	80.7	88.7	92.7	13.7	8.2
	En-Zh	22.6M	67.0	58.7	70.7	79.0	12.1	6.4
Mid-resource	**Ro-En**	3.9M	68.8	50.1	76.0	92.3	10.7	6.7
	Et-En	880K	64.4	40.5	72.0	89.3	13.8	9.4
Low-resource	**Si-En**	647K	51.4	26.0	51.3	77.7	13.4	8.7
	Ne-En	564K	37.7	23.3	33.7	49.0	11.5	5.9

Table 1: Multilingual QE dataset. Size of the NMT training corpus (size) and summary statistics for the raw DA scores (average, 25th percentile, median, and 75th percentile). As an indicator of annotators' consistency, the last two columns show the mean (avg) and standard deviation (std) of the absolute differences (diff) between the scores assigned by different annotators to the same segment.

and standard deviation of the average differences between the scores assigned by different annotators to each segment, as an indicator of annotation consistency. First, we observe that, as expected, the amount of training data per language pair correlates with the average quality of an NMT system. Second, we note that the distribution of human scores changes substantially across language pairs. In particular, we see very little variability in quality for En-De, which makes QE for this language pair especially challenging (see §5). Finally, as shown in the right-most columns, annotation consistency is similar across language pairs and comparable to existing work that follows DA methodology for data collection. For example, Graham et al. (2013) report an average difference of 25 across annotators' scores.

Data Splits To enable comparison between supervised and unsupervised approaches to QE, we split the data into 7K training partition, 1K development set, and two test sets of 1K sentences each. One of these test sets is used for the experiments in this paper, the other is kept blind for future work.

Additional Data To support our discussion of the effect of NMT training on the correlation between predictive probabilities and perceived translation quality presented in §6, we trained various alternative NMT system variants, translated and annotated 400 original Estonian sentences from our test set with each system variant.

The data, the NMT models, and the DA judgments are available at `https://github.com/facebookresearch/mlqe`.

5 Experiments and Results

Below we analyze how our unsupervised QE indicators correlate with human judgments.

5.1 Settings

Benchmark Supervised QE Systems We compare the performance of the proposed unsupervised QE indicators against the best performing supervised approaches with available open-source implementation, namely, the Predictor-Estimator (PredEst) architecture (Kim et al., 2017b) provided by OpenKiwi toolkit (Kepler et al., 2019b), and an improved version of the BiRNN model provided by DeepQuest toolkit (Ive et al., 2018), which we refer to as BERT-BiRNN (Blain et al., 2020).

PredEst. We trained PredEst models (see §2) using the same parameters as in the default configurations provided by Kepler et al. (2019b). Predictor models were trained for 6 epochs on the same training and development data as the NMT systems, while the Estimator models were trained for 10 epochs on the training and development sets of our dataset (see §4). Unlike Kepler et al. (2019b), the Estimator was not trained using multitask learning, as our dataset currently does not contain any word-level annotation. We use the model corresponding to the best epoch as identified by the metric of reference on the development set: perplexity for the Predictor and Pearson correlation for the Estimator.

BERT-BiRNN. This model, similarly to the recent SOTA QE systems (Kepler et al., 2019a), uses a large-scale pre-trained BERT model to obtain token-level representations that are then

Downloaded from http://www.mitpressjournals.org/doi/pdf/10.1162/tacl_a_00330?cookieSet=1 by guest on 28 March 2024

fed into two independent bidirectional RNNs to encode both the source sentence and its translation independently. The two resulting sentence representations are then concatenated as a weighted sum of their word vectors, using an attention mechanism. The final sentence-level representation is then fed to a sigmoid layer to produce the sentence-level quality estimates. During training, BERT was fine-tuned by unfreezing the weights of the last four layers along with the embedding layer. We used early stopping based on Pearson correlation on the development set, with a patience of 5.

Unsupervised QE For the dropout-based indicators (see §3.2), we use dropout rate of 0.3, the same as for training the NMT models (see §4). We perform $N = 30$ inference passes to obtain the posterior probability distribution. N was chosen following the experiments in related work (Dong et al., 2018; Wang et al., 2019). However, we note that increasing N beyond 10 results in very small improvements on the development set. The implementation of stochastic decoding with MC dropout is available as part of the fairseq toolkit (Ott et al., 2019) at `https://github.com/pytorch/fairseq`.

