Poor Man's Quality Estimation:
Predicting Reference-Based MT Metrics Without the Reference

Vilém Zouhar[E] Shehzaad Dhuliawala[E] Wangchunshu Zhou[E] Nico Daheim[D]
Tom Kocmi[M] Yuchen Eleanor Jiang[E] Mrinmaya Sachan[E]

[E]ETH Zürich [M]Microsoft [D]TU Darmstadt
{vzouhar,msachan}@ethz.ch

Abstract

Machine translation quality estimation (QE) predicts human judgements of a translation hypothesis without seeing the reference. State-of-the-art QE systems based on pretrained language models have been achieving remarkable correlations with human judgements yet they are computationally heavy and require human annotations, which are slow and expensive to create. To address these limitations, we define the problem of *metric estimation* (ME) where one predicts the automated metric scores also without the reference. We show that even without access to the reference, our model can estimate automated metrics (ρ=60% for BLEU, $\bar{\rho}$=51% for other metrics) at the sentence-level. Because automated metrics correlate with human judgements, we can leverage the ME task for pre-training a QE model. For the QE task, we find that pre-training on TER is better (ρ=23%) than training for scratch (ρ=20%).

github.com/zouharvi/mt-metric-estimation

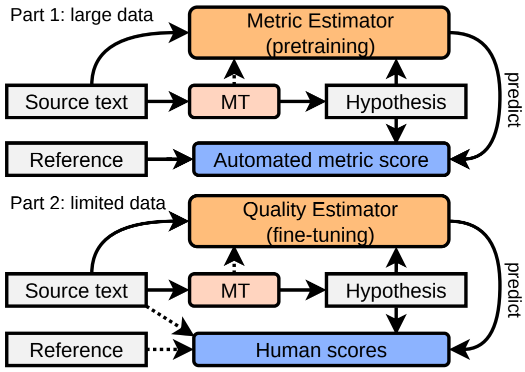

Figure 1: Pipeline for pre-training on automated metrics (top) and fine-tuning on limited quality estimation data (bottom). Dotted lines are optional dependency.

1 Introduction

Quality estimation (QE) is often used in machine translation (MT) production pipelines where we need to make decisions based on the quality of an MT output but where the reference is unavailable (Specia et al., 2020, 2021). For example, QE is used in translation companies to decide whether to send a specific MT output for post-editing to a human translator or whether to use it directly (Tamchyna, 2021; Zouhar et al., 2021; Murgolo et al., 2022). In this scenario, an accurate QE system has the potential to save expensive translator effort. However, training QE models usually requires human-annotated judgements of the translation quality (Specia et al., 2013; Rubino and Sumita, 2020; Rei et al., 2020a). These human annotations are scarce and costly, especially for low-resource language directions, and may need to be replicated for new MT systems and domains (de Souza et al., 2014)

We investigate if automated MT metrics can be used to reduce the cost of learning QE models. Automated metrics can be run with no additional costs and can be used to generate large amounts of QE data. If these metrics correlate well with human judgements, this larger data can be used as a partial substitute for human data during training. Data augmentation and synthetic QE data via automated metrics has already been explored (Heo et al., 2021; Baek et al., 2020; Cui et al., 2021), though never in the pre-training & fine-tuning fashion.

Our work is guided by a simple intuition. Intuitively, human judgements can be thought of as functions depending on the target sequence and (optionally) the source and reference(s): $f_{\text{HUMAN}}([s], h, [r])$. The task of QE is to model f_{HUMAN} based only on the source s and hypothesis h. Because the function arguments of $f_{\text{HUMAN}}([s], h, [r])$ resemble those of automated metrics for MT: $f_{\text{METRIC}}([s], h, r)$, we can use the automated metrics to guide learning of the human quality judgements which are hard to obtain and

Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, pages 1311–1325
May 2-6, 2023 ©2023 Association for Computational Linguistics

replicate. Generating automated metric scores is limited only by the amount of parallel data which is more abundant. Because f_{METRIC} correlates with human judgement f_{HUMAN} (Ma et al., 2019; Mathur et al., 2020), we can start by estimating f_{METRIC} and only later fine-tune to f_{HUMAN}. We refer to estimating $f_{\text{METRIC}}(s, h, r)$ as **metric estimation (ME)** as a parallel task to **quality estimation (QE)**.

We illustrate our idea of pre-training on the automatic metrics and fine-tuning on human assesments in Figure 1. Our model uses a BiLSTM with the source and hypothesis as the input (with several more features like the decoder confidence and hypothesis variance) to output a single number (metric or quality score).

Experimentally, we find the idea of mitigating data limitation for QE with ME pretraining challenging. Thus, we structure our investigation around a set of research questions. First, we try to establish that it is possible to robustly predict automated metrics and explore the associated data requirements. Then, concerned with the application and deployment of the ME model, we also check how transferrable the model is between different MT systems. We break the research down into the following questions.

RQ1: Can automated reference-based MT metrics be reliably predicted without the reference?
RQ2: What is the effect of data size on predicting automated reference-based MT metrics?
RQ3: How detrimental is the transfer between different MT systems for metric estimation models?
RQ4: Can metric estimation be used as a pre-training step for quality estimation?

We answer the research questions with experiments on the English → German language direction and replicate the main findings on 12 language pairs in total in Appendix C. We confirm that BLEU is predictable with $\rho = 60.4\%$ sentence-level Pearson's correlation and other metrics with $\bar{\rho} = 51.3\%$ (**RQ1**, Section 4.2). Authentic parallel data is needed for ME models but this can be alleviated by using more hypothesis from beam search (**RQ2**, Section 4.3). It is possible to train the ME system on one MT system and then use it on a different MT system with only a slight loss in performance (**RQ3**, Section 4.4). We find that pre-training on Translation Edit Rate (TER) (Snover et al., 2006) leads to better results than training on the QE data directly, though this approach does not outperform the state-of-the-art in QE (**RQ4**, Section 4.5).

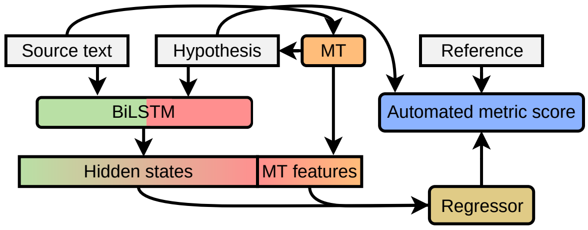

Figure 2: Metric/quality estimation model architecture.

2 Our Metric Estimation Model

Notation. Given a source sentence s and its translation, h which is an output of a MT system ($h = \text{MT}(s)$), we build a regressive ME model which outputs a numeric score that is close to the output of an automated metric that is further dependent on the reference: $f_{\text{METRIC}}(s, h, r)$. We distinguish two cases based on our level of access to the MT system: **blackbox** setting (where we assume access only to the MT system output) and **glassbox** setting (where we have access to entire MT model). In the later case, we may leverage features from the MT model to improve the ME capabilities.

