MATESE: Machine Translation Evaluation as a Sequence Tagging Problem

Stefano Perrella[1], **Lorenzo Proietti**[1], **Alessandro Scirè**[1,2]
Niccolò Campolungo[1] and **Roberto Navigli**[1]
[1]Sapienza NLP Group, Sapienza University of Rome
[2]Babelscape, Italy
{stefano.perrella, l.proietti, alessandro.scire}@uniroma1.it
campolungo@di.uniroma1.it navigli@diag.uniroma1.it

Abstract

Starting from last year, WMT human evaluation has been performed within the Multidimensional Quality Metrics (MQM) framework, where human annotators are asked to identify error spans in translations, alongside an error category and a severity. In this paper, we describe our submission to the WMT 2022 Metrics Shared Task, where we propose using the same paradigm for automatic evaluation: we present the MATESE metrics, which reframe machine translation evaluation as a sequence tagging problem. Our submission also includes a reference-free metric, denominated MATESE-QE. Despite the paucity of the openly available MQM data, our metrics obtain promising results, showing high levels of correlation with human judgements, while also enabling an evaluation that is interpretable. Moreover, MATESE-QE can also be employed in settings where it is infeasible to curate reference translations manually.

1 Introduction and Related Work

For many years, Machine Translation (MT) has mainly been evaluated using untrained evaluation techniques, such as BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005) and CHRF (Popović, 2015), which rely heavily on lexical-level matching of either token, or character, n-grams. Unfortunately, these metrics present two major drawbacks: i) it is not possible to carry out the evaluation without manually-curated references and, most importantly, ii) the evaluation is too dependent on the surface form of the translation, and its reference. More recently, attempts have been made to address these problems using machine-learned metrics, which have shown better correlations with human judgements (Mathur et al., 2020). More specifically, last year's WMT Metrics Shared Task saw C-SPEC$_{\text{PN}}$ (Takahashi et al., 2021), BLEURT-20[1] and COMET-MQM_2021 (Rei et al., 2021) emerge as distinctly better than the other participants (Freitag et al., 2021b). These metrics consist of regression models trained to mimic human annotators by directly assigning quality scalar scores to candidate translations. In detail, COMET-MQM_2021 is based on the Estimator architecture introduced by Rei et al. (2020), where features extracted from the embeddings of the source sentence, candidate translation, and reference translation are passed to a feed-forward regressor; C-SPEC first concatenates the embeddings derived from paired inputs of candidate-source and candidate-reference, and then passes the resulting vector to a multi-layer perceptron; BLEURT, instead, feeds the candidate translation and its reference to Rebalanced mBERT (Chung et al., 2021), and regresses on the representation provided by the [CLS] token. Moreover, BLEURT and C-SPEC add automatically-generated negative pairs to the standard training data: BLEURT applies random token perturbations, while C-SPEC uses Word Attribute Transfer to replace words in the translations. Although undoubtedly effective, regression metrics have the major drawback of not being interpretable, meaning that users are not able to gauge the quality of assessments that are returned, which is of paramount importance for an evaluation metric.

Recently, Freitag et al. (2021a) have proposed a shift in the standard practices for human machine translation evaluation, employing the Multidimensional Quality Metrics framework (Lommel et al., 2014, MQM), and moving away from Direct Assessments (Graham et al., 2013, DA), which were computed via requiring (even non-expert) annotators to assign a scalar value to a candidate translation, given a reference. Furthermore, Freitag et al. (2021a) pointed out the limitations of non-professional Direct Assessments, also show-

[1]BLEURT-20 is the retrained version of the previous year's BLEURT submission (Sellam et al., 2020).

Proceedings of the Seventh Conference on Machine Translation (WMT), pages 569–577
Abu Dhabi, December 7–8, 2022. ©2022 Association for Computational Linguistics

ing their unreliability compared to MQM. Indeed, differently from Direct Assessments, annotators who follow the MQM guidelines look at the source sentence rather than the reference, and are expected to tag the spans of the candidate translations that contain errors,[2] together with their error category (e.g., `Fluency/Grammar`, `Fluency/Punctuation` or `Style/Awkward`) and severity (e.g., `Major` or `Minor`), which, combined, determine the score associated with the error span. Finally, a scalar quality score for the entire sentence is derived from the various annotated spans.

