Return to the Source: Assessing Machine Translation Suitability

Francesco Fernicola[1,2], **Silvia Bernardini**[1], **Federico Garcea**[1]
Adriano Ferraresi[1] and **Alberto Barron-Cedeño**[1]
[1] Università di Bologna, Forlì, Italy
[2] Institute for Applied Linguistics, Eurac Research, Bolzano/Bozen, Italy
[francesco.fernicola2, silvia.bernardini, federico.garcea2]@unibo.it
[adriano.ferraresi, a.barron]@unibo.it

Abstract

We approach the task of assessing the suitability of a source text for translation by transferring the knowledge from established MT evaluation metrics to a model able to predict MT quality *a priori* from the source text alone. To open the door to experiments in this regard, we depart from reference English–German parallel corpora to build a corpus of 14,253 source text–quality score tuples. The tuples include four state-of-the-art metrics: cushLEPOR, BERTScore, COMET, and TransQuest. With this new resource at hand, we fine-tune XLM-RoBERTa, both in a single-task and a multi-task setting, to predict these evaluation scores from the source text alone. Results for this methodology are promising, with the single-task model able to approximate well-established MT evaluation and quality estimation metrics —without looking at the actual machine translations— achieving low Root Mean Square Error values in the [0.1–0.2] range and Pearson's correlation scores up to 0.688.

1 Introduction

There are many factors in play when assessing the suitability of a text for machine translation (MT). Readability might account for part of the problem, but the metrics designed for its estimation aim at assessing the level of education necessary to understand a given text, from a monolingual perspective (Gunning, 1969). As evidenced by Vanroy et al. (2019), there is a clear-cut distinction between *translatability*, "the difficulty of a translation task", and *readability*, "the difficulty of a monolingual text". They argue that, although the two might overlap in some regards, a translation task cannot be solely defined based on monolingual features. Their study is centred on human translation (HT), but given that MT and post-editing (PE) represent the strongest future trend for both industry and academia, according to the latest ELIS language industry report (European Language Industry Survey Research, 2022), our work seeks to advance the discussion in the field of MT.

In fact, although quality improvements over the last few years have indeed been significant, the translation world has expressed a need, time and time again, for new methods and technologies to properly assess its quality (Kocmi et al., 2021). Most of the previous work in this regard has focused on the target translation; both in the reference-based machine translation evaluation (MTE), where the machine-translated segment is compared against a human reference, and in the more recent quality estimation techniques (QE), where the machine-translated segment is evaluated without any reference (Freitag et al., 2021; Specia et al., 2021).

This paper seeks a different perspective, switching the focus to the source text, to assess whether a given segment will produce a high quality machine translation. We define this task as Machine Translation Suitability. Existing MTE and QE techniques either use a reference translation or an MT output, meaning they both require to first translate all the segments with MT system in order to obtain a quality evaluation. Many such segments will inevitably not meet the desired quality and will be discarded, constituing a net loss. Given that most commercial MT systems are paid by word, our approach would

serve to reduce the costs of the overall system by avoiding to send certain segments to MT, thus creating a more efficient production pipeline. Moreover, recent studies have also pointed towards a lower lexical variation of post-edited MT segments, as well as an overall lower quality of those segments with respect to translations from scratch (Volkart and Bouillon, 2022), while others highlight the challenges of generating comprehensive guidelines for post-editors, especially regarding what constitutes an error in a given scenario and how to correctly provide quality assurance for such segments (Nunziatini and Marg, 2020). Therefore, the presence of an additional evaluation step before generating the machine-translated segments would help avoid having to undergo an expensive PE step or reroute to human translation. Lastly, applying such a model could reduce the pipeline's carbon footprint, because it would not need to compute a translation using large, resource heavy models.

With the purpose of advancing research in this field, we thus formulate the following research question:

> ***RQ:*** *Is it possible to accurately predict the MTE or QE score of a translation from the source text alone?*

In order to give light to the RQ, we compile an ad-hoc corpus pairing source segments with the evaluation scores of their automatic translations in the English–German language pair from one of the most prominent MT engines available: ModernMT[1]. We select two reference-based evaluation metrics and two quality estimation metrics: cushLEPOR, BERTScore, COMET, and TransQuest, according to the state of the art (Freitag et al., 2021; Specia et al., 2021). We frame the task as a regression problem and fine-tune our model to reproduce the evaluation score by looking at the source text alone. The experiments are conducted using the multilingual model XLM-RoBERTa (XLM-R) (Conneau et al., 2020)[2] and approach the task in two different settings: single-task and multi-task. In the former, a model is fine-tuned on each evaluation score individually, whereas in the latter, a model is trained on all four scores to exploit the shared knowledge among the different metrics.

By achieving low RMSE values in the [0.1–0.2] range and Pearson correlation scores up to 0.688, our results are promising and indicate that it is indeed possible to distil the knowledge acquired from different MT evaluation metrics into a model trained solely on the source text, thus confirming our RQ.

