Rethinking Round-trip Translation for Automatic Machine Translation Evaluation

Terry Yue Zhuo1 and Qiongkai Xu2* and Xuanli He3 and Trevor Cohn2
1 Monash University, Clayton, VIC, Australia
2 The University of Melbourne, Carlton, VIC, Australia
3 University College London, London, United Kingdom
terry.zhuo@monash.edu
qiongkai.xu,trevor.cohn@unimelb.edu.au
zodiac.he@gmail.com

Abstract

A parallel corpus is generally required to automatically evaluate the translation quality using the metrics, such as BLEU, METEOR and BERTScore. While the reference-based evaluation paradigm is widely used in many machine translation tasks, it is difficult to be applied to translation with low-resource languages, as those languages suffer from a deficiency of corpora. Round-trip translation provides an encouraging way to alleviate the urgent requirement of the parallel corpus, although it was unfortunately not observed to correlate with forwarding translation in the era of statistical machine translation. In this paper, we firstly observe that forward translation quality consistently correlates to corresponding round-trip translation quality in the scope of neural machine translation. Then, we carefully analyse and unveil the reason for the contradictory results on statistical machine translation systems. Secondly, we propose a simple yet effective regression method to predict the performance of forward translation scores based on round-trip translation scores for various language pairs, including those between very low-resource languages. We conduct extensive experiments to show the effectiveness and robustness of the predictive models on 1,000+ language pairs. Finally, we test our method on challenging settings, such as predicting scores: i) for unseen language pairs in training and ii) on real-world WMT shared tasks but in new domains. The extensive experiments demonstrate the robustness and utility of our approach. We believe our work will inspire works on very low-resource multilingual machine translation.

1 Introduction

Thanks to the recent progress of neural machine translation (NMT) and large-scale multilingual corpora, machine translation (MT) systems have achieved remarkable performances on high- to medium-resource languages (Goyal et al., 2022a; Fan et al., 2021; Pan et al., 2021). However, the development of MT technology on low-resource language pairs still suffers from insufficient data for training and evaluation (Aji et al., 2022; Siddhant et al., 2022).

Traditional automatic MT evaluation metrics (Papineni et al., 2002; Snover et al., 2006) require parallel corpora to provide ground-truth references for estimating the translation quality. However, these approaches are impractical for evaluating under-represented languages without high-quality parallel corpora. Other work attempts to tackle this issue via reference-free MT evaluation, mainly focusing on quality estimation (QE). Most still require a large-scale training dataset with intensive direct-assessed or post-edited human annotation at the sentence level (Rei et al., 2020b; Martins et al., 2017). Such practice is especially resource-consuming when it is adapted to i) hundreds of underrepresented low-resource (Aji et al., 2022; Joshi et al., 2019; Bird and Chiang, 2012) and uncovered language pairs in human evaluation benchmarks (Specia et al., 2021; Freitag et al., 2021); and ii) multiple specific application domains for translation (Li et al., 2020; Müller et al., 2020).

In order to mitigate the aforementioned issues, we explore the feasibility of evaluating translation quality without a parallel corpus for references. We propose to utilise the round-trip translation (RTT) results to predict the forward translation (FT) performance and adapt the prediction to unseen language pairs or usage scenarios. The automatic evaluation of (forward) Translation Score (TRANS-Score) using RTT is demonstrated in Figure 1. Given a sentence A, TRANS-Score is calculated based on its (aligned) gold reference B and the system prediction $B'$. However, when B is absent, or the alignment is missing, shown as a dashed line in Figure 1.b, we can only access the RTT re-
results, e.g., a sentence $A$ is translated into $B^*$, then $B^*$ is translated back into $A'$. A self-evaluation score ($\text{SELF-SCORE}$) compares the original sentence $A$ with its RTT result $A'$. We formulate the new challenge as predicting the translation performance $\text{TRANS-Score}$ merely using $\text{SELF-SCORE}$ from RTT.\footnote{Note that our work could be applied to the setting of non-English-centric low-resource (LR) language pairs, \emph{i.e.}, LR-LR, which is less discussed than English-centric low-resource translation, \emph{i.e.}, EN-LR or LR-EN.}

In our preliminary experiments, we observe strong correlations between $\text{TRANS-Scores}$ and $\text{SELF-Scores}$ on various MT systems. While early studies (Huang, 1990; Zaanen and Zwarts, 2006; Koehn, 2005; Somers, 2005) on statistical machine translation (SMT) suggest that RTT is inappropriate for estimating the translation quality in the automatic evaluation settings, recent works (Moon et al., 2020; Crone et al., 2021) show the effectiveness of RTT to rank NMT systems, and hence can be used as a metric of quality estimation. We design comprehensive experiments to analyse and demonstrate the factors to such contradictory. The findings set the basis of using regression models to predict $\text{TRANS-Score}$ via $\text{SELF-Score}$. We construct an experimental basis for the new task, including a multilingual corpus \textsc{Flores-}AE\textsc{33} and corresponding translation results by several state-of-the-art multilingual MT systems. We train regression models on some language pairs and verify their ability to predict $\text{TRANS-Score}$. Because no training instances in target language pairs are required for inference, we consider our setting zero-shot learning. Moreover, we challenge the predictors under the settings of new MT systems, unseen language pairs during training, and transferred usage domains. The experimental results show our models are robust to these challenging settings. Finally, we demonstrate that our models can appropriately rank the competing MT systems on a series of real-world \textsc{Wmt} machine translation shared tasks.

Our main contributions are:

\begin{itemize}
  \item We propose a new task, zero-shot automatic machine translation evaluation, which does not rely on a parallel corpus on the tested language pairs.
  \item We observe strong correlation between $\text{TRANS-Score}$ and $\text{SELF-Score}$, and clarify the factor of its contradictory results on SMT.
  \item We propose using RTT to generate features and regression models to predict $\text{TRANS-Scores}$. Then, we conduct extensive experiments to show the robustness and usability of our approach. Our experiments show the regression models could predict $\text{TRANS-Scores}$ robustly on unseen language pairs and new MT systems. The predicted $\text{TRANS-Scores}$ could also appropriately rank MT systems for \textsc{Wmt} shared tasks in new domains.
\end{itemize}
The dataset and code will be available online. 2

2 Related Work

Reference-based Machine Translation Evaluation Metric. Designing high-quality automatic evaluation metric for translation is one of the fundamental challenges in MT research. Most existing metrics largely rely on parallel corpora to provide aligned texts as references (Papineni et al., 2002; Lin, 2004). One can compare translated outputs against references to estimate the performance of MT systems. The string-based metrics incorporate lexical matching rate for translation quality, such as BLEU (Papineni et al., 2002), ChrF (Popović, 2015) and TER (Snover et al., 2006). In addition, metrics using pre-trained language models to estimate the semantic relevance of texts, such as BERTScore (Zhang et al., 2020b) and BLEURT (Sellam et al., 2020), are demonstrated matching human evaluation (Kocmi et al., 2021). In the meantime, some reference-based evaluation metrics require supervised training to work well (Rei et al., 2020a; Mathur et al., 2019). While these automatic evaluation metrics are widely applied in MT evaluation, they fail in the low-resource language translation scenarios where there are no ground-truth parallel references (Mathur et al., 2020). Our work relaxes the constraint of the evaluating corpus to non-parallel or even monolingual.