Model. Our main model for ME/QE, shown in Figure 2, starts with Byte-Pair-Encoding (Gage, 1994; Sennrich et al., 2016) the source s and hypothesis h. It is followed by BiLSTM on top of concatenated source and hypothesis with a separator ($s\,\texttt{[SEP]}\,t$). The last hidden state (denoted as BiLSTM_{-1}) is extracted and fused together via concatenation with the internal MT system and other features (see the following list). This is then used in a simple feed-forward layer (FFNN) to output a single number:

$$\begin{aligned} \text{source } s, \text{hypothesis } h &:= \text{MT}(s) \\ \Phi &= \text{BiLSTM}_{-1}(\text{BPE}(s\,\texttt{[SEP]}\,h)) \\ \Psi &= \text{Features}(\text{MT}_{\text{conf.}}(s,h), s, h) \\ \text{ME}_{\text{all}}(s,h) &= \text{FFNN}[\Phi, \Psi] \\ \text{ME}_{\text{text}}(s,h) &= \text{FFNN}[\Phi] \end{aligned}$$

The first ME model is glassbox and the second is blackbox. In the first case, we utilize hand-crafted features and also those from the MT system (function features). Both of these models are optimized with mean-squared error against a particular metric. That is, we train separate models for each target metric (COMET, ChrF, BLEURT, BLEU, METEOR, TER) or human judgements.

$$\mathcal{L} = \frac{1}{n} \sum (\text{ME}(s_i, h_i) - f_{\text{METRIC}}(s_i, h_i, r_i))^2$$

The additional features Ψ are:

- Decoder confidence (prob and logprob).
- Source and target lengths and their relation. This is included as the distribution of errors may be different for various sentence lengths.
- Average distance and variance between hypotheses as measured by an automatic metric.

Decoder confidence. Low probability MT outputs have overall lower quality (Specia et al., 2018; Yankovskaya et al., 2018; Fomicheva et al., 2020). The decoder confidence is the hypothesis probability as defined by the model $\prod_i p(t_i|t_{<i}, s)$ which is in practice usually computed in the logspace $\sum_i \log p(t_i|t_{<i}, s)$.[1]

Hypotheses variance. Intuitively, there are many ways to generate a wrong translation but only a few correct ones (Xu et al., 2011). Similar to Fomicheva et al. (2020), we hypothesise that larger variance between the hypotheses correlates negatively with quality. We therefore use the distances between hypotheses as features for our system. Specifically, as shown in Figure 7 and formalized with the following, we use the mean distance and also the variance between distances as features. We first consider distances from the current hypothesis to be estimated (`H1`) to all other hypotheses, and then all hypothesis pairs.

$$\text{Avg or Var}(\{\text{BLEU}(h_1, h_j)|h_j \in H\})$$
$$\text{Avg or Var}(\{\text{BLEU}(h_i, h_j)|h_i, h_j \in H, i \neq j\})$$

Baselines. We use multiple baselines for comparison in the ME task. Apart from the individual features, they are all optimized to minimize the MSE loss with a specific target metric.

- Linear regression on TF-IDF features (with limited max features, see Appendix B): $\text{Lin.Reg.}\big[\text{TF-IDF}(s, h)\big]$
- Linear regression on all text & MT features: $\text{Lin.Reg.}[\Psi]$
- Fine-tuned mBERT (Devlin et al., 2019) with identical regression head architecture on top of last layer `[CLS]` hidden state[2]: $\text{FFNN}\big[\text{mBERT}(s\,\texttt{[SEP]}\,h)\big]$

We also use an off-the-shelf QE model used in WMT21 QE task `wmt21-comet-qe-mqm` (Rei et al., 2020b).[3] We do not fine-tune the model to the available data, but since we use correlation as an evaluation metric, the mean is subtracted and the output rescaled to unit variance, same as human judgements.

Automated metrics. For ME we use the following MT metrics. BLEU (Papineni et al., 2002), ChrF (Popović, 2015), TER (Snover et al., 2006) and METEOR (Banerjee and Lavie, 2005) work with lexical or character-level units, commonly in word or character n-grams. COMET (Rei et al., 2020a) uses pre-trained encoders to evaluate the hypothesis at a deeper level. BLEURT (Sellam et al., 2020) is another learned metric for text generation which uses pseudo-label. While automated metrics usually yield only sentence-level scores, QE is done for multiple levels: word, phrase and sentence. However, because of the automatic metric restrictions, we also focus on sentence-level QE in this work.

3 Experiment Setup

Pipeline & data. We start by translating 500k English→German sentences of the WMT14 dataset (Bojar et al., 2014) and computing the automated metrics of these translations.[4] We use a pre-trained WMT19 model by Ng et al. (2019).

For human scores, we use the train data of WMT21 Sentence-Level Quality Estimation Shared Task (Specia et al., 2021) which contains 14k human-direct-assessment annotated segments (5k unique). For the human score prediction, we do not have access to the features in the hypothesis space (because the hypotheses were not generated by the MT system to which we have access) and use forced decoding of the pre-trained model to get a confidence estimate. Note that since the data comes from a different MT system, there is a distribution mismatch which may negatively influence the results. We address this in Section 4.4.

We refer to the two datasets as ME and QE, respectively and show the distribution of automated metric scores and human judgements in Figure 6. We perform ME on both but QE only on the latter because of the human annotations availability.

[1] The justification for using both conf_t and $\exp(\text{conf}_t)$ is that the non-linear transformation improves correlation.

[2] See Appendices A and B for details.

[3] This model was better than `wmt21-comet-qe-da`. Note the difference between the automated metric COMET and the QE system COMET-QE.

[4] We are not limited by the relatively small size of this dataset because we are considering only its subset and study the effect of available data size in Section 4.3.

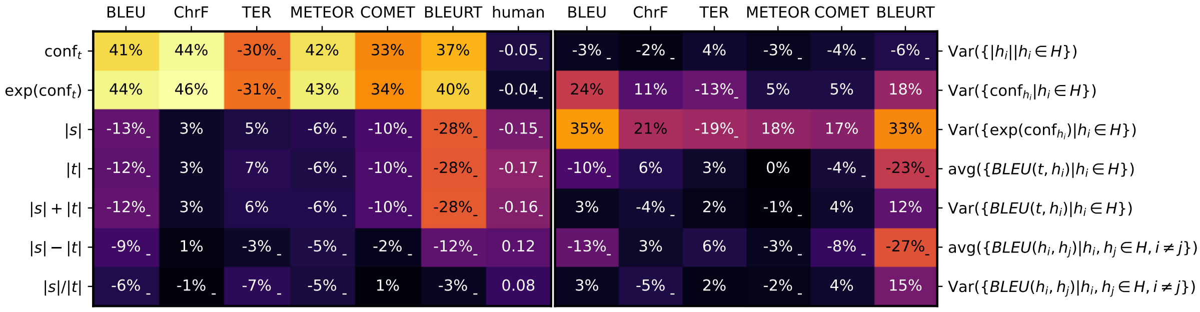

Figure 3: Feature correlations on 500k English → German sentences with reference-based metrics and human judgement. Colour is based on absolute values to show contained relevant information. Cells with negative correlations are marked with "-".