2 Related Work

Nowadays, the state of the art is divided between MTE metrics, similar to BLEU (Marie et al., 2021; Papineni et al., 2002; Post, 2018), which employ the source text, target text and a reference translation, and QE metrics which assess quality without looking at a reference (Specia et al., 2021).

Some of the most prominent reference-based metrics include cushLEPOR, an n-gram based metric whose parameters are automatically tuned using pre-trained language models (Han et al., 2021), and BERTScore, which exploits embedding similarity and has been shown to highly correlate with human judgments on sentence-level and system-level evaluation (Zhang et al., 2020; Freitag et al., 2021).

Being somewhat new, the field of QE achieved impressive results in the past few years by employing multilingual pre-trained representations from very large language models to generate their predictions. Nevertheless, it instead appears to have no single metric being consistently deployed to production in either the industry or institutions, with the only exception being COMET, which has consistently achieved top scores for three years in a row in the annual WMT QE shared task (Specia et al., 2021; Zerva et al., 2022).

Both MTE and QE metrics, though, depend on the underlying target translation produced by an MT engine and research specifically focused on the source text has been limited. Vanroy et al. (2019) aimed at developing a "translatability prediction system". It assigns a global difficulty score to a source text and identifies which passages are more problematic for translation. Albeit promising, this work solely addressed human translation difficulty and no study tailored to MT has been published yet.

SmartLQA (Smart Linguistic Quality Assessment), aims at analysing the impact of the source text on MT (Yanishevsky, 2021). It handles the prediction of *at-risk* content prior to translation, identifying the most problematic linguistic aspects within the source text via linguistic features and readability tests, such as the Flesch–Kincaid met-

[1] https://github.com/modernmt/modernmt

[2] We use a multilingual model instead of a monolingual one in order to have a realistic baseline and to facilitate future work in multiple language pairs.

ric (Kincaid et al., 1975). They conclude that poor source-text quality leads to poor target-text quality. To the best of our knowledge, no predictive model using these features has been proposed.

Additional work in this direction was carried out by Cambra and Nunziatini (2022), who use the source segment and MT training data to approximate translation quality without the target. Their method is based on the assumption that a similarity can be found between the source segment to be translated and the underlying data seen by the MT system. They employ either a bag-of-word representation or the "all-mpnet-base-v2" sentence transformer model (Song et al., 2020) to encode both the source and the training segments and apply similarity metrics on their vectorial representations, also accounting for words unknown to the MT system. Their technique achieves results comparable to QE metrics. Similarly, Tezcan (2022) shows how fuzzy matches retrieved from the training data can be highly informative for predicting sentence-level quality of a given MT model.

Another recent paper instead proposed a new task, called PreQuEL: Pre-Quality-Estimation Learning (Don-Yehiya et al., 2022), namely predicting the likelihood of an MT system to correctly translate a sentence in a given target language. They, too, entirely focus on the input text and their method also proposes to learn to predict quality evaluation metrics from the source text alone and for this they employ Direct Assessment (DA) scores from the WMT shared task on QE (Zerva et al., 2022). Additionally, they use the open-source Marian-MT (Junczys-Dowmunt et al., 2018) rather than commercial systems. Although we recognize that using quality DA scores would lead to more reliable target scores, these are not available for commercial systems, as the authors also point out. While we share the same objective, our attempt bypasses the need for manual evaluation to understand whether a large transformer model would be able to predict state-of-the-art MTE/QE scores, and instead uses a small pool of automatically scored data. Additionally, they employ the monolingual RoBERTa architecture, which limits their experiments to be carried out on English source texts (Liu et al., 2019). Hence, we opt for the multilingual XLM-R to create a solid baseline which could be easily extended to multiple language pairs and directions.

3 Corpus

In order to produce our corpus, we departed from a collection of parallel segments from OPUS (Tiedemann, 2012), including Europarl[3], Ubuntu[4] and News-commentary v16[5]. We target the English–German language pair because it is especially prominent for both MTE and QE (Specia et al., 2021; Freitag et al., 2021).

Although these corpora have been already extensively used in the literature, their pre-processing is done automatically, without any type of manual corrections. To ensure their quality for our experiments, two additional filtering steps have been carried out on the translation units (TUs), following Koehn et al. (2020). It involved the removal of both very long and very short segments from the corpora, set to a minimum length of 25 characters and a maximum length relative to each corpus and language. We removed outliers with respect to each subcorpus, since we do not deem them informative for modeling translation difficulty in a real use-case. The maximum allowed TU length is determined as:

$$\text{MaxLength} = \frac{1}{n}\sum_{i=1}^{n}\text{len}_i + \sigma\,, \qquad (1)$$

where n is the number of segments in the corpus, len_i is the length of the i-th segment and σ is one unit of the standard deviation over the corpus. Additionally, we applied an adaption of the filtering approach from the open-source version of ModernMT[6]. A TU is also discarded if either the source or the target-segment character length exceeds the length of the other segment by more than 50%. In order to prevent the filter from discarding short valid sentence pairs, an arbitrary value of 15 is added to the initial character count.