Reference-free Quality Estimation. In recent years, there has been a surge of interest in designing QE metrics, which aims to predict translation quality from human expert judgement without the access to parallel reference translations in the run-time (Specia et al., 2010, 2013; Bojar et al., 2014; Zhao et al., 2020). Recent focus on QE is mainly based on human evaluation approaches, direct assessment (DA) and post-editing (PE), where researchers intend to train models on numerous human evaluation score features to estimate MT quality. Despite few unsuccessful early QE works towards predicting automatic evaluation metric (Blatz et al., 2004), current QE metrics generally require human-annotated DA and PE data at sentence level for training on the target languages pairs. YiSi-2 (Lo, 2019) and COMET-QE-MQM (Rei et al., 2021) are recent progress, which demonstrate their effectiveness on WMT shared tasks. Our work follows a zero-shot setting for low-resource translation quality evaluation, which means there is no need of data in the tested language pairs for training our predictors.

Quality Estimation via Round-trip Translation

Round-trip translation has been widely used in data augmentation in various NLP tasks (Edunov et al., 2018; Hoang et al., 2018; Edunov et al., 2020). However, there is no consistent agreement if it could be applied to indicate the translation quality. In order to assess the quality, there are two types of evaluation paradigms, automatic evaluation and human judgement. As the attempts on QE via RTT lie in the area of automatic evaluation, metrics such as BLEU and BERTScore are frequently considered in the experiments. Somers (2005) firstly concluded that there was no relationship between RTT and forward translation quality by ranking several online statistical MT systems on two language pairs against RTT BLEU scores. Later, Koehn (2005) confirmed previous findings with the experiments testing one MT system for 10 English-centric language pairs. Our work rectify the transferred misunderstanding from SMT to NMT.

3 Methodology

3.1 Problem Statement

Given machine translation systems, \( T_{A \rightarrow B} \) and \( T_{B \rightarrow A} \), between two languages (\( L_A \) and \( L_B \)), we evaluate the systems using datasets in these two languages, \( D_A = \{ a_i \}_{i=1}^N \) and \( D_B = \{ b_i \}_{i=1}^M \). \( D_A \) and \( D_B \) could be collected from different domains. The corresponding MT systems’ translation performance scores \( \text{TRANS-SCORE}_{A \rightarrow B} \) and \( \text{TRANS-SCORE}_{B \rightarrow A} \) are evaluated by:

\[
f(T_{A \rightarrow B}, T_{B \rightarrow A}, D_A, D_B). \tag{1}
\]

We consider the conventional MT evaluation paradigm with a parallel corpus, a special case of our problem, when \( D_B \) is aligned with \( D_A \) and they have the same number of samples \( N = M \). The parallel corpus is noted as \( D_{A||B} = \{ (a_i, b_i) \}_{i=1}^N \).

3.2 Challenges and Solutions

An example of traditional MT evaluation method with regard to an evaluation metric \( M \) for transla-
Another example is that translation systems generate BLEU
intensive. Hence, we investigate another setting (or task, collecting a corpus in low-resource language domain. When organising a new WMT and corpora generally have different abilities in different domains, we are aware of some real-world cases that limit us from accessing the corpora from both languages. For example, Matukar Panau (an Austronesian language) (Barth et al., 2019) and Warlpiri (an aboriginal language in Australia) (Bavin, 1992) have small groups of speakers and very limited digital corpora. Another example is that translation systems generally have different abilities in different domains, and corpora A and B ideally should be in the same domain. When organising a new WMT-style shared task, collecting a corpus in low-resource language B (or A), especially in the same domain as corpus A (or B), could be challenging and resource-intensive. Hence, we investigate another setting that utilises merely the monolingual corpora in language A or B to predict Trans-Score,

\[
\text{TRANS-SCORE}_{A\rightarrow B}^M \approx f_M(\text{SELF-SCORE}_{A\rightarrow B}^M),
\]

\[
\text{TRANS-SCORE}_{A\rightarrow B} \approx f_M(\text{SELF-SCORE}_{A\rightarrow B}^M). \tag{5}
\]

We will compare and discuss this setting in our experiments on WMT.

3.3 Regression Model

We introduce our linear regression model for predicting Trans-Score,

\[
f_M(S) = W_1 \cdot S_{A\rightarrow B}^M + W_2 \cdot S_{B\rightarrow A}^M + \beta \tag{6}
\]

where \(S_{A\rightarrow B}^M\) and \(S_{B\rightarrow A}^M\) are Self-Score features used as inputs of the regression model\(^3\). \(W_1, W_2\) and \(\beta\) are the parameters of the prediction model optimised by supervised training. We use Residual Sum of Squares (RSS) loss to train the predictor,

\[
\mathcal{L}_{RSS} = \sum_{i=1}^{K} (t_i^M - f_M(S_i))^2, \tag{7}
\]

where \(K\) is the total number of training samples and \(t_i^M\) is the target Trans-Score of the \(i\)-th sample according to metric \(M\).

4 Preliminary Study on Statistical Machine Translation

4.1 Experimental Setup

We simulate the experiments by Somers (2005) and Koehn (2005) by using datasets and SMT architecture at the time.

Datasets. We train our SMT systems on News-Commentary v8 (Tiedemann, 2012), as suggested by WMT 2008 organisers. We test our systems on the competition tracks in WMT 2008 Translation Shared Tasks (Callison-Burch et al., 2008), namely news track WMT2008-News, which covers more diverse languages and domains than the ones in 2006 and 2007.