5.2 Correlation with Human Judgments

Table 2 shows Pearson correlation with DA for our unsupervised QE indicators and for the supervised QE systems. Unsupervised QE indicators are grouped as follows: **Group I** corresponds to the measurements obtained with standard decoding (§3.1); **Group II** contains indicators computed using MC dropout (§3.2); and **Group III** contains the results for attention-based indicators (§3.3). **Group IV** corresponds to the supervised QE models presented in §5.1. We use the Hotelling-Williams test to compute significance of the difference between dependent correlations (Williams, 1959) with p-value < 0.05. For each language pair, results that are not significantly outperformed by any method are marked in bold; results that are not significantly outperformed by any other method from the same group are underlined.

We observe that the simplest measure that can be extracted from NMT, sequence-level probability (TP), already performs competitively, in particular for the medium-resource language pairs. TP is consistently outperformed by D-TP, indicating that NMT output probabilities are not well calibrated. This confirms our hypothesis that estimating model uncertainty improves correlation with perceived translation quality. Furthermore, our approach performs competitively with strong supervised QE models. Dropout-based indicators significantly outperform PredEst and rival BERT-BiRNN for four language pairs.[6] These results position the proposed unsupervised QE methods as an attractive alternative to the supervised approach in the scenario where the NMT model used to generate the translations can be accessed.

For both unsupervised and supervised methods performance varies considerably across language pairs. The highest correlation is achieved for the medium-resource languages, whereas for high-resource language pairs it is drastically lower. The main reason for this difference is a lower variability in translation quality for high-resource language pairs. Figure 2 shows scatter plots for Ro-En, which has the best correlation results, and En-De with the lowest correlation for all quality indicators. Ro-En has a substantial number of high-quality sentences, but the rest of the translations are uniformly distributed across the quality range. The distribution for En-De is highly skewed, as the vast majority of the translations are of high quality. In this case capturing meaningful variation appears to be more challenging, as the differences reflected by the DA may be more subtle than any of the QE methods is able to reveal.

The reason for a lower correlation for Sinhala and Nepalese is different. For unsupervised indicators it can be due to the difference in model capacity[7] and the amount of training data. On the one hand, increasing depth and width of the model may negatively affect calibration (Guo et al., 2017). On the other hand, due to the small amount of training data the model can overfit, resulting in inferior results both in terms of translation quality and correlation. It is noteworthy, however, that supervised QE system suffers a larger drop in performance than unsupervised indicators, as its

[6]We note that PredEst models are systematically and significantly outperformed by BERT-BiRNN. This is not surprising, as large-scale pretrained representations have been shown to boost model performance for QE (Kepler et al., 2019a) and other natural language processing tasks (Devlin et al., 2019).

[7]Models for these languages were trained using Transformer-Big architecture from Vaswani et al. (2017).

Downloaded from http://www.mitpressjournals.org/doi/pdf/10.1162/tacl_a_00330?cookieSet=1 by guest on 28 March 2024

	Method	Low-resource		Mid-resource		High-resource	
		Si-En	Ne-En	Et-En	Ro-En	En-De	En-Zh
I	TP	0.399	0.482	0.486	0.647	0.208	0.257
	Softmax-Ent (-)	0.457	0.528	0.421	0.613	0.147	0.251
	Sent-Std (-)	0.418	0.472	0.471	0.595	**0.264**	0.301
II	D-TP	0.460	0.558	**0.642**	0.693	**0.259**	**0.321**
	D-Var (-)	0.307	0.299	0.356	0.332	0.164	0.232
	D-Combo (-)	0.286	0.418	0.475	0.383	0.189	0.225
	D-Lex-Sim	**0.513**	**0.600**	0.612	0.669	0.172	**0.313**
III	AW : Ent-Min (-)	0.097	0.265	0.329	0.524	0.000	0.067
	AW : Ent-Avg (-)	0.10	0.205	0.377	0.382	0.090	0.112
	AW : best head/layer (-)	0.255	0.381	0.416	0.636	0.241	0.168
IV	PredEst	0.374	0.386	0.477	0.685	0.145	0.190
	BERT-BiRNN	**0.473**	0.546	**0.635**	**0.763**	**0.273**	**0.371**

Table 2: Pearson (r) correlation between unsupervised QE indicators and human DA judgments. Results that are not significantly outperformed by any method are marked in bold; results that are not significantly outperformed by any other method from the same group are underlined.