We randomly selected a subset of the resulting TUs and generated their automatic translations, on which we could obtain the quality scores to be learned by the model. We used the out-of-the-box NMT system ModernMT, based on the state-of-the-art transformer architecture and trained on a large pool of parallel data (Bertoldi et al., 2018). In order to score the resulting automatic translations, we considered four evaluation metrics:

[3] https://opus.nlpl.eu/Europarl.php
[4] https://opus.nlpl.eu/Ubuntu.php
[5] https://opus.nlpl.eu/News-Commentary.php
[6] https://github.com/modernmt/DataCollection/blob/dev/baseline/filter_hunalign_bitext.py

corpus	train	test	length
Europarl	4,223	528	151.5±90.5
News	4,223	528	137.5±69.3
Ubuntu	4,223	528	33.2±74.6
Total	12,669	1,584	107.4±78.1

Table 1: Statistics of the full corpus, incl. number of instances and average character length of the source segments with their respective standard deviation.

hLEPOR. We used cushLEPOR, a version of hLEPOR with optimised settings for the en>de language pair (Han et al., 2021):[7] Alpha = 2.95, Beta = 2.68, n = 2, weight_elp = 2.95, weight_pos = 11.29, weight_pr = 1.87.

BERTScore. We adopted the official repository release (Zhang et al., 2020).[8]

COMET. Even if the most recent release turns the score within a $[0, 1]$ range, we opted for the early release `wmt20-comet-qe-da`, which provides an unbounded score (Rei et al., 2020).[9]

TransQuest. We used the en>de version `monotransquest-da-en_de-wiki` instead of the multilingual model because of its better performance, as reported in (Ranasinghe et al., 2020a; Ranasinghe et al., 2020b).[10]

For our MT suitability experiments, the source text segments are paired with their respective quality scores by combining only the `source text` and the `scores`. Our objective is to produce a model to predict the quality score from the source text alone. With such a model, it would be possible to know how well an MT engine would translate that segment in advance and thus how "suitable" would it be for machine translation. Figure 1 represents a possible pipeline, including the rerouting step from source text to either MT, MT+PE or HT, depending on the expected quality —suitability— of the machine translation. We partition the corpus into two: 12,669 instances for training and 1584 instances for testing. Table 1 shows its statistics.

Since the original corpora used for this work are open-source and specifically designed for NMT training (Tiedemann, 2012), it is likely that they have already been seen by ModernMT during training. This would be problematic because an attempt at learning MT suitability using these corpora would not necessarily be applicable to unseen texts. Hence, we compare the distributions of the training corpus to those of a new, smaller corpus, whose texts have surely not been seen by the system. If the scores' distribution of this secondary corpus were very similar to that of the training corpus, it would mean that there is no significant difference in the scores of unseen and already seen TUs.

To test this hypothesis, we performed a Mann-Whitney U test on all 4 independent variables (Mann and Whitney, 1947) between our corpus and a collection of texts from Globalvoices for which we had guarantees of not having been used for the training of the MT model. Appendix A contains all the details of the test. In summary, there was no significant difference ($p>0.05$) between the training and the Globalvoices dataset for all metrics except for TransQuest. This gives confidence that both corpora belong to the same non-gaussian distribution, meaning there is no significant difference in the quality scores obtained by texts translated using our training corpus and a corpus containing texts not seen by the MT system.

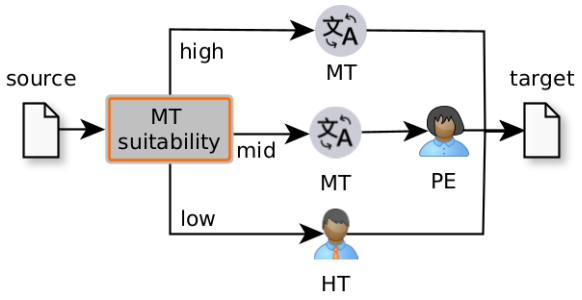

Figure 1: The MT Suitability workflow. A source segment is evaluated by the suitability module and then directed to the appropriate workflow based on quality: MT (high quality), MT+PE (mid quality) or HT (low quality).

4 Experiments

We perform two sets of experiments: once in a single-task setting and once in a multi-task setting. The single-task experiment involves one training session per evaluation metric, thus resulting in four distinct models.