Machine Translation Systems. We utilise Moses (Koehn and Hoang, 2009) to train phrase-based MT systems (Koehn et al., 2003), which were popular back to those years. We follow the

\(^3\)We use \(M^* = M\) as our primary setting, as it is the most straightforward and effective method to construct features. In addition, we discuss the possibility of improve the regression model by involving more features, see RQ 5 in Section 6.
suggested baseline setup in a Moses tutorial.\footnote{http://www2.statmt.org/moses/?n=Moses.Baseline} Kendall’s $\tau$ is used to verify whether the scores could correctly rank the MT systems coherently.

4.2 Experiments and Analysis

We study a few factors which potentially affect the quality of RTT and FT.

| Lang. Pair | $\tau$ w/o cp | $\tau$ w/ cp | $\Delta$ |
|------------|----------------|--------------|---------|
| de-en      | 1.00           | -1.00        | -2.00   |
| en-de      | 0.60           | -1.00        | -1.40   |
| fr-en      | 0.60           | -1.00        | -1.40   |
| en-fr      | 1.00           | -1.00        | -2.00   |

Table 1: Comparison between systems with copying (w/ cp) and without copying (w/o cp) unknown words using Kendall’s $\tau$ correlation score, on four language pairs. We vary the systems by using different phrase lengths (Phrase Len.) and word probability thresholds (Word Prob.).

![Figure 2: Comparison between SELF-SCORE\textsuperscript{BLEU} (FT) and TRANS-SCORE\textsuperscript{BLEU} (RTT) on four language pairs, based on SMT varying by the phrase probability thresholds.

| Lang. Pair | Avg #. cp | Avg %. cp |
|------------|-----------|-----------|
| SMT        | NMT       | SMT       | NMT       |
| de-en      | 7.00      | 2.20      | 30.96     | 10.33     |
| en-de      | 6.52      | 4.88      | 27.27     | 19.96     |

Table 2: Comparison of word copy phenomenon between SMT and NMT on two language pairs. We calculate the average number of copy (Avg #. cp) and average percentage of copy (Avg %. cp) per sentence.

vary the phrase table probability from 0.1 to 0.5 with the phrase table containing phrases no longer than 4. The comparison between with and without unknown word copy on the SMT systems is shown in Table 1. Our observation is that the quality correlations between RTT and FT have been significantly decrease after applying word copy. The common explanation of negative correlations when applying unknown word copy is that when the size of the phrase table decreases, the MT system has higher chances to copy unknown words and keeps the copies during RTT. The visualisation of SMT systems with default unknown word copy is given in Figure 2.

**Word Copy in SMT and NMT.** Inspired by our previous findings, we further investigate the word copy phenomenon in NMT, compared to SMT. We choose mBART50-m2m as our NMT system. To mine the copied words for each sentence, we tokenize the translation output with the default tokenizer in Moses toolkit. From our observation in Table 2, we find that the copy phenomenon is much more frequent in SMT. Despite the fact that NMT systems may copy words during translation, we observe that most of these copied word are proper nouns, which do not have translations. In contrast, the copied words in SMT are more diversified and tend to be common nouns.

**Takeaways.** The copy mechanism is a significant negative factor to the correlation of SELF-SCORE and TRANS-SCORE. The conflict conclusion between NMT and SMT settings are likely due to the fact that SMT copies more tokens than NMT when some phrases are not recorded in the dictionary.

5 Experimental Setup

5.1 Datasets

We construct our dataset based on two resources, FLORES-101 a multilingual benchmark and WMT machine translation shared tasks.
FLORES-101. FLORES-101 (Goyal et al., 2022a) is a benchmark for low-resource and multilingual MT, including 101 languages. The 101 languages are classified into high-, medium-, low- and very low-resource branches, according to the amount of resources available in OPUS (Zhang et al., 2020a). The benchmark dataset consists of 2,009 (train set + test set) selected sentences from English Wikipedia with various domains and topics, which are translated to the other 100 languages by professional translators. We utilise this corpus mainly for training and validation purposes.

FLORES-AE33. We extract FLORES-AE33, which contains parallel data among 33 languages, covering 1,056 (33×32) language pairs, from a subset of FLORES-101. We select these languages based on two criteria: i) We rank languages given the scale of their bi-text corpora; ii) We prioritise the languages covered by WMT2020-News and WMT2020-Bio. As a result, FLORES-AE33 includes 7 high-resource languages, 16 medium-resource languages and 10 low-resource languages, with more details in Appendix A.

Then, we separate these 33 languages into two sets, i) the languages that are utilised in training our models (TRAIN+TEST) and ii) the others are not utilised for training the predictors but considered for test purpose only (TEST). We include 20 languages to TRAIN+TEST, with 7 high-resource, 7 medium-resource and 6 low-resource. The rest 13 languages fall into TEST, with 9 medium-resource and 4 low-resource. Combining these two categories of languages, we obtain three types of language pairs in FLORES-AE33. Type I contains pairs of languages in TRAIN+TEST, where a train set and a test set are collected and utilised independently. For each language pairs, we collect 997 training samples and 1,012 test samples. The test set of Type II is more challenge than that of Type I set, where the language pairs in this set are composed of one language from TRAIN+TEST set and the other language from TEST set. Type III’s test set is the most challenging one, as all its language pairs are derived from TEST languages. Type II and Type III sets are designed for test purpose and they will not be used for training predictors. Overall, Type I, Type II and Type III sets contain 380, 520, and 156 language pairs, respectively.

WMT. The conference on Machine Translation (WMT) covers various types of MT shared tasks. We collect corpora from the translation track to evaluate multiple MT systems on the same test sets. We consider their ranking based on TRANS-Score with metric $M$ as the ground truth. We choose the competition tracks in WMT 2020 Translation Shared Tasks (Barrault et al., 2020), namely news track WMT2020-News and biomedical track WMT2020-Bio. We consider news and bio new domains, compared to Wikipedia contents in FLORES-101.

5.2 Machine Translation Systems

We experiment with five MT systems which support most of the languages appearing in FLORES-AE33 and WMT. mBART50-M2M (Tang et al., 2020), which covers 50 languages translation, is a many-to-many multilingual MT model using English as the pivot language in training. M2M-100-BASE and M2M-100-LARGE (Fan et al., 2021) are proposed to conduct many-to-many MT without explicit pivot languages, supporting 100 languages.

Google-TRANS (Wu et al., 2016; Bapna et al., 2022) is a commercial translation API, which was considered as a baseline translation systems in many previous competitions (Barrault et al., 2020). Meanwhile, we also include a family of bilingual MT models, OPUS-MT (Tiedemann and Thottingal, 2020), sharing the same model architecture MARIAN-NMT (Junczys-Dowmunt et al., 2018). We provide more details about these MT systems in Appendix B.