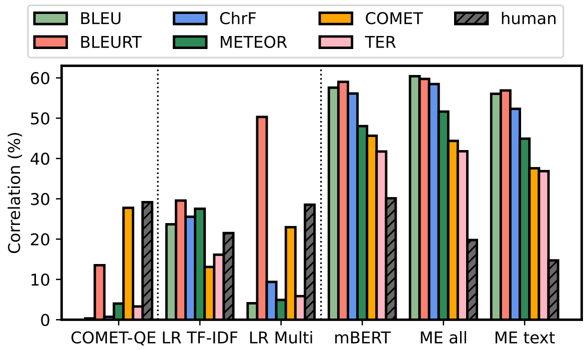

Figure 4: Correlations with metrics and human judgement of baseline and main metric estimation (ME) models. Each bar is a separate model trained to predict a particular metric or human judgement. Individual features are shown in Figure 3.

Evaluation. We evaluate ME model performance with Pearson's coefficient with the target metric on a dev set of 10k sentences from WMT14 on the segment-level. Note that we care about the *magnitude (absolute value) of the correlation* and not whether it is positive or negative. For example, TER is expected to correlate negatively with human ranking because higher TER means more errors while higher human scores mean higher translation quality. Correlations with human judgement are evaluated on 1k WMT21 Sentence-Level QE data. Specifically, we estimate human z-scores which were computed per-annotator.[5]

4 Results

This section first studies single-feature baselines (Section 4.1) and then the possibility of robust ME model (Section 4.2) and data size requirements (Section 4.3). The model is then checked on a different MT system output to see transferability between systems and architectures (Section 4.4). Fine-tuning and evaluation on human data (QE) is done in Section 4.5. A natural follow-up experiment on using joint prediction to improve the ME model is documented in Section 4.6.

4.1 Feature analysis

We show the correlations between individual features and metrics in Figure 3. An immediate observation is that confidence-based features correlate much more with automatic metrics than the other features. Some metrics (BLEU and COMET) and especially human z-scores are highly correlated with the source and target lengths. As a negative result, very few of the hypothesis space metrics correlate highly, with an exception of $\rho(\text{BLEU}, \text{Var}(\text{BLEU}(H)))$. We still use all features later on because (1) despite low individual correlations, they may still be useful in combination or for the full model, whose input is the text, and (2) we did not encounter any overfitting issues.

BLEU	BLEURT	ChrF	METEOR	COMET	TER
11.1%	16.5%	12.3%	-12.0%	11.5%	34.6%

Table 1: Pearson correlations between human judgement (human z-scores) and automated metrics.

We show the correlations between the automated metrics and humans in Table 1. For most of them, the correlations are very low. Outliers with the highest correlation are BLEURT and COMET, which are known to be strong-performant metrics (Kocmi et al., 2021). One of the reasons is that they were specifically trained to correlate well with humans.

[5] Z-score of a variable has zero mean and unit variance. They are possibly unbounded but on the other hand, slightly alleviate the effect of individual annotator differences.

4.2 Metric estimation performance

We show the baselines with comparison to the main models trained only on the target data (either WMT News or WMT QE) in Figure 4. Every bar is a separate model trained to predict a specific metric and the bar magnitude shows its correlation. The systems are described formally in Section 2. Notably, *ME text* has access to only the source and hypothesis texts while *ME all* in addition fuses in extra hand-crafted features.

A simple linear regression based on features from Figure 3 is able to achieve $> 40\%$ correlations with automated metrics and $\sim$13% correlation with human judgement. These features seem important as demonstrated by the comparatively lower correlations of a TF-IDF featurizer. This is also documented by the difference between *ME all* and *ME text*. The former model consistently outperforms *ME text*, possibly because it has access to all the extra features while the latter model only works with the source and target texts. The pre-training of the *mBERT* model on language modelling helps only marginally, given that it performs only slightly better than *ME text*. The COMET-QE model almost does not correlate with automated metrics at all apart from its related reference-based metric COMET. Of all models, it also correlates the most with human judgement.

> **RQ1**: Best-performing metric estimation models (mBERT, ME all, ME text) achieve sound correlations with automated metrics (60.4% with BLEU, $\sim 51.3\%$ for others).

We can interpret this correlation with respect to the individual features performances (max $\rho = 46\%$) which shows that the model was able to learn more predictive patterns. On the selected datasets, the ME task is easier than the QE task where our moidels have a consistently lower performance.

4.3 Data requirements for ME

We are interested in how much data we need to train the ME models. This is a practical question that helps us understand the behaviour and requirements of the models.

Naturally, models utilizing just a handful of continuous non-trainable features require much less data than models whose inputs are the raw texts. This is demonstrated in Figure 5 where the linear regression gains very little even if 500× more data is used. For the main ME model, a larger amount of data is required (line H1). So far we have been considering only the highest-scored hypothesis (in terms of decoder score) among 5 generated by beam search: (s, h_1). In a low-resource scenario, it may be beneficial to create more pairs from the hypotheses provided by the beam-search decoder: $\{(s, h_i) | h_i \in H\}$. Through this hypothesis expansion, we may obtain more parallel data in data-restricted settings. Again we show the results in Figure 5 with a substantial gain of the model trained on expanded data (H2) over just the top hypothesis (H1). This effect is quickly diminishing with larger data ($\geq$ 10k) but for lower data remains a useful tool.

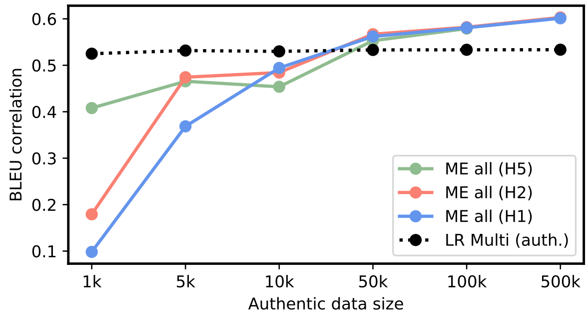

Figure 5: Model dependency on training data size. H2 and H5 expand the data by top-2 and top-5 decoder hypotheses, respectively. Note the non-linear x-axis.

> **RQ2**: The metric estimation model requires $\sim$500k parallel sentences and the associated metric scores before the performance starts to plateau. For very low-resource scenarios, it is possible to reduce the ME model data requirements by utilizing multiple hypotheses for a single source sentence.

	ME All		ME Text	
Model	**Transfer**	**Auth.**	**Transfer**	**Auth.**
Original	60.4%		56.1%	
W16Conv.	54.2%	57.3%	48.4%	51.0%
W16Trans.	54.2%	57.1%	47.9%	50.7%
W17Conv.	58.7%	60.8%	55.1%	55.7%
T5	19.8%	72.9%	43.8%	70.9%

Table 2: BLEU estimation correlation for texts translated by different models than WMT19 Transformer (original).

4.4 Generalization of ME across MT systems

In this section, we examine whether our ME model overfits on the specific errors the used MT system is doing or whether it generalizes and is applicable to also other MT systems. This is important in deployment so that the possible ME/QE model is not dependent on a specific MT system.