In addition to attempting to learn each of the four metrics independently, we also experiment with Multi-Task Learning (Caruana, 1997) to link the various label representations together instead of training separate models. This approach has been

[7] https://github.com/poethan/cushLEPOR
[8] https://github.com/Tiiiger/bert_score
[9] https://github.com/Unbabel/COMET
[10] https://huggingface.co/TransQuest/monotransquest-da-en_de-wiki

applied to multiple areas of NLP, ranging from the estimation of the check-worthiness of claims in political debates (Vasileva et al., 2019), to a demographic classifier based on features extracted from tweets (Vijayaraghavan et al., 2017) and fine-tuning of transformer models to improve performance on the GLUE benchmark (Karimi Mahabadi et al., 2021). Appendix B includes details on the batch size and other model settings for the multi-task approach, constrained by design decisions and the hardware at hand. Figure 2 offers a representation of the model.

We used `xlm-roberta-base` (Wolf et al., 2020) for our architecture, which has a total of 125 million parameters.[11] While it may be possible to achieve a higher performance with a monolingual English-only model, we believe that this would not accurately reflect the potential performance on other languages, because high-quality transformer models are not available for all languages. Furthermore, our choice is in line with the current trend in the WMT Shared Task on Quality Estimation, where XLM-R is one of the most commonly used transformer architectures (Specia et al., 2021; Zerva et al., 2022). All the experiments used a learning rate of $2e-5$ and employed the AdamW optimiser. We explored an effective training batch size $\in [2, 16, 32]$ and epochs $\in [1, 5, 10]$, as suggested for XLM-R by a recent study on the performance of multilingual language models by Hu et al. (2020).[12]

Additionally, for our use case, we used HuberLoss as the loss function (Huber, 1992).[13] This loss combines the advantages of both the MSELoss and the L1Loss because it employs a squared term if the absolute element-wise error falls below a predefined δ and a δ-scaled L1Loss otherwise (we use the default value for δ), making HuberLoss less sensitive to outliers.

We use Root Mean Square Error (RMSE) for the evaluation (lower values correspond to a better performance). Since it is scale-dependent, and the distributions of the labels fall within different ranges, the RMSE is not comparable across tasks. This makes it only informative with respect to the original distribution. In order to obtain a value which is not only comparable but also easily interpretable across tasks, all model predictions and gold labels are reshaped into the range $[0, 1]$. We also compute both Pearson's and Spearman's correlation coefficients (Cohen et al., 2009; Spearman, 1987) between the predicted outputs and the original predictions, similarly to what is done in the ranking of WMT tasks, except that we use MTE/QE scores as reference values instead of human evaluations (Zerva et al., 2022).

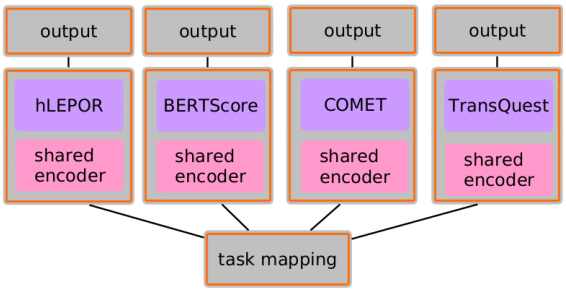

Figure 2: Representation of the multi-task model. Each box represents a separate encoder with a different prediction head, one for every MTE and QE score, each one connected via an external task mapping module.

Table 2 shows the RMSE results for both the single-task and multi-task XLM-R model, trained on a batch size of 2. The multi-task model performs poorly on all tasks except for BERTScore, for which it shows significant improvements over the single-task model, which instead converges to the mean value (0.7229). All models show an increased performance at smaller epochs, suggesting that with such a small batch size the models are likely overfitting. The only exception appears to be COMET, whose best model can actually be found at 5 epochs. Overall, though, the performance is generally poor, which is also confirmed by the extremely low values of Pearson's R and Spearman's ρ, which all approach 0, except for the single-task model (see Table 3).

Table 2 also shows the results for the single-task XLM-R models using the same learning rate as before but exploring a batch size of 16 and 32, respectively. Scaling to higher batch sizes yields better performance, as attested by the overall smaller RMSE values. All models show significant signs of learning as early as the first epoch, ramping up but remaining very close with respect to the RMSE value from 5 to 10 epochs. These results are confirmed by the correlation values, which are significantly higher for all tasks, showing definite correlation with values as high as 0.688 for TransQuest. This is especially evident at 5 epochs, where the overall strongest correlation is found (see Table 3).

[11] `https://huggingface.co/xlm-roberta-base`
[12] `https://github.com/JunjieHu/xtreme-dev/issues/2`
[13] `https://pytorch.org/docs/stable/generated/torch.nn.HuberLoss.html`.