5.3 Automatic Evaluation Metrics for Translation

We consider sacreBLEU (Post, 2018), spBLEU (Goyal et al., 2022b), chrF (Popović, 2015) and BERTScore (Zhang et al., 2020b) as the primary automatic evaluation metrics for translation (Freitag et al., 2020). All these metrics will be used and tested for both input features and target TRANS-Score. The first two metrics are differentiated by their tokenizers, where sacreBLEU uses Moses (Koehn and Hoang, 2010) and spBLEU uses SentencePiece (Kudo and Richardson, 2018). Both evaluation metrics were

---

5The train and test sets here refers to dev and devtest sets in FLORES-101.

6Both train and test sets of our corpus will have these languages.

7We queried GOOGLE-TRANS API in March, 2022.
officially used in WMT21 Large-Scale Multilingual Machine Translation Shared Task (Wenzek et al., 2021). While sacreBLEU works for most language tokenizations, spBLEU shows superior effectiveness on various language tokenizations, especially the performance on low-resource languages (Goyal et al., 2022a). For BERTScore, Deberta-xlarge-mnli (He et al., 2021) is used as the backbone pre-trained language model, as it is reported to have decent correlation with human evaluation in WMT16. While the first three metrics are string-based, the last one is model-based. The selection of these metrics is on the basis that they should directly reflect the translation quality. We calculate those scores via open-source toolboxes, EASYNMT\(^8\), SACREBLEU-TOOLKIT\(^9\) and BERTSCORE\(^10\). We use word-level 4-gram for sacreBLEU and spBLEU, character-level 6-gram for chrF, and \(F_1\) score for BERTScore by default.

6 Experiments and Analysis

In this section, we first test our hypothesis whether \(\text{TRANS-Score}\) is positively correlated to \(\text{SELF-Score}\). Then, we evaluate the performance of our proposed model on target translation metrics. Furthermore, we discuss the robustness of our model on unseen language pairs and new MT systems. Finally, we demonstrate the practicability of our approach on ranking MT models on WMT shared tasks in new domains.

RQ1: Is \(\text{TRANS-Score}\) positively correlated to its corresponding \(\text{SELF-Score}\)?

We investigate the correlation between \(\text{TRANS-Score}\) and \(\text{SELF-Score}\) across various settings.

Settings. We experiment with mBART50-M2M and M2M-100-BASE on Type I test set of FLORES-AE33 by comparing their \(\text{SELF-Score}_{M}^{A:B}\), \(\text{SELF-Score}_{M}^{B:A}\) and \(\text{TRANS-Score}_{M}^{A:B}\) using multiple translation metrics \(\text{BLEU}\), \(\text{spBLEU}\), \(\text{chrF}\) and BERTScore. We measure their correlations by computing Pearson’s \(r\) (Benesty et al., 2009) of \(\text{(SELF-Score}_{M}^{A:B}, \text{TRANS-Score}_{M}^{A:B})}\) and \(\text{(SELF-Score}_{M}^{B:A}, \text{TRANS-Score}_{M}^{A:B})}\). Note that our experiment is beyond English-centric, as all languages are permuted and equally considered.

\(^8\)https://github.com/UKPLab/EasyNMT.
\(^9\)https://github.com/mjpost/sacrebleu.
\(^10\)https://github.com/Tiiiger/bert_score.

Results. The overall correlation scores are reported in Table 3. Our results indicate at least moderately positive correlations between all pairs of \(\text{SELF-Scores}\) and \(\text{TRANS-Scores}\). Moreover, we observe that \(\text{TRANS-Score}_{M}^{B:A}\) is generally more correlated to \(\text{TRANS-Score}\) than \(\text{TRANS-Score}_{M}^{A:B}\), leading to strongly positive correlation scores. We attribute the advantage by the fact that \(T_{A\rightarrow B}\) serves as the last translation step in \(\text{TRANS-Score}_{B:A}\). We visualise one correlation experiment of \(\text{TRANS-Score}\) against \(\text{SELF-Score}\) on 380 language pairs in Figure 3. More detailed correlation experiments are illustrated in Appendix E.1.

Takeaways. The preliminary experiment shows the positive correlations between \(\text{TRANS-Scores}\) and \(\text{SELF-Score}\), which motivates us to train regression models to predict \(\text{TRANS-Scores}\).

RQ2: Can we train regression models to predict \(\text{TRANS-Scores}\) of unseen MT systems?

Settings. We train our regression model on Type I train set of FLORES-AE33 based on the translation scores from mBART50-M2M. Then, we test three models, mBART50-M2M, M2M-100-BASE and Google-TRANS, on Type I test set. We vary our experiment on four metrics, \(\text{sacreBLEU}\), \(\text{spBLEU}\), \(\text{chrF}\) and BERTScore. Mean Absolute Error (MAE), Root Mean Square Error (RMSE) and Pearson’s \(r\) are calculated to evaluation the difference between the predicted scores and the original \(\text{TRANS-Scores}\).

Results. In Table 4, we present the performance of our regression models across various translation systems and evaluation metrics. We first analyse
We further verify the language transferability of our regression models by testing them on language pairs with unseen languages in training.

**RQ3: Are the predictors robust to unseen language pairs?**

We continue our experiments using the same models trained in RQ2. In order to evaluate the transfer capability of our models on unseen languages, we consider TRANS-SCOREs of MBART50-M2M (a seen MT system in training) and M2M-100-BASE (an unseen MT system in training) on Type II and Type III test sets in FLORES-AE33. Type II and Type III language pairs respectively include one and two unseen languages for each pair.

**Results and Takeaways.** Our models obtain decent language transferability results on MAE and RMSE, as demonstrated in Table 5. Their Pearson’s $r$ scores are competitive to those reported on Type I in Table 4. The predictors are robust to language transfer in our experiments.

**RQ4: Can we improve performance of the predictors using more features?**

**Settings.** We introduce two extra features, MAX-4 COUNT and REF LENGTH,\(^1\) to enhance the prediction of spBLEU. MAX-4 COUNT is the counts of correct 4 grams, and REF LENGTH is the cumulative reference length. We follow the similar procedure in RQ2, using the same measurements to evaluate the predictor performance on MBART50-M2M and M2M-100-BASE across three types of test sets in FLORES-AE33.