We translate the same data using the following English → German models and store the decoder features: T5-small (Raffel et al., 2020), WMT16 Convolutional (Wu et al., 2019), WMT16 Transformer (Ott et al., 2018), WMT17 Convolutional (Gehring et al., 2017). We then run the ME model trained on the outputs of the original WMT19 system to predict metric scores for the outputs of these models. We show the results in Table 2, for both text-only and feature-enriched models. The evaluation of translations by different MT systems shows a varying decrease in correlation with the automated metrics. For most systems, the drop was ∼2–3%, which means that the metric estimator generalizes well across MT systems. However, an exception is using a completely different model, the prompted T5 LM, for which the transfer mostly failed. When training metric estimators on the T5 translated data, they achieved ∼71% correlation but when models trained on outputs of a different MT system were used, the correlation dropped to 19.8% and 43.8% for *ME All* and *ME Text*, respectively. One of the reasons may be vastly different extra features, as documented by the noticeably higher correlation for the *ME Text* model.

RQ3: It is possible to train a metric estimation model on sentences produced by MT_1 and use it to estimate metric scores of MT_2 with a drop in correlation of $\sim 3\%$.

4.5 From ME to QE

This section verifies empirically whether pre-training on ME helps on the QE task. For this, we use models trained on estimating automatic metrics and either: correlate their outputs with human judgements (zero-shot), or fine-tune them to predict the human judgements directly (fine-tuning). The fine-tuning was done with the same setup as in Appendix B with all model parameters updated.

The results in Table 3 show that fine-tuning definitely improved the performance over zero-shot. However, only TER was able to outperform training on z-scores from scratch. This can be attributed to it being the only metric with reasonable absolute correlation in the zero-shot. Notably, COMET was better than the rest of the automated metrics with the worst being BLEU. Despite the fine-tuning, we were not able to construct a QE system that would outperform the standard baseline of COMET-QE ($\rho = 29.2\%$).

Pre-train metric	Zero-shot	Fine-tuning
BLEU	-3.2%	-0.9%
BLEURT	5.9%	-5.2%
ChrF	-6.9%	6.5%
METEOR	-4.2%	6.0%
COMET	1.1%	10.4%
TER	-12.5%	22.8% ⋆
human	19.8%	

Table 3: Correlations with human judgement (z-scores) from models (with extra features fusion) which were pre-trained on metric estimation. Only the magnitude (absolute value) of the correlation is important.

RQ4: Pre-training on TER (large data) and fine-tuning on human scores (small data) is better than training only on human scores (small data).

Further experiments with limited target-domain data (Figure 8) show that the proposed pre-training & fine-tuning regime does not perform well even with less fine-tuning data. The same figure also shows fine-tuning sensitivity to the selected data. Variance is caused by both the optimization process and data subsampling. Even though we include confidence intervals, in deployment one would start multiple runs and use the best-performing one. A striking observation is that very little human-annotated data is needed for training. Further research should more closely examine the relationship between model capacity, data requirements and QE performance.

	ME		QE	
Metric	**Single**	**Multi**	**Single**	**Multi**
BLEU	60.4%	47.0%	15.5%	23.9%
BLEURT	59.7%	5.7%	26.7%	5.7%
ChrF	58.5%	42.6%	23.7%	24.0%
METEOR	51.6%	36.5%	22.6%	24.5%
COMET	44.4%	23.1%	13.1%	21.0%
TER	37.4%	36.8%	7.6%	18.7%
human	-	-	19.8%	5.5%

Table 4: Pearson correlations between system outputs and automated metrics or human z-scores for either multiple single-metric models (Single) or a single multi-target model (Multi).

4.6 Joint prediction of multiple metrics

In this section, we investigate using all the automated metrics at the same time in a single model instead of multiple individual models.

For both the WMT News and QE datasets (individually), all the metric scores for a single segment can be predicted at the same time. Instead of training 6+7 individual models to predict each metric, we train two models (for ME and QE data) that predict all available metrics at once (similar to BLEURT pre-training phase) using different regression heads. The only difference in the architecture from Appendix B is that the last linear layer has 6 or 7 output neurons instead of one. The loss for model f is then defined as $\sum_{m \in \text{metrics}} \text{MSE}(f_m(s,t), m(s,t,r))$. Having multiple targets in a single training can provide more signal and better representation (Aho et al., 2012; Korneva and Blockeel, 2020). The results shown in Table 4 demonstrate that for the smaller dataset (QE), joint learning mostly helps in metric prediction but not in human z-score prediction. This may be because of a loss imbalance of 6 target outputs optimizing on automated metrics and only 1 target output optimizing on human z-scores.

5 Complexity & Fluency Estimation

Currently, our model was dependent on mostly the source and the hypothesis. If it had access to only the hypothesis, it could still consider its fluency and other factors in estimating the metric. Likewise, having access to only the source would correspond to sentence difficulty/complexity estimation. Similarly to Wan et al. (2022), we explored both of these modes and found very high sentence-level correlations.

$$\rho(\{(f_{\text{ME}}(h), f_{\text{BLEU}}(h,r))\}) = 55.5\%$$
$$\rho(\{(f_{\text{ME}}(s), f_{\text{BLEU}}(h,r))\}) = 55.3\%$$
$$h, s, r \in \mathcal{D}$$

These high correlations, which are close to the full text-only model's performance ($\rho = 56.5\%$) show that our model is not able to utilize the relationship between the source and the hypothesis and that a more elaborate models should be considered. These results are in line with general findings of Behnke et al. (2022). The models' inadequacy is also shown by the imperfect performance when given access to the hypothesis and the reference, just as the metric has:

$$\rho(\{(f_{\text{ME}}(s,h,r), f_{\text{BLEU}}(h,r))\}) = 60.7\%$$
$$\rho(\{(f_{\text{ME}}(h,r), f_{\text{BLEU}}(h,r))\}) = 61.5\%$$
$$h, s, r \in \mathcal{D}$$

However, the focus solely on the hypothesis or just the source itself has been confirmed for also other QE systems (Sun et al., 2020) and our model is not an outlier.

6 Related Work

This section discusses how our proposed ME task fits in the field of QE.

Confidence estimation. The task of ME has a connection to an older task of confidence estimation (Blatz et al., 2004), which predates QE (Specia et al., 2013). In confidence estimation, the goal is to predict the probability of the output being correct. Blatz et al. (2004) define correctness as a binary class which is based on two thresholded MT metrics: word error rate and NIST (Doddington, 2002). This is in contrast to the ME task which is a regression task (predicting e.g. 0.7 instead of `GOOD` and segment-level).

More recent works use the term confidence estimation more freely to mean essentially the QE task with full training data and model access (Chelba et al., 2020). Because this term is used also in other contexts, such as in calibration (Wan et al., 2020; Wang et al., 2020), we define the task metric estimation to avoid ambiguity.

Feature-based QE models. Specia et al. (2010, 2013) pioneered the work of mainstream MT QE. The QuEst model uses support vector regression on top of features such as source & target lengths, the number of translations in a phrase table or target sentence language model probability. Further research has been devoted to devising good features for MT QE models, such as grammatical ones (Felice and Specia, 2012), ones based on the decoder (Avramidis, 2012; Fomicheva et al., 2020) or based on the model embeddings spaces (Shah et al., 2016; Chen et al., 2017).