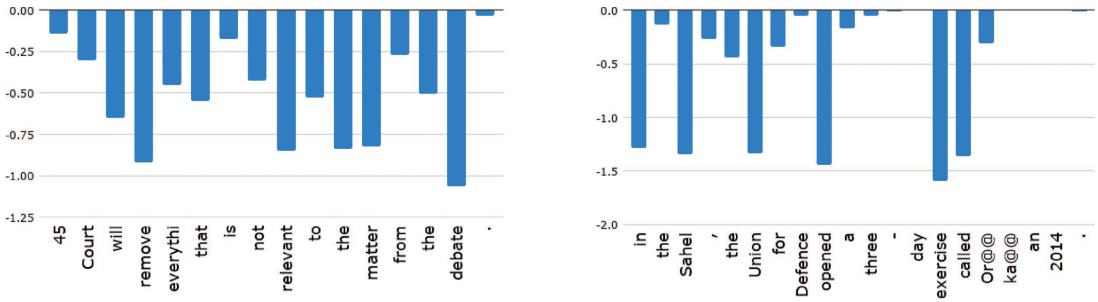


Figure 1: Token-level probabilities of high-quality (left) and low-quality (right) Et-En translations.

predictor component requires large amounts of parallel data for training. We suggest, therefore, that unsupervised QE is more stable in low-resource scenarios than supervised approaches.

We now look in more detail at the three groups of unsupervised measurements in Table 2.

Group I Average entropy of the softmax output (Softmax-Ent) and dispersion of the values of token-level probabilities (Sent-Std) achieve a significantly higher correlation than TP metric for four language pairs. Softmax-Ent captures uncertainty of the output probability distribution, which appears to be a more accurate reflection of the overall translation quality. Sent-Std captures a pattern in the sequence of token-level probabilities that helps detect low-quality translation illustrated in Figure 1. Figure 1 shows two Et-En translations that have drastically different absolute DA scores of 62 and 1, but the difference in their sentence-level log-probability is negligible: -0.50 and -0.48 for the first and second translations, respectively. By contrast, the sequences of token-level probabilities are very different, as the second sentence has larger variation in the log-probabilities for adjacent words, with very high probabilities for high-frequency function words and low probabilities for content words.

Group II The best results are achieved by the D-Lex-Sim and D-TP metrics. Interestingly, D-Var has a much lower correlation, because

Downloaded from http://www.mitpressjournals.org/doi/pdf/10.1162/tacl_a_00330?cookieSet=1 by guest on 28 March 2024

Low Quality	**Original**	Tanganjikast püütakse niiluse ahvenat ja kapentat.
	Reference	Nile perch and kapenta are fished from Lake Tanganyika.
	MT Output	There is a silver thread and candle from Tanzeri.
	Dropout	There will be a silver thread and a penny from Tanzer.
		There is an attempt at a silver greed and a carpenter from Tanzeri.
		There will be a silver bullet and a candle from Tanzer.
		The puzzle is being caught in the chicken's gavel and the coffin.
High Quality	**Original**	Siis aga võib tekkida seesmise ja välise vaate vahele lõhe.
	Reference	This could however lead to a split between the inner and outer view.
	MT Output	Then there may be a split between internal and external viewpoints.
	Dropout	Then, however, there may be a split between internal and external viewpoints.
		Then, however, there may be a gap between internal and external viewpoints.
		Then there may be a split between internal and external viewpoints.
		Then there may be a split between internal and external viewpoints.

Table 3: Example of MC dropout for a low-quality (top) and a high-quality (bottom) MT outputs.

by only capturing variance it ignores the actual probability estimate assigned by the model to the given output.[8]

Table 3 provides an illustration of how model uncertainty captured by MC dropout reflects the quality of MT output. The first example contains a low quality translation, with a high variability in MT hypotheses obtained with MC dropout. By contrast, MC dropout hypotheses for the second high-quality example are very similar and, in fact, constitute valid linguistic paraphrases of each other. This fact is directly exploited by the D-Lex-Sim metric that measures the variability between MT hypotheses generated with perturbed model parameters and performs on pair with D-TP. Besides capturing model uncertainty, D-Lex-Sim reflects the potential complexity of the source segments, as the number of different possible translations of the sentences is an indicator of their inherent ambiguity.[9]

Group III While our attention-based metrics also achieve a sensible correlation with human judgments, it is considerably lower than the rest of the unsupervised indicators. Attention may not provide enough information to be used as a quality indicator of its own, since there is no direct mapping between words in different languages, and, therefore, high entropy in attention weights does not necessarily indicate low translation quality. We leave experiments with combined attention and probability-based measures to future work.