In this work, we introduce the MATESE and MATESE-QE metrics, reframing the evaluation of machine-translated text as a sequence tagging problem based on the MQM framework, in an attempt to develop metrics that are interpretable, while also displaying high levels of correlation with human judgements.

2 MATESE Metrics

Inspired by the novel MQM evaluation framework, our work aims at employing a similar paradigm for automatic evaluation. We propose the MATESE metrics which, given a candidate translation and its reference (or source, for MATESE-QE), assign a label to each token of the candidate. These labels identify error spans, together with their severity, chosen among `Major` and `Minor`. Finally, in order to associate a score with the entire tagged sentence, we follow a weighting scheme similar to the one presented by Freitag et al. (2021a) for MQM-based human evaluation: we assign a score to an entire error span based on its severity, i.e., -5 and -1 for `Major` and `Minor`, respectively. The score assigned to a translation is the sum of the scores assigned to its error spans, with a minimum total score of -25. Following Freitag et al. (2021a), we compute a corpus-level score by averaging the scores of the sentences in the corpus. Although human MQM annotators are asked to report a maximum of 5 errors per translation,[3] we decided to let our metrics detect as many errors as they can find; nevertheless, in order to keep our scores in the same range as those computed on gold MQM annotations, we set a minimum score of -25, which is equal to the sum of 5 `Major` errors. Figure 1 shows an example of the annotations returned by our metrics.

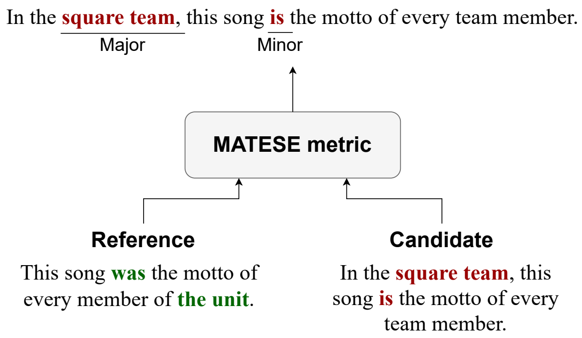

Figure 1: Example of the annotation returned by the MATESE metrics, given a candidate translation and its reference. The final score of the translation is -6, that is the sum of -5 and -1, assigned to the `Major` and `Minor` errors, respectively.

2.1 Data pre-processing

According to the MQM guidelines, mistranslated spans are tagged with an error category and a severity. To reduce the granularity of the annotations, we apply some transformations to the original data, which we report below:

1. We discard annotations of the `Non-translation` category, since they are weighted -25 by Freitag et al. (2021a), and would have required a special treatment, but are too scarce ($< 0.1\%$ of the whole data) for the model to learn how to assign them;
2. We discard annotations referring to either `Accuracy/Omission` or `Source error` categories, since in these cases the annotation might be in the source sentence, while our models are trained to tag the candidate translation only;
3. We discard annotations of errors with `Neutral` severity, since they are highly subjective and do not participate in the computation of the final quality score (Freitag et al., 2021a);
4. We replace `Critical` severity labels with `Major`, in order to make the English→Russian dataset conform to the rest of the data;
5. We discard all the MQM error categories, leaving only information about error severity. While we believe error categorization to

[2] In a few cases, the source sentence might also be annotated. An example of this is with omission errors, where annotators report the spans of the source sentence which are missing from the candidate translation.

[3] This holds only for the MQM guidelines released by Freitag et al. (2021a).

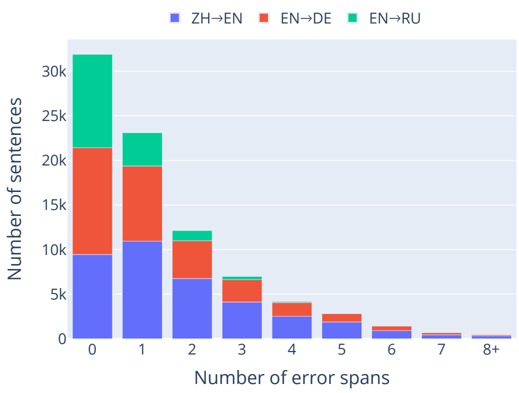

Figure 2: Distribution of the number of error spans over sentences in both training and test data of WMT 2021 Metrics Shared Task, after the pre-processing we described in Section 2.1.