**Results.** Table 6 shows the results of those models with additional features. Both features consistently improve our basic models, and the performance can be further boosted by incorporating both features.

**Takeaways.** Our models could be improved by incorporating simple features. We believe that more carefully designed features and regression models could potentially boost the performance of our predictors.

\(^{1}\)MAX-4 COUNT and REF LENGTH are “counts” and “ref_len” in https://github.com/mjpost/sacrebleu/blob/master/sacrebleu/metrics/bleu.py.
Table 5: The results of predicted Trans-Scores of MBART50-M2M (a seen MT system) and M2M-100-BASE (an unseen MT system) on Type II and Type III (with unseen languages) test sets based on different translation evaluation metrics (Trans. Metric).

| MT System     | Trans. Feature | Type I | Type II | Type III |
|---------------|----------------|--------|---------|----------|
|               |                | MAE ↓  | RMSE ↓  | r ↑     | MAE ↓  | RMSE ↓  | r ↑     |
| MBART50-M2M   | sacreBLEU      | 1.36   | 1.97    | 0.93    | 0.81   | 0.95    | 0.96    |
|               | spBLEU         | 1.61   | 2.19    | 0.93    | 1.20   | 1.38    | 0.94    |
|               | chrF           | 3.80   | 4.89    | 0.95    | 3.04   | 3.89    | 0.95    |
|               | BERTScore      | 5.26   | 7.14    | 0.88    | 5.63   | 7.67    | 0.86    |
| M2M-100-BASE  | sacreBLEU      | 3.10   | 4.16    | 0.95    | 2.99   | 3.76    | 0.94    |
|               | spBLEU         | 3.24   | 4.18    | 0.96    | 3.18   | 3.88    | 0.95    |
|               | chrF           | 5.53   | 6.70    | 0.95    | 5.42   | 6.54    | 0.93    |
|               | BERTScore      | 4.79   | 6.74    | 0.82    | 4.76   | 7.20    | 0.77    |

Table 6: The results of using auxiliary features to spBLEU for training predictors. We test the performance of MBART50-M2M and M2M-100-BASE cross language pairs in Type I, Type II and Type III of FLORES-AE33.

RQ5: Does the predictor manage to rank the translation quality of competing MT systems in shared tasks?

Settings. We conduct more experiments on two series of WMT shared tasks, WMT2020-News and WMT2020-Bio, including 10 and 12 language pairs respectively. For all shared tasks, we have involved five MT systems, MBART50-M2M, M2M-100-BASE, M2M-100-LARGE, OPUS-MT and GOOGLE-TRANS. In addition to MAE, RMSE and Pearson’s r, we introduce Kendall’s τ (Kendall, 1938) to measure the rank correlation coefficient of the MT systems, via comparing the ranking of our model predictions and the actual ranking based on Trans-Score. We are aware of the cases that collecting corpora in target languages for competitions might be significantly complex, which means only a monolingual corpus is available for evaluation. Thus, we train the predictor \( f' \) using single Self-Scores in Equation 5 on Type I train set. Note that this experiment covers several challenging settings, such as transferred MT systems, unseen languages in training, single source features, and transferred application domains.

Results. In Table 7, we display the results on WMT2020-News\(^1\). Although MAE and RMSE vary among experiments for different language pairs, the overall correlation scores are favourable. Pearson’s \( r \) values on all language pairs are above 0.5, showing their strong ranking correlations. The results of the experiments using \( A \odot B \) are competitive to those using both \( A \odot B \) and \( B \odot A \) features, showing the feasibility of predicting Trans-Score using monolingual data.

Takeaways. The regression-based predictors can be practically used to rank MT systems for their translation performance, when merely monolingual data is available. This setting could be adapted to future WMT shared tasks on low-resource MT.

7 Conclusion

This paper investigates the problem of predicting automatic evaluation scores for translation quality without accessing a parallel test corpus. We link Trans-Scores to Self-Scores by discussing their correlation and devise prediction models for Trans-Score. We conduct comprehensive experiments to verify the reliability of the predicted scores and the robustness of the results using new MT systems, unseen languages, transferred usage

\(^1\)We have contacted the competitors to WMT2020-News. However, we have not received enough valid MT systems to increase the number of competitors. We will show the robustness of our method to a larger number of pseudo-competitors in Appendix E.2.

\(^1\)The results on WMT2020-Bio are reported in Appendix E.3.
domains, and real-world shared tasks. We believe our work would potentially motivate the research on low-resource and non-English-centric machine translation by arming them with an appropriate evaluation tool.

Limitations

Although we observe positive correlation between \textsc{Trans-Score} and \textsc{Self-Score}, the relation could be complicated and non-linear. We encourage future research to investigate various \textsc{Self-Score} features and more complex machine learning models for better prediction models. We have not tested those very low-resource languages that are really without parallel corpus. We suggest auditing our model on a small validation dataset. Our research cannot be directly used to train or improve a multilingual MT system on low-resource language, but an evaluation tool will be beneficial to those research.

References

Alham Aji, Genta Indra Winata, Fajri Koto, Samuel Cahyawijaya, Ade Romadhony, Rahmad Mahendra, Kemal Kurniawan, David Moeljadi, Radityo Eko Prasoyo, Timothy Baldwin, et al. 2022. One country, 700+ languages: Nlp challenges for underrepresented languages and dialects in indonesia. In \textit{Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)}, pages 7226–7249.

Ankur Bapna, Isaac Caswell, Julia Kreutzer, Orhan Firat, Daan van Esch, Aditya Siddhant, Mengmeng Niu, Pallavi Baljekar, Xavier Garcia, Wolfgang Macherey, et al. 2022. Building machine translation systems for the next thousand languages. \textit{arXiv preprint arXiv:2205.03983}.

Loïc Barrault, Magdalena Biesalska, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Matthias Huck, Eric Joanis, Tom Kocmi, Philipp Koehn, Chi-kuo Lo, Nikola Ljubešić, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Santanu Pal, Matt Post, and Marcos Zampieri. 2020. Findings of the 2020 conference on machine translation (WMT20). In \textit{Proceedings of the Fifth Conference on Machine Translation}, pages 1–55, Online. Association for Computational Linguistics.

Danielle Barth et al. 2019. Matukar panau language documentation. \textit{PARADISEC (Pacific and Regional Archive for Digital Sources in Endangered Cultures)}.

Edith L Bavin. 1992. The acquisition of warlpiri. In \textit{The crosslinguistic study of language acquisition}, pages 309–371. Psychology Press.