The use of multihead attention with multiple layers in Transformer may also negatively affect

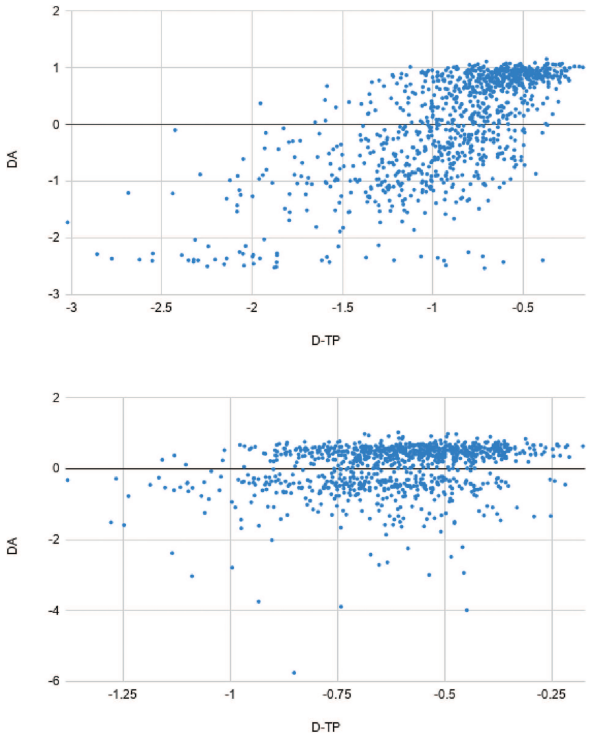


Figure 2: Scatter plots for the correlation between D-TP (x-axis) and standardized DA scores (y-axis) for Ro-En (top) and En-De (bottom).

[8]This is in contrast with the work by Wang et al. (2019) where D-Var appears to be one of the best performing metric for NMT training with back-translation demonstrating an essential difference between this task and QE.

[9]Note that D-Lex-Sim involves generating N additional translation hypotheses, whereas the D-TP only requires re-scoring an existing translation output and is thus less expensive in terms of time.

Downloaded from http://www.mitpressjournals.org/doi/pdf/10.1162/tacl_a_00330?cookieSet=1 by guest on 28 March 2024

the results. As shown by Voita et al. (2019), different attention heads are responsible for different functions. Therefore, combining the information coming from different heads and layers in a simple way may not be an optimal solution. To test whether this is the case, we computed attention entropy and its correlation with DA for all possible combinations of heads and layers. As shown in Table 2, the best head/layer combination (AW : best head/layer) indeed significantly outperforms other attention-based measurements for all language pairs suggesting that this method should be preferred over simple averaging. Using the best head/layer combination for QE is limited by the fact that it requires validation on a dataset annotated with DA and thus is not fully unsupervised. This outcome opens an interesting direction for further experiments to automatically discover the best possible head/layer combination.

6 Discussion

In the previous section we studied the performance of our unsupervised quality indicators for different language pairs. In this section we validate our results by looking at two additional factors: domain shift and underlying NMT system.

6.1 Domain Shift

One way to evaluate how well a model represents uncertainty is to measure the difference in model confidence under domain shift (Hendrycks and Gimpel, 2016; Lakshminarayanan et al., 2017; Snoek et al., 2019). A well-calibrated model should produce low confidence estimates when tested on data points that are far away from the training data.

Overconfident predictions on out-of-domain sentences would undermine the benefits of unsupervised QE for NMT. This is particularly relevant given the current wide use of NMT for translating mixed domain data online. Therefore, we conduct a small experiment to compare model confidence on in-domain and out-of-domain data. We focus on the Et-En language pair. We use the test partition of the MT training dataset as our in-domain sample. To generate the out-of-domain sample, we sort our Wikipedia data (prior to sentence sampling stage in §4) by distance to the training data and select the top 500 segments with the largest distance score. To compute distance scores we follow the strategy of Niehues and Pham (2019) that measures the test/training data distance based on the hidden states of NMT encoder.

We compute model posterior probabilities for the translations of the in-domain and out-of-domain sample either obtained through standard decoding, or using MC dropout. TP obtains average values of -0.440 and -0.445 for in-domain and out-of-domain data, respectively, whereas for D-TP these values are -0.592 and -0.685. The difference between in-domain and out-of-domain confidence estimates obtained by standard decoding is negligible. The difference between MC-dropout average probabilities for in-domain vs. out-of-domain samples was found to be statistically significant under Student's t-test, with p-value < 0.01. Thus, expectation over predictive probabilities with MC dropout indeed provides a better estimation of model uncertainty for NMT, and therefore can improve the robustness of unsupervised QE on out-of-domain data.

6.2 NMT Calibration across NMT Systems

Findings in the previous section suggest that using model probabilities results in fairly high correlation with human judgments for various language pairs. In this section we study how well these findings generalize to different NMT systems. The list of model variants that we explore is by no means exhaustive and was motivated by common practices in MT and by the factors that can negatively affect model calibration (number of training epochs) or help represent uncertainty (model ensembling). For this small-scale experiment we focus on Et-En. For each system variant we translated 400 sentences from the test partition of our dataset and collected the DA accordingly. As baseline, we use a standard Transformer model with beam search decoding. All system variants are trained using Fairseq implementation (Ott et al., 2019) for 30 epochs, with the best checkpoint chosen according to the validation loss.