> be of great importance, we decided to remove it because of the limited availability of training data and to avoid making the classification problem too sparse.

Furthermore, in the MQM data released by Freitag et al. (2021a), every sentence has been annotated 3 times, each one by a different rater. In order to yield a single sample per sentence and maximize the number of annotations, we merge the annotations of the different raters into a single annotated sentence;[4] in the case when there is even only a partial overlap between two annotated spans, we discard the one associated with the `Minor` error in favor of the `Major`, or pick one or the other randomly if they have the same severity. We decided to keep `Majors` over `Minors` because Freitag et al. (2021a) obtained almost the same ranking of MT systems when considering only `Major` errors, compared to the full MQM score.

2.2 Hypothesis and Target Span Hit metrics

Typically, MT evaluation metrics' quality is assessed through their correlations with human judgements. Nevertheless, our novel formulation of MT evaluation as a sequence tagging problem allows us to estimate the quality of our metrics also via the produced error spans. Specifically, we are interested in determining how well our metrics are able to flag, even partially, a true error span, regardless of its severity or length. However, existing span-level metrics, such as Span Precision, Span Recall and Span F1, focus on exact overlaps between predicted spans and target ones. Moreover, correlations with MQM scores paint only a partial picture, since the final score assigned to a translation depends only on the number of error spans (with their severity), but not on their position in the sentence. For instance, if a system flagged a span as a `Major` error, but the target annotation had a different span tagged as `Major`, the MQM scores would be identical despite the tagging error.

[4]Therefore, in our merged sentences the number of error spans per translation can be greater than 5. Figure 2 reports the distribution of error spans in our entire data.

Figure 3: An example of evaluation with the Hypothesis and Target Span Hit metrics. The **turquoise line —** (below) and **amber line —** (above) represent the hypothesis and target annotation, respectively. HSH $= 2/3$ (2 out of 3 spans are *hit*), TSH $= 2/2$ (2 out of 2 spans are *hit*).

To address these issues, we introduce the Hypothesis Span Hit (HSH) and Target Span Hit (TSH) metrics: HSH represents the percentage of predicted error spans that are also, at least partially, true; instead, TSH represents the percentage of true error spans that the metric has predicted, even partially. An example of their assessments is given in Figure 3.

Formal definition Let us consider a candidate translation $\boldsymbol{c}$ as a sequence of tokens $(c_1, c_2, \ldots, c_n)$; moreover, let us define an error span s as a set of contiguous tokens in $\boldsymbol{c}$, e.g., $\{c_1, c_2, c_3\}$, and an error annotation A as a set of disjoint error spans, i.e., that satisfies $\bigcap_{s' \in A} s' = \emptyset$. Furthermore, we define the Span Hit Indicator as

$$\mathrm{SHI}(s, A) = \mathbb{I}(s \cap \sigma(A) \neq \emptyset)$$

where $\sigma(A) = \bigcup_{s' \in A} s'$, i.e., the set of all tokens in annotation A. In simpler terms, $\mathrm{SHI}(s, A)$ is 1 if at least one of the tokens in s belongs to the set of all tokens of the error spans in A.

Finally, let us take two error annotations: A_h represents the hypothesis spans produced by a model, while A_t represents the target spans that $\boldsymbol{c}$ was originally annotated with. We define the Hypothesis Span Hit and Target Span Hit metrics as follows:

$$\mathrm{HSH}(A_h, A_t) = \frac{\sum_{s_h \in A_h} \mathrm{SHI}(s_h, A_t)}{|A_h|}$$

$$\mathrm{TSH}(A_t, A_h) = \frac{\sum_{s_t \in A_t} \mathrm{SHI}(s_t, A_h)}{|A_t|}$$

Both metrics are defined as the average number of span hits of one error annotation with respect to the other. To compute the metrics for an entire dataset we employ micro-averaging, i.e., we concatenate all hypotheses into a single one, do the same for the targets, and then measure Span Hit metrics on the newly-created pair of hypothesis and target. We avoid averaging the single results because the number of spans varies widely across samples (Figure 2).