Deeper QE models. QUETCH (Kreutzer et al., 2015), NuQE (Martins et al., 2016), DeepQuest (Ive et al., 2018) and others (Kim and Lee, 2016; Li et al., 2018) regress directly from the source and hypothesis texts into a score. Notably some systems approach sentence-level QE by aggregating or otherwise utilizing previously-estimated word-level QE predictions (Kepler et al., 2019).

Pre-training QE models. Closest to our work is computing TER between the hypothesis and its post-edited version (Heo et al., 2021). QE is then trained jointly on the artificial and authentic QE data. Other models are first trained on artificial (pre-training) and then authentic data (Baek et al., 2020; Cui et al., 2021; Yankovskaya and Fishel, 2021). The pre-training task does not have to be tied to QE. Large pre-trained language models have also been used for QE (Hu et al., 2020; Moura et al., 2020; Nakamachi et al., 2020; Eo et al., 2021). In contrast, our pre-training aims not only to acquire better sentence representations in general but specifically to acquire better sentence representations for translation quality score estimation.

QE models outside of MT The idea to estimate the quality of a prediction given only itself and the input has also been applied to other NLP tasks. RUBER (Tao et al., 2018) uses an RNN-based model to compute a score for a context-response pair in a dialog that is combined with a reference-based score to obtain the final metric. BERT-RUBER extends this by using pre-trained representations and GRADE uses the unreferenced scorer in combination with a module that constructs a conversational graph using ConceptNet concepts and reasons over it (Huang et al., 2020). Recently different works evaluate responses based on specific quality attributes, for example, groundedness (Honovich et al., 2021), or combine them into one quality score (Pang et al., 2020). For open-ended tasks, reference-free metrics, similar to QE, are more desirable because they make fewer assumptions about how the hypothesis should look like. Compared to those tasks, the admissible hypothesis space in MT for a given source sentence is more constrained.

7 Discussion

The experiments have shown that the outputs of automated metrics can be predicted even without access to the reference. Pre-training only on TER and not any other automated metric outperformed training from scratch, which highlights the importance of exploring multiple metrics instead of just one or two of the most popular ones. Pre-training on TER helped because of the large absolute zero-shot correlation. We are, however, unable to provide an explanation for this zero-shot correlation in the first place.

Our ME approach can also be potentially used for improving models, even outside of the MT field, by providing additional signals through self-supervision. For example, the ME model could be used instead of the decoder probabilities for reranking when decoding with beam-search in generative models. This approach would be similar to using QE in decoding by Fernandes et al. (2022) but would not require a separate scorer model.

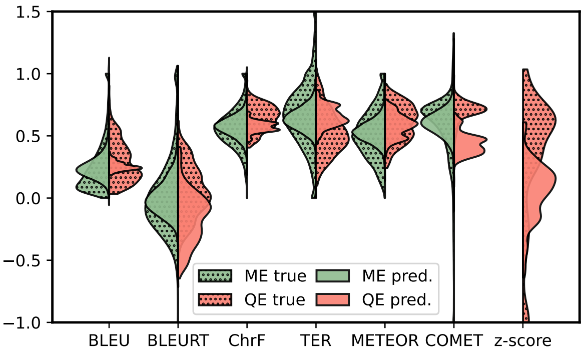

Figure 6: Distribution of metric and human judgement values in the used ME and QE datasets together with their predictions. Long tails clipped between -1 and 1.5 for higher resolution.

7.1 Error analysis

In this section, we examine three model predictions and comment on their comparison with the target scores. At the first glance, the model predictions are generally more conservative (lower variance and concentration around the average), as shown in Figure 6. This is however not an issue when evaluating with Pearson's correlation coefficient as the distributions are rescaled. While the error analysis serves as a good check for the model outputs, we are unable to clearly define specific failure modes of the model, which does not achieve 100% metric correlation.

> **Example 1** **BLEU** 0.08, **ME** 0.07
>
> **Source:** *Police try new, less-lethal tools as protests continue.*
> **Reference:** *Während die Proteste weitergehen, testet die Polizei weniger tödliche Geräte*
> **Reference (lit.):** *While the protest continue, the police is testing less-lethal devices.*
> **Hypothesis:** *Die Polizei probiert neue, weniger tödliche Werkzeuge aus, während die Proteste anhalten.*
> **Hypothesis (lit.):** *The police is testing new, less-lethal tools, as the protest persist.*

We first examine cases in which we compare our model's prediction with the true metric value.[6] Example 1 is an almost exact match in BLEU and model prediction. Although the hypothesis is reasonable, it is too literal of a translation. This is a failure for the automated metric but a success for our model because it predicted the metric accurately.

[6]We report true BLEU scores and not percentages (i.e. scale is 0-1, not 0-100).

> **Example 2** **BLEU** 0.67, **ME** -0.51
>
> **Source:** *It didn't work.*
> **Reference:** *Das hat nicht funktioniert.*
> **Reference (lit.):** *That did not function.*
> **Hypothesis:** *IT hat nicht funktioniert.*
> **Hypothesis (lit.):** *THAT/IT did not function.*

In some error cases, the model prediction was a better quality estimate than the metric it was trained to estimate. In Example 2, the word *it* is mistaken for a named entity which distorts the sentence meaning: *it* - pronoun, *IT* - abbreviation *Information Technology* (can also mean the tech support department).

> **Example 3** **Human** -4.72, **QE** -0.42
>
> **Source:** *There's mask-shaming and then there's full on assault.*
> **Reference:** *Masken-Shaming ist eine Sache, Körperverletzung eine andere.*
> **Reference (lit.):** *Mask-shaming is one thing, body assault is a different one.*
> **Hypothesis:** *Es gibt Masken-Beschämen und dann gibt es voll auf Angriff.*
> **Hypothesis (lit.):** *There is mask-humiliation and then there is full-on assault.*

In certain cases, the model output is very far from human judgement. In Example 3, the hypothesis contains a phrase *voll auf Angriffe* which seems like a good translation but only at the first sight and is, in fact, incorrect (word-wise translation *full - voll, on - auf, assault - Angriff*). This may be a reason for the low score by the human annotator and it was not captured at all by our model.

7.2 Negative results

We also attempted to leverage mBERT representations in the main LSTM-based model via concatenation fusion with the last hidden state, identical to the approach of Zouhar et al. (2022) for language modelling. However, the results (ρ = 60.0% for BLEU) were on par with the main LSTM-based model alone (ρ = 60.4% for BLEU).

We experimented with the expressivity of the used model architecture in Section 5 by granting the model access to the source, the reference and the hypothesis. This should provide sufficient information to be able to learn the specific metrics, however, the model was unable to fit it perfectly. Approaches in future work should therefore use models with larger capacities and explore specific mechanics that could be useful for predicting automated metrics relying on n-gram overlaps.

[6]Training one full model takes $\sim$3 hours on 500k sentences on NVIDIA GeForce GTX 1080 Ti. Machine translation of the same data takes 120 GPU hours on the same model.