	2b@1*	2b@5*	2b@10*	2b@1	2b@5	2b@10
hLEPOR	0.4006	0.3800	0.4611	0.1361	0.1498	0.1601
BERTScore	0.2676	0.3063	0.3075	0.3500	0.6030	0.4215
COMET	0.3910	0.2439	0.3354	0.2972	0.1461	0.2248
TransQuest	0.3019	0.2035	0.2281	0.2010	0.2212	0.2127
	16b@1	16b@5	16b@10	32b@1	32b@5	32b@10
hLEPOR	0.1342	0.1292	0.1387	0.1456	**0.1260**	0.1386
BERTScore	0.3359	0.1931	**0.1747**	0.3381	0.2069	0.1833
COMET	0.2731	0.1161	0.1419	0.1598	0.1309	**0.1126**
TransQuest	0.1493	0.1339	0.2116	0.1543	0.1569	**0.1338**

Table 2: Results using a training batch size of 16 and 32 at different epochs $[1, 5, 10]$, only using single-task models. The score is reported as normalized RMSE value and the best performances are highlighted in bold.

	hLEPOR	BERTScore	COMET	TransQuest
e=1				
multi	-0.017	-0.014	0.019	0.008
2b	0.546	0.357	0.395	0.549
e=5				
16b	0.565	0.415	0.475	**0.688**
32b	**0.589**	**0.420**	0.444	0.660
e=10				
16b	0.521	0.412	**0.477**	0.596
32b	0.519	0.381	0.446	0.686

Table 3: Correlation values between the predictions of the most accurate models and the original evaluation metrics. The score is calculated using Pearson's R. The best result on each metric is in **bold**.

	hLEPOR	BERTScore	COMET	TransQuest
e=1				
multi	-0.033	-0.007	0.023	0.009
2b	0.335	0.340	0.464	0.434
e=5				
16b	0.358	0.404	0.503	**0.652**
32b	0.352	**0.416**	0.487	0.629
e=10				
16b	0.374	0.402	**0.516**	0.546
32b	**0.379**	0.378	0.515	0.643

Table 4: Correlation values between the predictions of the most accurate models and the original evaluation metrics. The score is calculated using Spearman's ρ. The best result on each metric is in **bold**.

5 Discussion

The obtained results are promising. Given that, on average, the reported RMSE values of the best models lie in the $[0.11, 0.17]$ range, whereas their correlation scores are in the $[0.420, 0.688]$ range for Pearson's R and in the $[0.379, 0.652]$ range for Spearman's ρ. This means that all single-task models are able to reproduce the MTE/QE fairly accurately starting from the source text alone, which corroborates our RQ.

Overall, the best performing batch size for the single-task model is 32, also thanks to its reduced training time, even though it is certainly more costly in terms of memory requirements.

Especially encouraging are the Pearson's correlation scores. Not only do they confirm the results obtained using the RMSE values, but they are also in line with the latest results of the WMT shared task in Quality Estimation for the English–German language pair, where the top-performing IST-Unbabel submission to the segment-level evaluation track has obtained a correlation score of 0.559 (Rei et al., 2022; Zerva et al., 2022). It is also interesting to note the higher correlation achieved by our model with QE scores in comparison to MTE scores, a division clearly visible in Tables 3 and 4. Given that in our case the model is completely blind to the target sentences, these results coult be connected to the findings of Sun et al. (2020), who show that QE metrics tend to assign higher scores to fluent translations or source segments with low complexity, regardless of their semantic similarity to the original source sentence. These correlations should be further investigated to better understand what are the implications for QE models with respect to the source text.

Considering all of the above, we conclude that the RQ is corroborated by the results obtained by the single-task model, meaning that it is possible to accurately predict evaluation scores from the source text alone.

With regards to which approach is better suited to the problem, the answer is indeed more challenging. Although the single-task model appears to be

decidedly better than the multi-task model in 3 out of 4 target scores, there certainly is room for improvement for the multi-task model, given that it never showed a tendency to converge to the mean, contrary to the single-task model, and especially on BERTScore, the knowledge transfer obtained by training on multiple metrics seemed to be beneficial. The results for all other metrics are overall stable, showing no noticeable sign of improvement past the 5-epoch margin (see Table 2). As stated in the previous section, this might be a sign of overfitting which, based on the current results and their stability, might be solved by scaling to bigger batch sizes, meaning the model could indeed experience an increased benefit from seeing multiple segments at once. In this regard, researching higher batch sizes would thus be the natural follow-up step to the current study.

The low error margins and the good correlation values shown in these experiments point towards the possibility to achieve an accurate estimate of the quality of MT based on the source text alone, without needing to even obtain a machine translated version of the given segment. Additionally, given that these automatic metrics are not perfect themselves, future research should focus on testing this model on either Human DA provided by WMT (Zerva et al., 2022), similarly to Don-Yehiya et al. (2022), or by assessing post-editing effort based on the scores produced, working towards the definition of thresholds to generate an actual implementation of the workflow sketched in Figure 1.

Nevertheless, it is also imperative to stress two limitations of this study. The corpus which was used in this study contains segments which have likely been seen by the MT system already during training. Although a set of exploratory experiments has shown no significant difference between unseen and seen texts, this remains an aspect that requires further attention, since it would be possible to argue that to properly learn how difficult a text was for a given system, this had to never be seen by the system during training in the first place.