3 Experimental Setup

In this Section, we describe the different architectures we experiment with, the data for training and evaluation, and the metrics we use to measure performances.

3.1 Architectures

Since it is rather convenient to have a single model capable of evaluating text in multiple languages, we leverage multilingual pre-trained models like XLM-RoBERTa (Conneau et al., 2020) and mBART (Liu et al., 2020). In order to compare the performances of multilingual models with their English-only counterparts, we also experiment with RoBERTa (Liu et al., 2019).[5]

Encoder-only models XLM-RoBERTa and RoBERTa models consist of only the encoder part of the standard Transformer architecture (Vaswani et al., 2017). The input we provide to the encoder models is the concatenation of the candidate translation and its reference (or source, for MATESE-QE), separated by a </s> token. Furthermore, we add two randomly-initialized encoder layers on top of the last layer, as well as a classification head. Due to computational constraints, we keep the embedding layer frozen.

Encoder-decoder model When experimenting with mBART, we feed the reference translation (or the source, for MATESE-QE) to the encoder, and the candidate to the decoder, so as to maintain similarity with the pre-training process. We highlight that we do not use the decoder autoregressively; instead, following the standard practice for sequence classification with encoder-decoder models, we force the candidate to be processed all at once, and collect the contextualized embeddings at the last layer. On top of the decoder, we add two randomly-initialized encoder layers, and a classification head. As with the encoder-only models, due to computational constraints the embedding layer is frozen.

[5]RoBERTa can be employed only for reference-based evaluation, and with language pairs that have English as target language: in our case, this is only Chinese→English.

3.2 Training and validation data

In order to perform our experiments employing all the existing MQM data, we experiment using a 90/10 training/validation split of the concatenation of the training set (which is the MQM data released by Freitag et al. (2021a)) and the test sets of WMT 2021 Metrics Shared Task (Freitag et al., 2021b).

Moreover, to make a fair comparison between the MATESE metrics and the ones submitted to the aforementioned Shared Task, we also retrain our systems using only the above-mentioned training set, with the same split. We dub these systems MATESE^{21} and MATESE-QE^{21}.

In both settings, we use only English→German and Chinese→English data. Moreover, we point out that the split is performed on *unique source sentences*: since each source sentence is translated by multiple systems, our split avoids having translations of the same source sentence be present in both the training and validation splits.

WMT Submission Training Split For our final submission to the WMT 2022 Metrics Shared Task, we include English→Russian data to the concatenation of the training and test sets of the WMT 2021 Metrics Shared Task. We split the whole data 5 times, each time taking 90% for training and 10% for validation, and train 5 different systems (10 if we also consider MATESE-QE). In our submission, each score is the median prediction of the systems trained on the 5 different data splits.

3.3 Evaluation metrics

The MATESE metrics tag the spans of a candidate translation that contain an error. Following the BIO scheme (Ramshaw and Marcus, 1995), we assign to each token a label in $L = \{\texttt{O}, \texttt{B-Minor}, \texttt{I-Minor}, \texttt{B-Major}, \texttt{I-Major}\}$; a final score for the annotated sentence is then obtained as the sum of the individual spans' scores. We can evaluate the performances of our metrics according to the final scores, as well as in terms of the produced annotations: indeed, we use the scalar scores to rank translations and measure the correlations with human judgements, and we measure the tagging accuracy with respect to the gold annotations. In the latter

	O	B-Minor	I-Minor	B-Major	I-Major
EN→DE	818,945	32,667	37,897	8516	25,192
ZH→EN	1,053,663	33,633	48,333	33,996	76,984
EN→RU	343,449	614	1015	7271	3189
ALL	2,216,057	66,914	87,245	49,783	105,365

Table 1: Distribution of the token-level gold annotations in the concatenation of the training and test sets of WMT 2021 Metrics Shared Task, after the pre-processing we described in Section 2.1.