Jacob Benesty, Jingdong Chen, Yiteng Huang, and Israel Cohen. 2009. Pearson correlation coefficient. In \textit{Noise reduction in speech processing}, pages 1–4. Springer.

Steven Bird and David Chiang. 2012. Machine translation for language preservation. In \textit{Proceedings of COLING 2012: Posters}, pages 125–134.

John Blatz, Erin Fitzgerald, George Foster, Simona Gandrabur, Cyril Goutte, Alex Kulesza, Alberto Sanchis, and Nicola Ueffing. 2004. Confidence estimation for machine translation. In \textit{Coling 2004: Proceedings of the 20th international conference on computational linguistics}, pages 315–321.

Ondřej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, et al. 2014. Findings of the 2014 workshop on statistical machine translation. In \textit{Proceedings of the ninth workshop on statistical machine translation}, pages 12–58.

Chris Callison-Burch, Philipp Koehn, Christof Monz, Josh Schroeder, and Cameron Shaw Fordyce. 2008. Proceedings of the third workshop on statistical machine translation. In \textit{Proceedings of the Third Workshop on Statistical Machine Translation}.
Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 489–500.

Sergey Edunov, Myle Ott, Marc’Aurelio Ranzato, and Michael Auli. 2020. On the evaluation of machine translation systems trained with back-translation. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 2836–2846, Online. Association for Computational Linguistics.

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

Markus Freitag, David Grangier, and Isaac Caswell. 2020. Bleu might be guilty but references are not innocent. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pages 61–71.

Markus Freitag, Ricardo Rei, Nitiha Mathur, Chi-kuo 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.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc’Aurelio Ranzato, Francisco Guzman, and Angela Fan. 2022a. The flores-101 evaluation benchmark for low-resource and multilingual machine translation. Transactions of the Association for Computational Linguistics, 10:522–538.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc’Aurelio Ranzato, Francisco Guzman, and Angela Fan. 2022b. The flores-101 evaluation benchmark for low-resource and multilingual machine translation. Transactions of the Association for Computational Linguistics, 10:522–538.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. In International Conference on Learning Representations.

Vu Cong Duy Hoang, Philipp Koehn, Gholamreza Haffari, and Trevor Cohn. 2018. Iterative back-translation for neural machine translation. In Proceedings of the 2nd Workshop on Neural Machine Translation and Generation, pages 18–24.

Xiuming Huang. 1990. A machine translation system for the target language inexpert. In COLING 1990 Volume 3: Papers presented to the 13th International Conference on Computational Linguistics.

Pratik Joshi, Christain Barnes, Sebasntin Santy, Simran Khanauja, Sanket Shah, Anirudh Srinivasan, Satwik Bhattachamrsha, Sunayana Sitaram, Monojit Choudhury, and Kalika Bali. 2019. Unsung challenges of building and deploying language technologies for low resource language communities. In Proceedings of the 16th International Conference on Natural Language Processing, pages 211–219, International Institute of Information Technology, Hyderabad, India. NLP Association of India.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germmn, Alham Bikri Aji, Nikolay Bogyochev, 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. Association for Computational Linguistics.

Maurice G Kendall. 1938. A new measure of rank correlation. Biometrika, 30(1/2):81–93.

Tom Kocmi, Christian Federmann, Roman Grundkiewicz, Marcin Junczys-Dowmunt, Hitokazu Matsuhita, 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.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In Proceedings of machine translation summit x: papers, pages 79–86.

Philipp Koehn and Hieu Hoang. 2009. Moses: Statistical machine translation system.

Philipp Koehn and Hieu Hoang. 2010. Moses. Statistical Machine Translation System, User Manual and Code Guide, page 245.

Philipp Koehn, Franz J Och, and Daniel Marcu. 2003. Statistical phrase-based translation. Technical report, University of Southern California Marina Del Rey Information Sciences Inst.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Rumeng Li, Xin Wang, and Hong Yu. 2020. Metamt, a meta learning method leveraging multiple domain data for low resource machine translation. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 34, pages 8245–8252.
Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In Text summarization branches out, pages 74–81.

Chi-kiu Lo. 2019. Yisi—a unified semantic mt quality evaluation and estimation metric for languages with different levels of available resources. In Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1), pages 507–513.

André FT Martins, Marcin Junczys-Dowmunt, Fabio N Kepler, Ramón Astudillo, Chris Hokamp, and Roman Grundkiewicz. 2017. Pushing the limits of translation quality estimation. Transactions of the Association for Computational Linguistics, 5:205–218.

Nitika Mathur, Timothy Baldwin, and Trevor Cohn. 2019. Putting evaluation in context: Contextual embeddings improve machine translation evaluation. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 2799–2808.

Nitika Mathur, Timothy Baldwin, and Trevor Cohn. 2020. Tangled up in BLEU: Reevaluating the evaluation of automatic machine translation evaluation metrics. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 4984–4997, Online. Association for Computational Linguistics.

Jihyung Moon, Hyunchang Cho, and Eunjeong L. Park. 2020. Revisiting round-trip translation for quality estimation. In Proceedings of the 22nd Annual Conference of the European Association for Machine Translation, pages 91–104, Lisboa, Portugal. European Association for Machine Translation.

Mathias Müller, Annette Rios, and Rico Sennrich. 2020. Domain robustness in neural machine translation. In Proceedings of the 14th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track), pages 151–164, Virtual. Association for Machine Translation in the Americas.

Xiao Pan, Mingxuan Wang, Liwei Wu, and Lei Li. 2021. Contrastive learning for many-to-many multilingual neural machine translation. 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 244–258, Online. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Weijing 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, Belgium, Brussels. Association for Computational Linguistics.

Ricardo Rei, Ana C Farinha, Chrysoula Zerva, Daan van Stigt, Craig Stewart, Pedro Ramos, Taisiya Glushkova, André FT Martins, and Alon Lavie. 2021. Are references really needed? unbabel-ist 2021 submission for the metrics shared task. In Proceedings of the Sixth Conference on Machine Translation, pages 1030–1040.

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, 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 911–920, Online. Association for Computational Linguistics.

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. Improving neural machine translation models with monolingual data. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pages 86–96.

Aditya Siddhant, Ankur Bapna, Orhan Firat, Yuan Cao, Mia Xu Chen, Isaac Caswell, and Xavier Garcia. 2022. Towards the next 1000 languages in multilingual machine translation: Exploring the synergy between supervised and self-supervised learning. arXiv preprint arXiv:2201.03110.