We also need to consider the issue of sustainability. In recent years the carbon footprint of large language models has become increasingly impactful and longer training times have been disincentivised by the research community (Anthony et al., 2020; Bannour et al., 2021). The multi-task model used for this study took around 32 hours to train, much longer than the single-task model, which took a fifth of the time, further decreased to 2:30 hours when scaling to higher batch sizes. Additionally, since it needs to load four distinct copies of the same XLM-R model, the total number of parameters used increases from 125 million to 500 million in training. This led to the experiments for the multi-task model to be only carried out on a batch size of 2 and, given the significant improvements obtained by the single-task model both in training time and performance, a greater batch size could therefore not only improve the performance of the multi-task model but also reduce its carbon footprint.

6 Conclusions

This work attempted to answer one main research question: is it possible to accurately predict the Machine Translation Evaluation or Quality Estimation score from the source text alone? It was motivated by the increasing need to automatically assess the quality of machine translation in a way that is both dynamic and scalable, without the limitation of providing very expensive reference translations.

While there exists a field entirely dedicated to reference-less metrics, namely Quality Estimation, this paper tried to explore innovative techniques that would focus entirely on the source text. Such an approach offers an alternative that could further reduce the costs of machine translation by streamlining the post-editing process without the need to first generate every time the machine-translated version of all the segments, given that many will be inevitably discarded, which constitues a net loss. In fact, it might even be beneficial to avoid having these low-quality segments undergo post-editing, since recent studies have pointed towards lower lexical variation of post-edited machine translation segments, leading to an overall lower quality of the resulting translation (Volkart and Bouillon, 2022).

Additionally, post-editing also leads to several challenges in liaising with the post-editors themselves, especially with respect to what constitutes an error in a given scenario and how to provide quality assurance, leading to increased costs (Nunziatini and Marg, 2020). In order to streamline these processes, reducing costs and improving efficiency, our proposed model can be integrated as part of a workflow which includes a Machine Translation Suitability module to reroute a source text to MT, PE or human translation (HT) depending on the assessed level of suitability (See Figure 1).

The scripts and corpora used for the experiments

are available for research purposes.[14] While further studies involving human evaluation are still needed, by obtaining an RMSE score as low as 0.11 and good correlations of up to 0.688 with MTE/QE metrics, we show a possible link between MT quality prediction and the source text. We also show that, while the multi-task model might be well-suited for this task, its performance is subpar when compared to the single-task model and there remain concerns regarding its computational cost and sustainability issues. Nevertheless, the results point toward the possibility of obtaining accurate machine translation evaluations starting from the source text alone, paving the way for further research in the field of MT Suitability.

Future research could improve many aspects touched by this work. Exploring correlations with Human DA scores, research on source text translatability for humans or assessing post-editing effort based on the scores produced are all paramount aspects to investigate in order to correctly define the thresholds for the workflow proposed in Figure 1. Moreover, since XLM-R is a multilingual model, an additional focus could be posed on extending the experiments to other language pairs, surveying significant differences among different language combinations and directions to further confirm the current findings. Especially interesting would be to expand the analysis on the higher correlation between our metric and the QE metrics when compared to MTE metrics, because it may shed further light on what state-of-the-art QE models are actually predicting. Additionally, adding a pipeline for terminology recognition in the source text could offer valuable information for the final prediction, given how terminology is still a problematic aspect for many MT systems (Dinu et al., 2019). Lastly, two main aspects could be improved in order to surpass the current limitations: the training corpus and the training methodology, especially by scaling the current architecture to greater batch sizes.

Acknowledgements

We wish to thank Francisco Guzmán (Facebook Research) for the early discussions on translation suitability, Alex Yanishevsky (Smartling) for our conversations on the impact of the source text on machine translation and post-editing, and the reviewers for their invaluable comments.

[14] https://github.com/TinfFoil/MTsweet

References

Anthony, Lasse F Wolff, Benjamin Kanding, and Raghavendra Selvan. 2020. Carbontracker: Tracking and predicting the carbon footprint of training deep learning models. *arXiv preprint arXiv:2007.03051*.

Bannour, Nesrine, Sahar Ghannay, Aurélie Névéol, and Anne-Laure Ligozat. 2021. Evaluating the carbon footprint of NLP methods: a survey and analysis of existing tools. In *Proceedings of the Second Workshop on Simple and Efficient Natural Language Processing*, pages 11–21, Virtual, November. Association for Computational Linguistics.

Bertoldi, Nicola, Davide Caroselli, and Marcello Federico. 2018. The ModernMT Project. *Proceedings of the 21st Annual Conference of the European Association for Machine Translation*, page 345.

Cambra, Jon and Mara Nunziatini. 2022. All you need is source! a study on source-based quality estimation for neural machine translation. In *Proceedings of the 15th Biennial Conference of the Association for Machine Translation in the Americas (Volume 2: Users and Providers Track and Government Track)*, pages 210–220, Orlando, USA, September. Association for Machine Translation in the Americas.