	P	R	F1	HSH	TSH
XLM-R^LARGE	47.38	**38.40**	**41.72**	57.73	**46.08**
XLM-R^BASE	46.64	34.12	37.93	**58.01**	38.70
mBART	**47.97**	31.94	36.01	55.85	32.66

Table 2: Comparison of different architectures in terms of Precision, Recall and F1-score in their macro versions; HSH and TSH are Hypothesis Span Hit and Target Span Hit metrics.

case, we rely on the standard classification metrics of Precision, Recall and F1-score, computed using TorchMetrics[6] modules. More specifically, given that our data is highly imbalanced (see Table 1), we employ macro versions of these metrics and, in particular, use the macro-F1 score to select the best checkpoint of the models on the validation set. Furthermore, in order to assess the span-level error detection capabilities of our systems, we employ the Hypothesis and Target Span Hit metrics as defined in Section 2.2.

4 Results

In this Section, we show the results obtained by our metrics. Unless explicitly specified, all experiments have been performed using reference-based systems.

4.1 Architectures comparison

We can see the results of comparing the aforementioned architectures in Table 2. The best performing architecture is XLM-R^LARGE, which attains the highest F1-score, as a consequence of achieving the best Recall. Considering the complexity of the task, and the imbalance of the data, we conjecture that the other architectures obtain high Precision and low Recall scores because they are able to predict only the errors that are easier to detect, while assigning `O`s more frequently. This is also confirmed by the TSH score which, ruling `O` labels out of the computation, exacerbates the difference between different architectures, with XLM-R^BASE and mBART clearly failing to detect a higher number of errors of the target annotation compared to XLM-R^LARGE. An additional interesting fact that emerges from this comparison is that XLM-R architectures perform better than mBART, with XLM-R^BASE outperforming it despite having less than half of its parameters.

[6]https://github.com/Lightning-AI/metrics

4.2 Monolingual-multilingual comparison

Table 3 reports the results of training the same XLM-R model using a single language pair at a time, or both. Moreover, we test whether an English language model like RoBERTa outperforms XLM-R, when dealing with English-only data. Our results show that training on the whole data is beneficial to the task, with XLM-R^ALL obtaining a higher Recall and Target Span Hit in both language pairs, and an F1-score that is higher, or on par with, its variants. Similarly to what happens with different architectures, we hypothesize that training on more data enables the models to detect a wider range of errors, even if the additional data is in a different language. We do not record significant differences in the results obtained by RoBERTa, compared to XLM-R^MONO on Chinese→English data.

4.3 MATESE-QE

A desirable feature of evaluation metrics is to function both in the presence and the absence of humanly-curated references. To achieve this, we investigate whether it is feasible to tag the errors in the candidate translation by looking at the source sentence only. Table 4 reports the results obtained by the best architecture, i.e., XLM-R^LARGE, trained on both English→German and Chinese→English, both when disposing of the reference sentence, and not.

MATESE outperforms MATESE-QE in terms of Recall, F1-score and Target Span Hit metrics. Clearly, the information found in the reference is easier to exploit, and the reference-based system is able to detect a much wider range of errors. At the same time, MATESE-QE proves to be a viable alternative in the absence of manually-curated references: it displays high levels of Precision and Hypothesis Span Hit, which means that it outputs predictions that are more accurate than those of MATESE, even if only for the range of errors that it is able to detect.

	EN→DE					ZH→EN				
	P	R	F1	HSH	TSH	P	R	F1	HSH	TSH
XLM-R$^{\text{ALL}}$	43.03	**33.09**	**35.89**	54.35	**44.54**	47.86	**39.39**	42.63	59.90	**47.04**
XLM-R$^{\text{MONO}}$	**43.98**	30.64	33.52	**56.56**	39.67	50.85	38.51	**42.77**	63.51	44.38
RoBERTa	–	–	–	–	–	**51.49**	37.91	42.41	**64.33**	42.81

Table 3: Model performances on monolingual and multilingual settings. XLM-R$^{\text{ALL}}$ is trained and evaluated on the concatenation of English→German (EN→DE) and Chinese→English (ZH→EN) datasets, while XLM-R$^{\text{MONO}}$ stands for two different models, each one trained and evaluated on a single dataset. RoBERTa is an English language model, and therefore can deal with ZH→EN data only.