8 Conclusion

We proposed the task of **metric estimation** for machine translation, as a parallel to quality estimation and attempted to solve it with a baseline BiLSTM model. We show that it is possible to predict the output of a metric without even seeing the reference (ρ = 60.4% for BLEU and $\bar{\rho}$ = 51.3% for other metrics). The main advantage of this task compared to QE is that the data for training ME models to predict a particular metric can be generated from any parallel corpus on which the metric can be run. While pre-training on TER outperformed training from scratch, it did not perform better than the commonly used baseline, COMET-QE.

Future work. Despite the negative results, features in the hypothesis space should be more explored for tasks beyond ME/QE, such as calibration of self-reported confidence in generative models. Metric and quality estimation could also be a part of the MT system itself (e.g. as a separate head) which would alleviate the need for an external ME/QE model. The ME models should also be evaluated for cross-domain performance, similarly to our cross-system evaluation as motivated by needs of production settings. Non-perfect correlation with the metrics when presented with the same input shows the imporants of exploring more complex architectures or optimization approaches.

Limitations

Although our model outperformed the simple baselines in ME and QE, it provides less explainability because a specific QE output can not be linked easily to input features. The model also required much longer training[6] while the baselines just need a simple featurizer, MT intrinsic features and can run the linear regression fitting on a CPU. Nevertheless, the largest computational bottleneck in this research has been running the MT system inference[6] rather than training the individual models.

Concerns have long been raised about using segment-level metrics/evaluations because of the large variance (Lavie, 2010). However, we find that for automated metrics, our models are still able to deal with this variance.

Ethics statement

Detailed error analysis should always be performed before deploying a quality estimation system in machine translation production pipelines.

Acknowledgements

We thank Ricardo Rei (Unbabel), Jonas Belouadi (Universität Bielefeld) and Florian Schottmann (TextShuttle) for their proofreading, discussions and comments on our work.

References

Timo Aho, Bernard Ženko, Sašo Džzeroski, Tapio Elomaa, and Carla Brodley. 2012. Multi-target regression with rule ensembles. *Journal of Machine Learning Research*, 13(8).

Eleftherios Avramidis. 2012. Quality estimation for machine translation output using linguistic analysis and decoding features. In *Proceedings of the seventh workshop on statistical machine translation*, pages 84–90.

Yujin Baek, Zae Myung Kim, Jihyung Moon, Hyunjoong Kim, and Eunjeong Park. 2020. Patquest: Papago translation quality estimation. In *Proceedings of the Fifth Conference on Machine Translation*, pages 991–998.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan.

Marta Bañón, Pinzhen Chen, Barry Haddow, Kenneth Heafield, Hieu Hoang, Miquel Esplà-Gomis, Mikel L. Forcada, Amir Kamran, Faheem Kirefu, Philipp Koehn, Sergio Ortiz Rojas, Leopoldo Pla Sempere, Gema Ramírez-Sánchez, Elsa Sarrías, Marek Strelec, Brian Thompson, William Waites, Dion Wiggins, and Jaume Zaragoza. 2020. ParaCrawl: Web-scale acquisition of parallel corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4555–4567, Online.

Hanna Behnke, Marina Fomicheva, and Lucia Specia. 2022. Bias mitigation in machine translation quality estimation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1475–1487.

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.

Ondrej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, Radu Soricut, Lucia Specia, and Aleš Tamchyna. 2014. Findings of the 2014 workshop on statistical machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 12–58, Baltimore, Maryland, USA.

Ciprian Chelba, Hideto Kazawa, Jeff Klingner, Junpei Zhou, Mengmeng Niu, and Music Li. 2020. Data troubles in sentence level confidence estimation for machine translation.

Zhiming Chen, Yiming Tan, Chenlin Zhang, Qingyu Xiang, Lilin Zhang, Maoxi Li, and Mingwen Wang. 2017. Improving machine translation quality estimation with neural network features. In *Proceedings of the Second Conference on Machine Translation*, pages 551–555.

Qu Cui, Shujian Huang, Jiahuan Li, Xiang Geng, Zaixiang Zheng, Guoping Huang, and Jiajun Chen. 2021. Directqe: Direct pretraining for machine translation quality estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 12719–12727.

José GC de Souza, Marco Turchi, and Matteo Negri. 2014. Machine translation quality estimation across domains. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 409–420.

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, Minneapolis, Minnesota.

George Doddington. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the second international conference on Human Language Technology Research*, pages 138–145.

Sugyeong Eo, Chanjun Park, Hyeonseok Moon, Jaehyung Seo, and Heuiseok Lim. 2021. Comparative analysis of current approaches to quality estimation for neural machine translation. *Applied Sciences*, 11(14):6584.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, et al. 2021. Beyond english-centric multilingual machine translation. *J. Mach. Learn. Res.*, 22(107):1–48.

Mariano Felice and Lucia Specia. 2012. Linguistic features for quality estimation. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 96–103.

Patrick Fernandes, António Farinhas, Ricardo Rei, José GC de Souza, Perez Ogayo, Graham Neubig, and André FT Martins. 2022. Quality-aware decoding for neural machine translation. *arXiv preprint arXiv:2205.00978*.

Marina Fomicheva, Shuo Sun, Lisa Yankovskaya, Frédéric Blain, Francisco Guzmán, Mark Fishel, Nikolaos Aletras, Vishrav Chaudhary, and Lucia Specia. 2020. Unsupervised quality estimation for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:539–555.

Philip Gage. 1994. A new algorithm for data compression. *C Users Journal*, 12(2):23–38.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin. 2017. Convolutional sequence to sequence learning. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1243–1252.

Taisiya Glushkova, Chrysoula Zerva, Ricardo Rei, and André FT Martins. 2021. Uncertainty-aware machine translation evaluation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3920–3938.

Dam Heo, WonKee Lee, Baikjin Jung, and Jong-Hyeok Lee. 2021. Quality estimation using dual encoders with transfer learning. In *Proceedings of the Sixth Conference on Machine Translation*, pages 920–927.

Or Honovich, Leshem Choshen, Roee Aharoni, Ella Neeman, Idan Szpektor, and Omri Abend. 2021. q^2: Evaluating factual consistency in knowledge-grounded dialogues via question generation and question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7856–7870, Online and Punta Cana, Dominican Republic.

Chi Hu, Hui Liu, Kai Feng, Chen Xu, Nuo Xu, Zefan Zhou, Shiqin Yan, Yingfeng Luo, Chenglong Wang, Xia Meng, et al. 2020. The niutrans system for the WMT20 quality estimation shared task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 1018–1023.

Lishan Huang, Zheng Ye, Jinghui Qin, Liang Lin, and Xiaodan Liang. 2020. GRADE: Automatic graph-enhanced coherence metric for evaluating open-domain dialogue systems. *arXiv preprint arXiv:2010.03994*.

Julia Ive, Frédéric Blain, and Lucia Specia. 2018. DeepQuest: A framework for neural-based quality estimation. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3146–3157.

Fábio Kepler, Jonay Trénous, Marcos Treviso, Miguel Vera, and André FT Martins. 2019. OpenKiwi: An open source framework for quality estimation. *ACL 2019*, page 117.