Matthew Snoover, 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.

Harold Somers. 2005. Round-trip translation: What is it good for? In Proceedings of the Australasian Language Technology Workshop 2005, pages 127–133.

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

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.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiaqiao Gu, and Angela Fan. 2020. Multilingual translation with extendible multilingual pretraining and finetuning. *arXiv preprint arXiv:2008.00401*.

Jörg Tiedemann and Santhosh Thottingal. 2020. Opusmt–building open translation services for the world. In *22nd Annual Conference of the European Association for Machine Translation*, page 479.

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, European Language Resources Association (ELRA).

David Vilar, Jia Xu, Luis Fernando d’Haro, and Hermann Ney. 2006. Error analysis of statistical machine translation output. In *Proceedings of the fifth international conference on language resources and evaluation (LREC’06)*.

Guillaume Wenzek, Vishrav Chaudhary, Angela Fan, Sahir Gomez, Naman Goyal, Somya Jain, Douwe Kiela, Tristan Thrush, and Francisco Guzmán. 2021. Findings of the wmt 2021 shared task on large-scale multilingual machine translation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 89–99.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google’s neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Menno van Zaanen and Simon Zwarts. 2006. Unsupervised measurement of translation quality using multi-engine, bi-directional translation. In *Australasian Joint Conference on Artificial Intelligence*, pages 1208–1214. Springer.

Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. 2020a. Improving massively multilingual neural machine translation and zero-shot translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1628–1639.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2020b. Bertscore: Evaluating text generation with bert. In *Proceedings of the Eighth International Conference on Learning Representations*.

Wei Zhao, Goran Glavaš, Maxime Peyrard, Yang Gao, Robert West, and Steffen Eger. 2020. On the limitations of cross-lingual encoders as exposed by reference-free machine translation evaluation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1656–1671.
A Dataset Construction

| Resource  | Language   | Scale | Usage       |
|-----------|------------|-------|-------------|
| High      | English    | -     | TRAIN+TEST  |
|           | Spanish    | 315M  | TRAIN+TEST  |
|           | French     | 289M  | TRAIN+TEST  |
|           | German     | 216M  | TRAIN+TEST  |
|           | Portuguese | 137M  | TRAIN+TEST  |
|           | Russian    | 127M  | TRAIN+TEST  |
|           | Italian    | 116M  | TRAIN+TEST  |
|           | Dutch      | 82.4M | TRAIN+TEST  |
|           | Turkish    | 41.2M | TRAIN+TEST  |
|           | Polish     | 40.9M | TRAIN+TEST  |
|           | Chinese    | 37.9M | TRAIN+TEST  |
|           | Romanian   | 31.9M | TRAIN+TEST  |
|           | Greek      | 23.7M | TRAIN+TEST  |
|           | Japanese   | 23.2M | TRAIN+TEST  |
|           | Czech      | 23.2M | TEST        |
|           | Finnish    | 15.2M | TEST        |
|           | Bulgarian  | 10.3M | TEST        |
|           | Lithuanian | 6.69M | TEST        |
|           | Estonian   | 4.82M | TEST        |
|           | Latvian    | 4.8M  | TEST        |
|           | Hindi      | 3.3M  | TEST        |
|           | Javanese   | 1.49M | TEST        |
|           | Icelandic  | 1.17M | TEST        |
| Low       | Tamil      | 992K  | TRAIN+TEST  |
|           | Armenian   | 977K  | TEST        |
|           | Azerbaijani| 867K  | TEST        |
|           | Kazakh     | 701K  | TRAIN+TEST  |
|           | Urdu       | 630K  | TEST        |
|           | Khmer      | 398K  | TRAIN+TEST  |
|           | Hausa      | 335K  | TRAIN+TEST  |
|           | Pashto     | 293K  | TRAIN+TEST  |
|           | Burmese    | 283K  | TEST        |
|           | Gujarati   | 160K  | TRAIN+TEST  |

Table 8: The statistics of FLORES-AE33. 20 languages are used in both training and test (TRAIN+TEST), the other 13 languages are used in test only (TEST).

We provide the statistics of all languages covered by FLORES-AE33, categorised by different scale of the resource (high, medium and low) and usage purpose (TRAIN+TEST and TEST) in Table 8. Scale is counted by the amount of bi-text data to English in FLORES-101 (Goyal et al., 2022a).

B Machine Translation Systems

**MBART50-M2M.** MBART50-M2M (Tang et al., 2020) is a multilingual translation model with many-to-many encoders and decoders. The model is trained on 50 publicly available language corpora with English as a pivot language.

**M2M-100-BASE & M2M-100-LARGE.** These two models are one of the first non-English-centric multilingual machine translation systems, which are trained on 100 languages covering high-resource to low-resource languages. Different from MBART50-M2M, M2M-100-BASE and M2M-100-LARGE (Fan et al., 2021) are trained on parallel multilingual corpora without an explicit centring language.

**OPUS-MT.** OPUS-MT (Tiedemann and Thottin-gal, 2020) is a collection of one-to-one machine translation models which are trained on corresponding parallel data from OPUS using MARIAN-NMT as backbone (Junczys-Dowmunt et al., 2018). The collection of MT models support 186 languages.

**GOOGLE-TRANS.** GOOGLE-TRANS (Wu et al., 2016; Bapna et al., 2022) is an online Translation service provided by Google Translation API, which supports 133 languages. The system is frequently involved as a baseline system by WMT shared tasks (Barrault et al., 2020).

C Implementation Details

**Regression Modeling.** We use the linear regression model tool by Scikit-Learn\(^{14}\) with the default setting of the API.

**MT Systems.** We adopt EasyNMT\(^{15}\) for loading MBART50-M2M, M2M-100-BASE, M2M-100-LARGE and OPUS-MT for translation.

**Computational Resource and Time.** In our experiment, we collect the translation results and compute their TRANS-SCORE and SELF-SCORE on multiple single-GPU servers with Nvidia A40. Overall, it cost us about three GPU months for collecting translation results by all the aforementioned MT systems.

D Measurement

We evaluate the performance of our predictive model via the following measurements:

**Mean Absolute Error (MAE)** is used for measuring the average magnitude of the errors in a set of predictions, indicating the accuracy for continuous variables.

\(^{14}\)https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html

\(^{15}\)https://github.com/UKPLab/EasyNMT
**Root Mean Square Error (RMSE)** measures the average magnitude of the error. Compared to MAE, RMSE gives relatively higher weights to larger error.