	P	R	F1	HSH	TSH
MATESE	47.38	**38.40**	**41.72**	57.73	**46.08**
MATESE-QE	**49.34**	34.53	38.89	**59.89**	36.84

Table 4: Comparison of our reference-based and reference-free systems, i.e., MATESE and MATESE-QE, respectively. The only difference between the two is that MATESE-QE uses the source sentence in place of the reference.

4.4 Correlations with Human Judgements

Tables 5a and 5b report the correlations with human judgements that our metrics attained on `newstest2021` (in-domain) and `TED` (out-of-domain) test sets of last year's WMT Metrics Shared Task: `w/ HT` means that manually-curated references have been scored together with system outputs, while `w/o HT` means that those references have been kept out of the evaluation. Aside from our systems, i.e., MATESE^{21} and MATESE-QE^{21}, we also report two additional baselines: #1 WMT and #2 WMT. These are the top-1 and top-2 results reported by Freitag et al. (2021b) in the corresponding tables (Tables 23, 24, 27 and 28). Since those positions are held by different systems, we assign each submission a unique symbol and report the mapping in Appendix A.

Generally speaking, for in-domain settings, we observe that, on English→German, MATESE^{21} and MATESE-QE^{21} achieve correlations on par or better than the top-2 WMT 2021 submissions, while on Chinese→English the results are slightly worse. Interestingly, in out-of-domain settings, we observe a sizeable drop in correlation on both translation directions. We attribute this drop to the very limited amount of training data, which probably hinders proper generalization capabilities to out-of-domain settings. Finally, we observe that MATESE-QE^{21} lags behind MATESE^{21} by a relatively small margin.

	EN→DE			ZH→EN		
	w/o HT	w/ HT	TED	w/o HT	w/ HT	TED
#1 WMT	‡0.938	$^{\perp}$0.823	$^{\|}$**0.818**	$^{\wedge}$**0.834**	$^{\wedge}$**0.727**	$^{\vee}$**0.421**
#2 WMT	†0.937	$^{\perp}$0.822	$^{\perp}$0.802	$^{|}$0.628	$^{|}$0.619	$^{\top}$0.403
MATESE^{21}	**0.946**	**0.863**	0.621	0.636	0.701	0.017
MATESE-QE^{21}	0.910	0.806	0.584	0.502	0.600	0.056

(a) System-level Pearson correlations.

	EN→DE			ZH→EN		
	w/o HT	w/ HT	TED	w/o HT	w/ HT	TED
#1 WMT	$^{\perp}$0.267	$^{\perp}$0.256	$^{\wedge}$**0.290**	$^{\perp}$**0.402**	$^{\perp}$**0.390**	$^{\wedge}$0.248
#2 WMT	$^{\perp}$0.266	$^{\perp}$0.254	$^{\perp}$0.285	$^{\perp}$0.401	$^{\perp}$0.388	$^{\perp}$0.241
MATESE^{21}	**0.323**	**0.310**	0.271	0.358	0.346	**0.257**
MATESE-QE^{21}	0.288	0.277	0.210	0.343	0.332	0.196

(b) Segment-level Kendall correlations.

Table 5: System- and segment-level correlations with human judgements as measured in WMT 2021 Metrics Shared Task (Freitag et al., 2021b). MATESE^{21} and MATESE-QE^{21} are MATESE metrics that have been re-trained using only the training set of the Shared Task. A legend of the other symbols is found in Appendix A.

5 Conclusions

In this paper, we described our submission to the WMT 2022 Metrics Shared Task: we presented the MATESE metrics, a new way of automatically assessing the quality of translations, putting forward evaluation techniques that are interpretable, while at the same time displaying high levels of correlation with human judgements. Scores are in the same ballpark of the best performing metrics proposed in the WMT 2021 Metrics Shared Task. Furthermore, the MATESE metrics can also be used in the absence of humanly-curated references, with MATESE-QE being slightly less accurate than its reference-based counterpart, but still presenting encouraging levels of correlation with human judgements. In future work, we plan to improve the MATESE metrics to also detect the type of errors, and not only their severity, in order to approximate even better the MQM annotation process.