Hyun Kim and Jong-Hyeok Lee. 2016. Recurrent neural network based translation quality estimation. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 787–792.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Tom Kocmi, Christian Federmann, Roman Grundkiewicz, Marcin Junczys-Dowmunt, Hitokazu Matsushita, and Arul Menezes. 2021. To ship or not to ship: An extensive evaluation of automatic metrics for machine translation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 478–494.

Evgeniya Korneva and Hendrik Blockeel. 2020. Towards better evaluation of multi-target regression models. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 353–362. Springer.

Julia Kreutzer, Shigehiko Schamoni, and Stefan Riezler. 2015. Quality estimation from scratch (quetch): Deep learning for word-level translation quality estimation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 316–322.

Alon Lavie. 2010. Evaluating the output of machine translation systems. In *Proceedings of the 9th Conference of the Association for Machine Translation in the Americas: Tutorials*.

Maoxi Li, Qingyu Xiang, Zhiming Chen, and Mingwen Wang. 2018. A unified neural network for quality estimation of machine translation. *IEICE TRANSACTIONS on Information and Systems*, 101(9):2417–2421.

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.

André FT Martins, Ramón Fernandez Astudillo, Chris Hokamp, and Fabio Kepler. 2016. Unbabel's participation in the wmt16 word-level translation quality estimation shared task. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 806–811.

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

Joao Moura, Miguel Vera, Daan van Stigt, Fabio Kepler, and André FT Martins. 2020. Ist-unbabel participation in the wmt20 quality estimation shared task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 1029–1036.

Elena Murgolo, Javad Pourmostafa Roshan Sharami, and Dimitar Shterionov. 2022. A quality estimation and quality evaluation tool for the translation industry. In *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, pages 305–306.

Akifumi Nakamachi, Hiroki Shimanaka, Tomoyuki Kajiwara, and Mamoru Komachi. 2020. Tmuou submission for wmt20 quality estimation shared task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 1037–1041.

Nathan Ng, Kyra Yee, Alexei Baevski, Myle Ott, Michael Auli, and Sergey Edunov. 2019. Facebook FAIR's WMT19 news translation task submission. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 314–319.

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 the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53.

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

Bo Pang, Erik Nijkamp, Wenjuan Han, Linqi Zhou, Yixian Liu, and Kewei Tu. 2020. Towards holistic and automatic evaluation of open-domain dialogue generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3619–3629, Online.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Maja Popović. 2015. chrF: character n-gram f-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020a. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020b. Unbabel's participation in the WMT20 metrics shared task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 909–918, Online.

Raphael Rubino and Eiichiro Sumita. 2020. Intermediate self-supervised learning for machine translation quality estimation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4355–4360.

Holger Schwenk, Guillaume Wenzek, Sergey Edunov, Édouard Grave, Armand Joulin, and Angela Fan. 2021. CCMatrix: Mining billions of high-quality parallel sentences on the web. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6490–6500.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. *ArXiv*, abs/1508.07909.

Kashif Shah, Fethi Bougares, Loïc Barrault, and Lucia Specia. 2016. Shef-lium-nn: Sentence level quality estimation with neural network features. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 838–842.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231.

Lucia Specia, Frédéric Blain, Marina Fomicheva, Erick Fonseca, Vishrav Chaudhary, Francisco Guzmán, and André F. T. Martins. 2020. Findings of the WMT 2020 shared task on quality estimation. In *Proceedings of the Fifth Conference on Machine Translation*, pages 743–764, Online.

Lucia Specia, Frédéric Blain, Marina Fomicheva, Chrysoula Zerva, Zhenhao Li, Vishrav Chaudhary, and André FT Martins. 2021. Findings of the WMT 2021 shared task on quality estimation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 684–725.

Lucia Specia, Dhwaj Raj, and Marco Turchi. 2010. Machine translation evaluation versus quality estimation. *Machine translation*, 24(1):39–50.

Lucia Specia, Carolina Scarton, and Gustavo Henrique Paetzold. 2018. Quality estimation for machine translation. *Synthesis Lectures on Human Language Technologies*, 11(1):1–162.

Lucia Specia, Kashif Shah, José GC 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.

Shuo Sun, Francisco Guzmán, and Lucia Specia. 2020. Are we estimating or guesstimating translation quality? In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 6262–6267.

Aleš Tamchyna. 2021. Deploying MT quality estimation on a large scale: Lessons learned and open questions. In *Proceedings of Machine Translation Summit XVIII: Users and Providers Track*, pages 291–305.

Chongyang Tao, Lili Mou, Dongyan Zhao, and Rui Yan. 2018. Ruber: An unsupervised method for automatic evaluation of open-domain dialog systems. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey.

Yu Wan, Dayiheng Liu, Baosong Yang, Haibo Zhang, Boxing Chen, Derek Wong, and Lidia Chao. 2022. UniTE: Unified translation evaluation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8117–8127.

Yu Wan, Baosong Yang, Derek F Wong, Yikai Zhou, Lidia S Chao, Haibo Zhang, and Boxing Chen. 2020. Self-paced learning for neural machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1074–1080.

Shuo Wang, Zhaopeng Tu, Shuming Shi, and Yang Liu. 2020. On the inference calibration of neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3070–3079.

Felix Wu, Angela Fan, Alexei Baevski, Yann N Dauphin, and Michael Auli. 2019. Pay less attention with lightweight and dynamic convolutions. *arXiv preprint arXiv:1901.10430*.

Haihua Xu, Daniel Povey, Lidia Mangu, and Jie Zhu. 2011. Minimum bayes risk decoding and system combination based on a recursion for edit distance. *Computer Speech & Language*, 25(4):802–828.

Elizaveta Yankovskaya, Andre Tättar, and Mark Fishel. 2018. Quality estimation with force-decoded attention and cross-lingual embeddings. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 816–821.

Lisa Yankovskaya and Mark Fishel. 2021. Direct exploitation of attention weights for translation quality estimation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 955–960.

Vilém Zouhar, Michal Novák, Matúš Žilinec, Ondřej Bojar, Mateo Obregón, Robin L Hill, Frédéric Blain, Marina Fomicheva, Lucia Specia, and Lisa Yankovskaya. 2021. Backtranslation feedback improves user confidence in MT, not quality. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 151–161.

Vilém Zouhar, Marius Mosbach, and Dietrich Klakow. 2022. Fusing sentence embeddings into LSTM-based autoregressive language models.

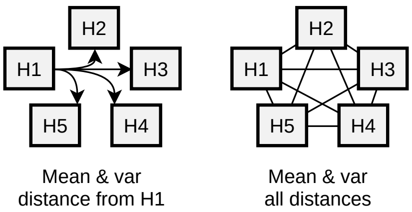

Figure 7: Four features based on properties of the generated hypothesis space: mean and variance of either H1 (top hypothesis) or all pairs. The used metric (any hyp-ref-based) considers one of the hypotheses in a pair as the reference.