First, we consider three system variants with differences in architecture or training: RNN-based NMT (Bahdanau et al., 2015; Luong et al., 2015), Mixture of Experts (MoE, He et al., 2018; Shen et al., 2019; Cho et al., 2019), and model ensemble (Garmash and Monz, 2016).

Shen et al. (2019) use the *MoE* framework to capture the inherent uncertainty of the MT task where the same input sentence can have multiple

Downloaded from http://www.mitpressjournals.org/doi/pdf/10.1162/tacl_a_00330?cookieSet=1 by guest on 28 March 2024

Method	r	DA
TP-Beam	0.482	58.88
TP-Sampling	0.533	42.02
TP-Diverse beam	0.424	55.12
TP-RNN	0.502	43.63
TP-Ensemble	0.538	61.19
TP-MoE	0.449	51.20
D-TP	0.526	58.88

Table 4: Pearson correlation (r) between sequence-level output probabilities (TP) and average DA for translations generated by different NMT systems.

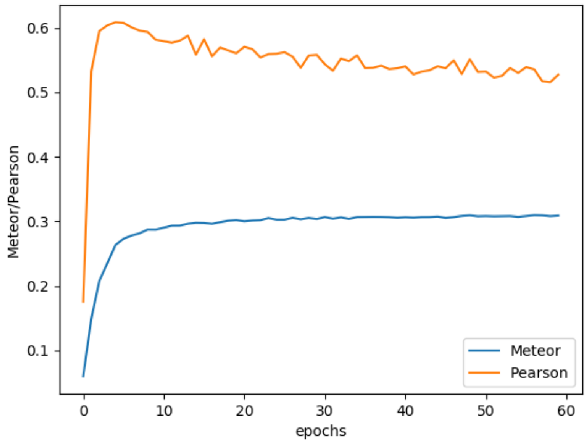


Figure 3: Pearson correlation between translation quality and model probabilities (orange), and Meteor (blue) over training epochs.

correct translations. A mixture model introduces a multinomial latent variable to control generation and produce a diverse set of MT hypotheses. In our experiment we use hard mixture model with uniform prior and 5 mixture components. To produce the translations we generate from a randomly chosen component with standard beam search. To obtain the probability estimates we average the probabilities from all mixture components.

Previous work has used *model ensembling* as a strategy for representing model uncertainty (Lakshminarayanan et al., 2017; Pearce et al., 2018).[10] In NMT, ensembling has been used to improve translation quality. We train four Transformer models initialized with different random seeds. At decoding time predictive distributions from different models are combined by averaging.

Second, we consider two alternatives to beam search: diverse beam search (Vijayakumar et al., 2016) and sampling. For sampling, we generate translations one token at a time by sampling from the model conditional distribution $p(y_j|\vec{y}_{<j}, \vec{x}, \theta)$, until the end of sequence symbol is generated. For comparison, we also compute the D-TP metric for the standard Transformer model on the subset of 400 segments considered for this experiment.

Table 4 shows the results. Interestingly, the correlation between output probabilities and DA is not necessarily related to the quality of MT outputs. For example, sampling produces much higher correlation although the quality is much lower. This is in line with previous work that indicates that sampling results in better calibrated probability distribution than beam search (Ott et al., 2018a). System variants that promote diversity in NMT outputs (diverse beam search and MoE) do not achieve any improvement in correlation over standard Transformer model.

The best results both in quality and QE are achieved by ensembling, which provides additional evidence that better uncertainty quantification in NMT improves correlation with human judgments. MC dropout achieves very similar results. We recommend using either of these two methods for NMT systems with unsupervised QE.

6.3 NMT Calibration across Training Epochs

The final question we address is how the correlation between translation probabilities and translation quality is affected by the amount of training. We train our base Et-En Transformer system for 60 epochs. We generate and evaluate translations after each epoch. We use the test partition of the MT training set and assess translation quality with Meteor evaluation metric. Figure 3 shows the average Meteor scores (blue) and Pearson correlation (orange) between segment-level Meteor scores and translation probabilities from the MT system for each epoch.

Interestingly, as the training continues test quality stabilizes whereas the relation between model probabilities and translation quality is deteriorated. During training, after the model is

[10] Note that MC dropout discussed in §3.2 can be interpreted as an ensemble model combination where the predictions are averaged over an ensemble of NNs (Lakshminarayanan et al., 2017).

Downloaded from http://www.mitpressjournals.org/doi/pdf/10.1162/tacl_a_00330?cookieSet=1 by guest on 28 March 2024

able to correctly classify most of the training examples, the loss can be further minimized by increasing the confidence of predictions (Guo et al., 2017). Thus longer training does not affect output quality but damages calibration.