Caruana, Rich. 1997. Multitask learning. *Machine learning*, 28(1):41–75.

Cohen, Israel, Yiteng Huang, Jingdong Chen, Jacob Benesty, Jacob Benesty, Jingdong Chen, Yiteng Huang, and Israel Cohen. 2009. Pearson correlation coefficient. *Noise reduction in speech processing*, pages 1–4.

Conneau, Alexis, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online, July. Association for Computational Linguistics.

Dinu, Georgiana, Prashant Mathur, Marcello Federico, and Yaser Al-Onaizan. 2019. Training neural machine translation to apply terminology constraints. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3063–3068, Florence, Italy, July. Association for Computational Linguistics.

Don-Yehiya, Shachar, Leshem Choshen, and Omri Abend. 2022. PreQuEL: Quality estimation of machine translation outputs in advance. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11170–11183, Abu Dhabi, United Arab Emirates, December. Association for Computational Linguistics.

European Language Industry Survey Research. 2022. European Language Industry Survey 2022. Technical report, March.

Freitag, Markus, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, George Foster, Alon Lavie, and Ondřej Bojar. 2021. Results of the WMT21 metrics shared task: Evaluating metrics with expert-based human evaluations on TED and news domain. In *Proceedings of the Sixth Conference on Machine Translation*, pages 733–774, Online, November. Association for Computational Linguistics.

Gunning, Robert. 1969. The Fog Index After Twenty Years. *Journal of Business Communication*, 6(2):3–13, January.

Han, Lifeng, Irina Sorokina, Gleb Erofeev, and Serge Gladkoff. 2021. cushLEPOR: customising hLEPOR metric using optuna for higher agreement with human judgments or pre-trained language model LaBSE. In *Proceedings of the Sixth Conference on Machine Translation*, pages 1014–1023, Online, November. Association for Computational Linguistics.

Hu, Junjie, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *International Conference on Machine Learning*, pages 4411–4421. PMLR.

Huber, Peter J. 1992. Robust estimation of a location parameter. In *Breakthroughs in statistics*, pages 492–518. Springer.

Junczys-Dowmunt, Marcin, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia, July. Association for Computational Linguistics.

Karimi Mahabadi, Rabeeh, Sebastian Ruder, Mostafa Dehghani, and James Henderson. 2021. Parameter-efficient Multi-task Fine-tuning for Transformers via Shared Hypernetworks. 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 565–576, Online, August. Association for Computational Linguistics.

Kincaid, J Peter, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. 1975. Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel. Technical report, Naval Technical Training Command Millington TN Research Branch.

Kocmi, Tom, 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, Online, November. Association for Computational Linguistics.

Koehn, Philipp, Vishrav Chaudhary, Ahmed El-Kishky, Naman Goyal, Peng-Jen Chen, and Francisco Guzmán. 2020. Findings of the WMT 2020 shared task on parallel corpus filtering and alignment. In *Proceedings of the Fifth Conference on Machine Translation*, pages 726–742, Online, November. Association for Computational Linguistics.

Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *CoRR*, abs/1907.11692.

Mann, Henry B and Donald R Whitney. 1947. On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*, pages 50–60.

Marie, Benjamin, Atsushi Fujita, and Raphael Rubino. 2021. Scientific credibility of machine translation research: A meta-evaluation of 769 papers. 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 7297–7306, Online, August. Association for Computational Linguistics.

Nunziatini, Mara and Lena Marg. 2020. Machine translation post-editing levels: Breaking away from the tradition and delivering a tailored service. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 309–318, Lisboa, Portugal, November. European Association for Machine Translation.

Papineni, Kishore, 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, Philadelphia, Pennsylvania, USA, July. Association for Computational Linguistics.

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

Ranasinghe, Tharindu, Constantin Orasan, and Ruslan Mitkov. 2020a. TransQuest at WMT2020: Sentence-Level Direct Assessment. In *Proceedings of the Fifth Conference on Machine Translation*, pages 1049–1055, Online, November. Association for Computational Linguistics.

Ranasinghe, Tharindu, Constantin Orasan, and Ruslan Mitkov. 2020b. TransQuest: Translation Quality Estimation with Cross-lingual Transformers. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5070–5081, Barcelona,

Spain (Online), December. International Committee on Computational Linguistics.

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

Rei, Ricardo, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. 2022. CometKiwi: IST-Unbabel 2022 Submission for the Quality Estimation Shared Task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid), December. Association for Computational Linguistics.

Song, Kaitao, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. Mpnet: Masked and permuted pre-training for language understanding. *Advances in Neural Information Processing Systems*, 33:16857–16867.

Spearman, Charles. 1987. The proof and measurement of association between two things. *The American journal of psychology*, 100(3/4):441–471.