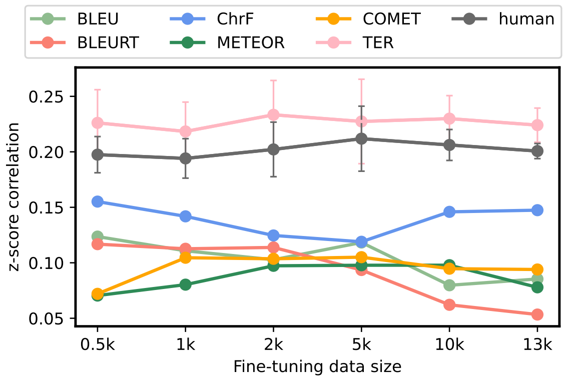

Figure 8: Finetuning on limited target-domain QE data. Note the non-linear x-axis. Each point is an average of 5 runs. Error bars show a 95% confidence interval of the mean. Error bars shown only for BLEU and human for clarity.

A Reproducibility

We compute metric scores using SacreBLEU (Post, 2018) with the following signatures:

- All: `nrefs:1 | version:2.2.0`
- BLEU: `case:mixed | eff:yes | tok:13a | smooth:exp`
- ChrF: `case:mixed | eff:yes | nc:6 | nw:0 | space:no`
- TER: `case:lc | tok:tercom | norm:no | punct:yes | asian:no`

For baseline experiments, multilingual version of BERT was used: `bert-base-multilingual-cased`. Translations using T5-small are done with the prefix "*translate English to German:*". For other translation models, we used the following models available on torch.hub under the `pytorch/fairseq` namespace (Ott et al., 2019):

- `dynamicconv.glu.wmt16.en-de`
- `conv.wmt17.en-de`
- `transformer.wmt16.en-de`
- `transformer.wmt18.en-de`

B Model details

The metric/quality estimation model specifics are shown in Table 5. Additionally, we concatenate all forward and backward hidden states from both LSTM layers. A non-standard choice was to use sigmoid as the final activation function which worked better than just the linear output.[7] However, we also rescaled and centered it in case of BLEU, ChrF and METEOR so that scores of 0 and 100 are attainable. We eventually did not use this in the main experiments so that a single model setup could be used for all metrics. All models are trained with early stopping of 10 epochs. The optimization loss is mean squared error. Our used model is fairly small in comparison to other models, such as those utilizing the Transformer architecture. This was an intentional choice with respect to the small amount of data used. The hyperparameters were chosen manually by best practices with respect to final metric correlation (5 trials).

Optimizer	Adam (Kingma and Ba, 2015)
Learning rate	10^{-6}
Batch size	10 (0-padded to longest)
Vocab size (BPE)	8192
Vocab embedding	512
LSTM	Hidden state 128
	2 bidirectional layers
LSTM dropout	20% inter-layer
	75% final hidden state
Fusion	Concatenate (512 + 6)
Linear	$518 \rightarrow 100$
Activation	ReLU
Linear	$100 \rightarrow 1$ (6/7 for multi)

Table 5: Metric/quality estimation model details.

The TF-IDF featurizer in the linear regression TF-IDF baseline uses variable maximum number of features and the best-performing one is chosen. The search is logarithmical from 2^4 to 2^{14}.

[7] This is dissimilar to the baseline model where linear regression worked better than logistic regression even in the case of metrics with bounded output range.

C Results for other language pairs

The paper used figures and examples from the English→German language direction. To replicate the findings, we translate 500k sentences for the following language directions, models and datasets:[8]

- German ↔ English (WMT14)
  `transformer.wmt19.{de-en,en-de}`
- German ↔ Polish (opus_paracrawl)
  `Helsinki-NLP/opus-mt-{de-pl,pl-de}`
- Chinese ↔ English (CCMatrix)
  `Helsinki-NLP/opus-mt-{zh-en,en-zh}`
- Czech ↔ English (WMT14)
  `Helsinki-NLP/opus-mt-{cs-en,en-cs}`
- Russian ↔ English (WMT14)
  `transformer.wmt19.{en-ru,ru-en}`
- French ↔ English (WMT14)
  `Helsinki-NLP/opus-mt-{fr-en,en-fr}`
- Hindi → English (CCMatrix)
  `Helsinki-NLP/opus-mt-{hi-en,en-hi}`

The results for metric estimation for all metrics and language directions are shown in Figure 9. In comparison to the other languages, the chosen language pair in the paper (En → De) is more conservative than the other language pairs, which achieve higher correlations across most metrics. Most of them achieve $> 50\%$ correlation, though specifically *COMET* appears to be more predictable for other language pairs. These results confirm the main results of predictability of the metric without having access to the reference.

D Confidence Estimation for Metric Estimation

Quantifying the confidence of the prediction can be crucial in downstream applications. For example, if a quality estimator has low confidence or is predicted to not be accurate, the decision regarding the quality of translation should be delegated to humans. We attempt to do this by training an auxiliary model to predict the confidence of the output score. We do this by training a logistic regression classifier which takes as input the final layer (λ) of our metric estimator and is trained to predict the binary label: if the predicted metric $f_{\text{BLEU}}(s, h, r)$ is close to the true metric $\text{ME}(s, h)$.

$$\mathbb{P}(|\text{ME} - f_{\text{BLEU}}| \in [0, \text{ME}(s,h) \times 10\%]) \\ = \sigma(W^T \cdot \lambda + b)$$

We, unfortunately, find that the classifier suffers from a very low accuracy 63.4% against a most common class (negative) baseline of 51.6%. It therefore cannot be meaningfully used to ascertain when our regressor is correct and when it is not. Glushkova et al. (2021) propose a more complex solution and exploit uncertainty methods for MT metric and quality estimation systems.

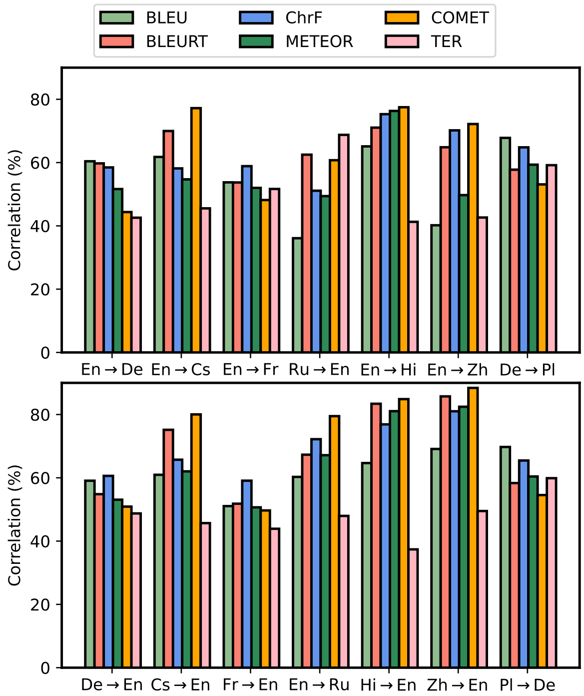

Figure 9: Correlations with metrics and human judgement of main metric estimation (ME) models on other languages. Each bar is a separate model trained to predict a particular metric or human judgement.

[8] WMT14 (Bojar et al., 2014), Opus Paracrawl (Tiedemann, 2012; Bañón et al., 2020), CCMatrix (Schwenk et al., 2021; Fan et al., 2021).