**Pearson’s r correlation** ([Benesty et al., 2009](#)) is officially used in WMT to evaluate the agreement between the automatic evaluation metrics and human judgement, emphasising on the translation consistency. In our paper, the metric evaluates the agreement between the predicted automatic evaluation scores and the ground truth.

**Kendall’s τ correlation** ([Kendall, 1938](#)) is another metric to evaluate the ordinal association between two measured quantities.

---

**Figure 4:** The first row is the correlations between $\text{SELF-SCORE}_{M \rightarrow B}$ and $\text{TRANS-SCORE}_{M \rightarrow B}$ on MBART50-M2M using (a) sacreBLEU, (b) spBLEU, (c) chrF and (d) BERTScore. The second row is the correlations between $\text{SELF-SCORE}_{B \rightarrow A}$ and $\text{TRANS-SCORE}_{A \rightarrow B}$ on MBART50-M2M using (e) sacreBLEU, (f) spBLEU, (g) chrF and (h) BERTScore. All experiments with overall Pearson’s $r$.

**Figure 5:** The first row is the correlations between $\text{SELF-SCORE}_{M \rightarrow B}$ and $\text{TRANS-SCORE}_{M \rightarrow B}$ on M2M-100-BASE using (a) sacreBLEU, (b) spBLEU, (c) chrF and (d) BERTScore. The second row is the correlations between $\text{SELF-SCORE}_{B \rightarrow A}$ and $\text{TRANS-SCORE}_{A \rightarrow B}$ on M2M-100-BASE using (e) sacreBLEU, (f) spBLEU, (g) chrF and (h) BERTScore. All experiments with overall Pearson’s $r$.

**E Supplementary Experiments**

**E.1 Statistics on FLORES–AE33 (RQ1)**
We visualise more detailed results of correlation between TRANS-SCOREs and SELF-SCOREs on Type I language pairs in FLORES–101, in Figure 4 (MBART50-M2M) and Figure 5 (M2M-100-BASE).

**E.2 WMT2020–News with Synthetic Competitors (RQ5)**
We increase the scale of competitors to WMT2020–News by introducing pseudo competitors. To mimic the number of a conventional WMT task, we vary 17 forward translation systems by randomly dropping 0% to 80% (with a step of 5%) tokens from the outputs of GOOGLE-TRANS. Then, we utilise the vanilla GOOGLE-TRANS
to translate these synthetic forward translation results back to the source language. We conduct experiments on de-fr, en-ta and zh-en, representing those non-En to non-En, En to non-En and non-En to En language pairs.

The results in Table 9 demonstrate the predictors’ performances on ranking the pseudo competitors on WMT2020-News based on spBLEU features. The overall ranking errors on 17 MT systems are small on all three selected language pairs.

The results in Table 9 demonstrate the predictors’ performances on ranking the pseudo competitors on WMT2020-News based on spBLEU features. The overall ranking errors on 17 MT systems are small on all three selected language pairs.

The overall ranking errors on 17 MT systems are small on all three selected language pairs.

| Language Pair | MAE ↓ | RMSE ↓ | τ ↑ | r ↑ |
|---------------|-------|--------|-----|-----|
| de-fr         | 2.21  | 2.67   | 1.00| 0.98|
| en-ta         | 0.88  | 0.98   | 1.00| 0.99|
| zh-en         | 1.69  | 2.37   | 1.00| 0.99|
| Average       | 1.59  | 2.01   | 1.00| 0.99|

Table 9: Results of prediction and ranking on translation quality of WMT2020-News synthetic data for three language pairs.

E.3 Ranking Experiments on WMT2020-Bio (RQ5)

We display the experimental results on WMT2020-Bio in the Table 10. The overall performance is positive, while it is relatively worse than the results of WMT2020-News reported in Table 7. We attribute this to the fact that the M used on WMT2020-Bio are calculated on document, while our regression models rely on sentence-level translation metrics in training. The large granularity difference of text may result in a distribution shift.

The results in Table 9 demonstrate the predictors’ performances on ranking the pseudo competitors on WMT2020-News based on spBLEU features. The overall ranking errors on 17 MT systems are small on all three selected language pairs.

The overall ranking errors on 17 MT systems are small on all three selected language pairs.

| Language Pair | MAE ↓ | RMSE ↓ | τ ↑ | r ↑ |
|---------------|-------|--------|-----|-----|
| de-fr         | 2.21  | 2.67   | 1.00| 0.98|
| en-ta         | 0.88  | 0.98   | 1.00| 0.99|
| zh-en         | 1.69  | 2.37   | 1.00| 0.99|
| Average       | 1.59  | 2.01   | 1.00| 0.99|

Table 9: Results of prediction and ranking on translation quality of WMT2020-News synthetic data for three language pairs.

E.4 Comparison with COMET (RQ5)

For WMT2020-News experiment, we provide the comparison between our predictor and COMET (Rei et al., 2020a) a neural evaluation tool for translation quality estimation. We consider Pearson’s correlation r and Kendall’s τ to evaluate the performance. We choose COMET-QE-MQM (Rei et al., 2021), a reference-free regression model for the experiment, which aligns to the setting in textbfRQ5. From Table 11, we observe that our predictor achieves overall better results to predict and rank translation quality than COMET, against the ground-truth Trans-Score^spBLEU^.

| Language Pair | Pearson’s r ↑ | Kendall’s τ ↑ |
|---------------|---------------|---------------|
|               | Our | COMET | Our | COMET |
| cs-en         | 0.91 | 0.95 | 0.60 | 0.60 |
| de-en         | 0.95 | 0.75 | 0.80 | 0.20 |
| de-fr         | 0.97 | 0.70 | 0.80 | 0.80 |
| en-cs         | 0.94 | 0.90 | 0.60 | 0.80 |
| en-de         | 0.92 | 0.78 | 1.00 | 0.40 |
| en-fr         | 0.85 | 0.91 | 0.40 | 1.00 |
| en-zh         | 0.80 | 0.91 | 0.80 | 0.80 |
| fr-de         | 0.94 | 0.92 | 1.00 | 0.80 |
| ru-en         | 0.85 | 0.71 | 0.80 | 0.40 |
| zh-en         | 0.50 | 0.53 | 0.20 | 0.40 |
| Average       | 0.86 | 0.81 | 0.70 | 0.62 |

Table 11: Comparison of our predictors and COMET on ranking the translation quality of the selected MT systems on WMT2020-News shared tasks.