7 Conclusions

We have devised an unsupervised approach to QE where no training or access to any additional resources besides the MT system is required. Besides exploiting softmax output probability distribution and the entropy of attention weights from the NMT model, we leverage uncertainty quantification for unsupervised QE. We show that, if carefully designed, the indicators extracted from the NMT system constitute a rich source of information, competitive with supervised QE methods.

We analyzed how different MT architectures and training settings affect the relation between predictive probabilities and translation quality. We showed that improved translation quality does not necessarily imply a stronger correlation between translation quality and predictive probabilities. Model ensemble have been shown to achieve optimal results both in terms of translation quality and when using output probabilities as an unsupervised quality indicator.

Finally, we created a new multilingual dataset for QE covering various scenarios for MT development including low- and high-resource language pairs. Both the dataset and the MT models needed to reproduce the results of our experiments are available at `https://github.com/facebookresearch/mlqe`.

This work can be extended in many directions. First, our sentence-level unsupervised metrics could be adapted for QE at other levels (word, phrase, and document). Second, the proposed metrics can be combined as features in supervised QE approaches. Finally, other methods for uncertainty quantification, as well as other types of uncertainty, can be explored.

Acknowledgments

Marina Fomicheva, Lisa Yankovskaya, Frédéric Blain, Mark Fishel, Nikolaos Aletras, and Lucia Specia were supported by funding from the Bergamot project (EU H2020 grant no. 825303).

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015*.

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

Luisa Bentivogli, Arianna Bisazza, Mauro Cettolo, and Marcello Federico. 2016. Neural versus phrase-based machine translation quality: A case study. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 257–267.

Frédéric Blain, Nikolaos Aletras, and Lucia Specia. 2020. Quality in, quality out: Learning from actual mistakes. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*.

John Blatz, Erin Fitzgerald, George Foster, Simona Gandrabur, Cyril Goutte, Alex Kulesza, Alberto Sanchis, and Nicola Ueffing. 2004. Confidence Estimation for Machine Translation. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, pages 315–321.

Ondřej Bojar, Yvette Graham, and Amir Kamran. 2017. Results of the WMT17 metrics shared task. In *Proceedings of the Second Conference on Machine Translation*, pages 489–513, Copenhagen, Denmark. Association for Computational Linguistics.

Ondřej Bojar, Yvette Graham, Amir Kamran, and Miloš Stanojević. 2016. Results of the WMT16 metrics shared task. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 199–231, Berlin, Germany. Association for Computational Linguistics.

Downloaded from http://www.mitpressjournals.org/doi/pdf/10.1162/tacl_a_00330?cookieSet=1 by guest on 28 March 2024

Sheila Castilho, Joss Moorkens, Federico Gaspari, Iacer Calixto, John Tinsley, and Andy Way. 2017. Is neural machine translation the new state of the art? *The Prague Bulletin of Mathematical Linguistics*, 108(1):109–120.

Jaemin Cho, Minjoon Seo, and Hannaneh Hajishirzi. 2019. Mixture content selection for diverse sequence generation. 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 3112–3122.

Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the EACL 2014 Workshop on Statistical Machine Translation*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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.

Li Dong, Chris Quirk, and Mirella Lapata. 2018. Confidence modeling for neural semantic parsing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 743–753. Association for Computational Linguistics, Melbourne, Australia.

Miquel Esplà, Mikel Forcada, Gema Ramírez-Sánchez, and Hieu Hoang. 2019. ParaCrawl: Web-scale parallel corpora for the languages of the EU. In *Proceedings of Machine Translation Summit XVII Volume 2: Translator, Project and User Tracks*, pages 118–119, Dublin, Ireland. European Association for Machine Translation.

Thierry Etchegoyhen, Eva Martínez Garcia, and Andoni Azpeitia. 2018. Supervised and unsupervised minimalist quality estimators: Vicomtechs participation in the WMT 2018 Quality Estimation Task. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 782–787.

Erick Fonseca, Lisa Yankovskaya, André F. T. Martins, Mark Fishel, and Christian Federmann. 2019. Findings of the WMT 2019 Shared Tasks on Quality Estimation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 1–10.

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

Ekaterina Garmash and Christof Monz. 2016. Ensemble learning for multi-source neural machine translation. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1409–1418.

Yvette Graham, Timothy Baldwin, Meghan Dowling, Maria Eskevich, Teresa Lynn, and Lamia Tounsi. 2016. Is all that glitters in machine translation quality estimation really gold? In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3124–3134.