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

Sun, Shuo, 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, Online, July. Association for Computational Linguistics.

Tezcan, Arda. 2022. Integrating fuzzy matches into sentence-level quality estimation for neural machine translation. *Computational Linguistics in the Netherlands Journal*, 12:99–123.

Tiedemann, Jörg. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey, May. European Language Resources Association (ELRA).

Vanroy, Bram, Orphée De Clercq, and Lieve Macken. 2019. Correlating process and product data to get an insight into translation difficulty. *Perspectives*, 27(6):924–941.

Vasileva, Slavena, Pepa Atanasova, Lluís Màrquez, Alberto Barrón-Cedeño, and Preslav Nakov. 2019. It Takes Nine to Smell a Rat: Neural Multi-Task Learning for Check-Worthiness Prediction. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 1229–1239, Varna, Bulgaria, September. INCOMA Ltd.

Vijayaraghavan, Prashanth, Soroush Vosoughi, and Deb Roy. 2017. Twitter demographic classification using deep multi-modal multi-task learning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 478–483, Vancouver, Canada, July. Association for Computational Linguistics.

Volkart, Lise and Pierrette Bouillon. 2022. Studying post-editese in a professional context: A pilot study. In *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, pages 71–79, Ghent, Belgium, June. European Association for Machine Translation.

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

Yanishevsky, Alex. 2021. Bad to the bone: Predicting the impact of source on MT. In *Proceedings of Machine Translation Summit XVIII: Users and Providers Track*, pages 175–199, Virtual, August. Association for Machine Translation in the Americas.

Zerva, Chrysoula, Frédéric Blain, Ricardo Rei, Piyawat Lertvittayakumjorn, José G. C. de Souza, Steffen Eger, Diptesh Kanojia, Duarte Alves, Constantin Orăsan, Marina Fomicheva, André F. T. Martins, and Lucia Specia. 2022. Findings of the WMT 2022 Shared Task on Quality Estimation. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 69–99, Abu Dhabi, United Arab Emirates (Hybrid), December. Association for Computational Linguistics.

Zhang, Tianyi, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. In *International Conference on Learning Representations*.

A Mann–Whitney U Test

In order to perform the Mann–Whitney U Test, we have selected three recently published texts available on the Globalvoices website[15] in both English and German, which have been manually extracted

[15] https://globalvoices.org/about/

and segmented. This website was selected because one of the subcorpora from OPUS, the News subcorpus, contains some texts from Globalvoices (Tiedemann, 2012).[16] The Mann-Whitney U test assesses whether two independent populations belong to the same distribution. In order to perform the test, four assumptions are needed: (1) the dependent variable should be measured at the ordinal or continuous level (evaluation metrics are continuous), (2) the independent variable should consist of two categorical, independent groups (i.e., the corpus with "seen" texts and the corpus with "unseen" texts), (3) there is independence of observations (there is no inherent relationship among the various segments), and (4) the two variables are not normally distributed.

	hLEPOR	BERTScore	COMET	TransQuest
glob	0.8555	0.6642	0.5651	0.7346
std	0.1548	0.1841	0.4232	0.0155
med	0.8875	0.6720	0.6941	0.7368
min	0.0	0.0	-2.4113	0.6548
max	1.0	1.0	1.3308	0.7759

Table 5: ModernMT corpus scores distribution

	U	p-value
hLEPOR	749851.0	0.0713
BERTScore	808062.0	0.4736
COMET	764728.0	0.1338
TransQuest	670510.5	0.0004

Table 6: Mann-Whitney U Test results for the comparison among the ModernMT and Globalvoices dataset distributions

[16]We do not use this corpus as a test set, because it is restricted to the "news" domain and only contains 128 TUs.

	hLEPOR	BERTScore	COMET	TransQuest
train	0.888	0.672	0.694	0.737
glob	0.879	0.671	0.714	0.740

Table 7: Median values for comparison between the training dataset and the Globalvoices dataset.

Our data adheres to these assumptions, as observed in Table 5. Tables 6 and 7 show the results of the test. There is no significant difference ($p>0.05$) between the training and the Globalvoices dataset for all metrics except for TransQuest. This gives confidence that both corpora belong to the same non-gaussian distribution, meaning we can safely proceed with assuming there is no difference in the quality scores obtained by texts translated using our training corpus and a corpus containing texts not seen by the MT system.

B Multi-task Setting Details

We test the multi-task architecture using the same settings as the single-label one, with the major difference being the effective training batch size. In order to generate the multi-task model, it is necessary to load four copies of the same language model simultaneously on the GPU. As a result, the total parameters see an increase from 125 million to 500 million. This led us to only test the multi-task model with an effective training batch size of 2 due to its significant computational cost. All experiments were carried out using an NVIDIA Quadro P4000 8 GB GPU; the training lasted 6 hours for each single-task model and 32 hours for the multi-task model.