Yvette Graham, Timothy Baldwin, and Nitika Mathur. 2015a. Accurate evaluation of segment-level machine translation metrics. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1183–1191.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2013. Continuous measurement scales in human evaluation of machine translation. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 33–41.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2015b. Can machine translation systems be evaluated by the crowd alone. *Natural Language Engineering*, pages 1–28.

Alex Graves. 2011. Practical variational inference for neural networks. In *Advances in Neural Information Processing Systems*, pages 2348–2356.

Downloaded from http://www.mitpressjournals.org/doi/pdf/10.1162/tacl_a_00330?cookieSet=1 by guest on 28 March 2024

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1321–1330. JMLR. org.

Francisco Guzmán, Peng-Jen Chen, Myle Ott, Juan Pino, Guillaume Lample, Philipp Koehn, Vishrav Chaudhary, and Marc'Aurelio Ranzato. 2019. 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 6097–6110, Hong Kong, China. Association for Computational Linguistics.

Hany Hassan, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mu Li, et al. 2018. Achieving human parity on automatic Chinese to English news translation. *arXiv preprint arXiv:1803.05567*.

Xuanli He, Gholamreza Haffari, and Mohammad Norouzi. 2018. Sequence to sequence mixture model for diverse machine translation. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 583–592.

Dan Hendrycks and Kevin Gimpel. 2016. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*.

Julia Ive, Frédéric Blain, and Lucia Specia. 2018. DeepQuest: a framework for neural-based quality estimation. In *Proceedings of COLING 2018, the 27th International Conference on Computational Linguistics: Technical Papers*. Santa Fe, New Mexico.

Alex Kendall and Yarin Gal. 2017. What uncertainties do we need in Bayesian deep learning for computer vision? In *Advances in Neural Information Processing Systems*, pages 5574–5584.

Fabio Kepler, Jonay Trénous, Marcos Treviso, Miguel Vera, António Góis, M Amin Farajian, António V. Lopes, and André F. T. Martins. 2019a. Unbabel's Participation in the WMT19 Translation Quality Estimation Shared Task. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 78–84.

Fábio Kepler, Jonay Trénous, Marcos Treviso, Miguel Vera, and André F. T. Martins. 2019b. OpenKiwi: An open source framework for quality estimation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics–System Demonstrations*, pages 117–122, Florence, Italy. Association for Computational Linguistics.

Hyun Kim, Jong-Hyeok Lee, and Seung-Hoon Na. 2017a. Predictor-estimator using multilevel task learning with stack propagation for neural quality estimation. In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Tasks Papers*, pages 562–568. Copenhagen, Denmark.

Hyun Kim, Jong-Hyeok Lee, and Seung-Hoon Na. 2017b. Predictor-estimator using multilevel task learning with stack propagation for neural quality estimation. In *Proceedings of the Second Conference on Machine Translation*, pages 562–568.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86.

Volodymyr Kuleshov and Percy S. Liang. 2015. Calibrated structured prediction. In *Advances in Neural Information Processing Systems*, pages 3474–3482.

Aviral Kumar and Sunita Sarawagi. 2019. Calibration of encoder decoder models for neural machine translation. *arXiv preprint arXiv:1903.00802*.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, pages 6402–6413.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on*

Downloaded from http://www.mitpressjournals.org/doi/pdf/10.1162/tacl_a_00330?cookieSet=1 by guest on 28 March 2024

Empirical Methods in Natural Language Processing, pages 1412–1421, Lisbon, Portugal. Association for Computational Linguistics.

Qingsong Ma, Ondřej Bojar, and Yvette Graham. 2018. Results of the WMT18 metrics shared task: Both characters and embeddings achieve good performance. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 671–688, Belgium, Brussels. Association for Computational Linguistics.

Qingsong Ma, Johnny Wei, Ondřej Bojar, and Yvette Graham. 2019. Results of the WMT19 metrics shared task: Segment-level and strong MT systems pose big challenges. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 62–90, Florence, Italy. Association for Computational Linguistics.

David J. C. MacKay. 1992. Bayesian Methods for Adaptive Models. Ph.D. thesis, California Institute of Technology.

Kristian Miok, Dong Nguyen-Doan, Blaž Škrlj, Daniela Zaharie, and Marko Robnik-Šikonja. 2019. Prediction uncertainty estimation for hate speech classification. Carlos Martín-Vide, Matthew Purver, and Senja Pollak, editors, In *Statistical Language and Speech Processing*, pages 286–298, Cham. Springer International Publishing.

Erwan Moreau and Carl Vogel. 2012. Quality estimation: an experimental study using unsupervised similarity measures. In *7th Workshop on Statistical Machine Translation*, page 120.

Khanh Nguyen and Brendan O'Connor. 2015. Posterior calibration and exploratory analysis for natural language processing models. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1587–1598, Lisbon, Portugal. Association for Computational Linguistics.

Jan Niehues and Ngoc-Quan Pham. 2019. Modeling confidence in sequence-to-sequence models. In *Proceedings of The 12th International Conference on Natural Language Generation*, pages 575–583.

Myle Ott, Michael Auli, David Grangier, and Marc'Aurelio Ranzato. 2018a. Analyzing uncertainty in neural machine translation. In *International Conference on Machine Learning*.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*.

Myle Ott, Sergey Edunov, David Grangier, and Michael Auli. 2018b. Scaling neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 1–9.

Tim Pearce, Mohamed Zaki, Alexandra Brintrup, and Andy Neel. 2018. Uncertainty in neural networks: Bayesian ensembling. *arXiv preprint arXiv:1810.05546*.

Maja Popović. 2012. Morpheme-and pos-based ibm1 scores and language model scores for translation quality estimation. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 133–137. Association for Computational Linguistics.

Matīss Rikters and Mark Fishel. 2017. Confidence through attention. In *Proceedings of MT Summit XVI*, pages 299–311. Nagoya, Japan.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2019. WikiMatrix: Mining 135M Parallel Sentences in 1620 Language Pairs from Wikipedia. *arXiv preprint arXiv:1907.05791*.

Tianxiao Shen, Myle Ott, Michael Auli, and Marc'Aurelio Ranzato. 2019. Mixture models for diverse machine translation: Tricks of the trade. *arXiv preprint arXiv:1902.07816*.

Jasper Snoek, Yaniv Ovadia, Emily Fertig, Balaji Lakshminarayanan, Sebastian Nowozin, D. Sculley, Joshua Dillon, Jie Ren, and Zachary Nado. 2019. Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift. In *Advances in Neural Information Processing Systems*, pages 13969–13980.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted

Downloaded from http://www.mitpressjournals.org/doi/pdf/10.1162/tacl_a_00330?cookieSet=1 by guest on 28 March 2024

human annotation. In *Proceedings of association for machine translation in the Americas*, volume 200.

Lucia Specia, Frédéric Blain, Varvara Logacheva, Ramón Astudillo, and André F. T. Martins. 2018. Findings of the WMT 2018 Shared Task on Quality Estimation. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 689–709.

Lucia Specia, Kashif Shah, José G. C. De Souza, and Trevor Cohn. 2013. QuEst-A Translation Quality Estimation Framework. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 79–84.

Lucia Specia, Marco Turchi, Nicola Cancedda, Marc Dymetman, and Nello Cristianini. 2009. Estimating the sentence-level quality of machine translation systems. In *13th Conference of the European Association for Machine Translation*, pages 28–37.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems*, pages 3104–3112.

Dustin Tran, Mike Dusenberry, Mark van der Wilk, and Danijar Hafner. 2019. Bayesian layers: A module for neural network uncertainty. In *Advances in Neural Information Processing Systems*, pages 14633–14645.

Shikhar Vashishth, Shyam Upadhyay, Gaurav Singh Tomar, and Manaal Faruqui. 2019. Attention interpretability across NLP tasks. *arXiv preprint arXiv:1909.11218*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

Jesse Vig and Yonatan Belinkov. 2019. Analyzing the structure of attention in a transformer language model. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 63–76.

Ashwin K. Vijayakumar, Michael Cogswell, Ramprasath R. Selvaraju, Qing Sun, Stefan Lee, David Crandall, and Dhruv Batra. 2016. Diverse beam search: Decoding diverse solutions from neural sequence models. *arXiv preprint arXiv:1610.02424*.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. 2019. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5797–5808, Florence, Italy. Association for Computational Linguistics.

Jiayi Wang, Kai Fan, Bo Li, Fengming Zhou, Boxing Chen, Yangbin Shi, and Luo Si. 2018. Alibaba submission for WMT18 quality estimation task. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 809–815, Belgium, Brussels. Association for Computational Linguistics.

Shuo Wang, Yang Liu, Chao Wang, Huanbo Luan, and Maosong Sun. 2019. Improving back-translation with uncertainty-based confidence estimation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 791–802.

Max Welling and Yee W. Teh. 2011. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 681–688.

Evan James Williams. 1959. *Regression Analysis*, 14, Wiley, New York.

Elizaveta Yankovskaya, Andre Tattar, and Mark Fishel. 2018. Quality estimation with force-decoded attention and cross-lingual embeddings. In *Proceedings of the Third Conference on Machine Translation, Volume 2: Shared Tasks Papers*. Brussels, Belgium.

Downloaded from http://www.mitpressjournals.org/doi/pdf/10.1162/tacl_a_00330?cookieSet=1 by guest on 28 March 2024