Acknowledgements

The authors gratefully acknowledge the support of the ERC Consolidator Grant MOUSSE No. 726487.

The authors also acknowledge the support of the PerLIR project (Personal Linguistic resources in Information Retrieval) funded by the MIUR Progetti di ricerca di Rilevante Interesse Nazionale programme (PRIN 2017).

Limitations

Poor generalization We expect the MATESE metrics' generalization capabilities to be hindered by the narrow range of errors that they are trained upon. Indeed, while the number of samples in the datasets is relatively large (around 80K annotated sentences), the number of unique sources is much smaller (around 6K), because the annotations are performed on the same source sentences translated by multiple MT systems. In fact, we observe a drop in performance in the out-of-domain setting, i.e., the TED dataset.

Computational requirements The MATESE metrics require a non-negligible computational budget, especially when compared to their untrained alternatives, such as BLEU, METEOR or CHRF. Given that the task we tackle is arguably challenging, and that we need semantically-rich representations of the analyzed sentences, we decided to rely upon a large Transformer encoder, which makes the evaluation computationally intensive. Unfortunately, the comparison between XLM-RoBERTa Large and its Base counterpart shows that a sizeable improvement is due to the increased size of the model.

References

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

Hyung Won Chung, Thibault Fevry, Henry Tsai, Melvin Johnson, and Sebastian Ruder. 2021. Rethinking embedding coupling in pre-trained language models. In *International Conference on Learning Representations*.

Alexis Conneau, 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. Association for Computational Linguistics.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021a. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, George Foster, Alon Lavie, and Ondřej Bojar. 2021b. 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. Association for Computational Linguistics.

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

Lifeng Han, 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. Association for Computational Linguistics.

Mateusz Krubiński, Erfan Ghadery, Marie-Francine Moens, and Pavel Pecina. 2021a. Just ask! evaluating machine translation by asking and answering questions. In *Proceedings of the Sixth Conference on Machine Translation*, pages 495–506, Online. Association for Computational Linguistics.

Mateusz Krubiński, Erfan Ghadery, Marie-Francine Moens, and Pavel Pecina. 2021b. MTEQA at WMT21 metrics shared task. In *Proceedings of the Sixth Conference on Machine Translation*, pages 1024–1029, Online. Association for Computational Linguistics.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pretraining for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Yinhan Liu, 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.

Arle Lommel, Aljoscha Burchardt, and Hans Uszkoreit. 2014. Multidimensional quality metrics (mqm): A framework for declaring and describing translation quality metrics. *Tradumàtica: tecnologies de la traducció*, 0:455–463.

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

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

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, Lisbon, Portugal. Association for Computational Linguistics.

Lance Ramshaw and Mitch Marcus. 1995. Text chunking using transformation-based learning. In *Third Workshop on Very Large Corpora*.

Ricardo Rei, Ana C Farinha, Chrysoula Zerva, Daan van Stigt, Craig Stewart, Pedro Ramos, Taisiya Glushkova, André F. T. 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, Online. Association for Computational Linguistics.

Ricardo Rei, 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. Association for Computational Linguistics.

Thibault Sellam, Amy Pu, Hyung Won Chung, Sebastian Gehrmann, Qijun Tan, Markus Freitag, Dipanjan Das, and Ankur Parikh. 2020. Learning to evaluate translation beyond English: BLEURT submissions to the WMT metrics 2020 shared task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 921–927, Online. Association for Computational Linguistics.

Matthew Snover, Bonnie Dorr, Rich 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, Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas.

Michal Stefanik, Vít Novotný, and Petr Sojka. 2021. Regressive ensemble for machine translation quality evaluation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 1041–1048, Online. Association for Computational Linguistics.

Kosuke Takahashi, Yoichi Ishibashi, Katsuhito Sudoh, and Satoshi Nakamura. 2021. Multilingual machine translation evaluation metrics fine-tuned on pseudo-negative examples for WMT 2021 metrics task. In *Proceedings of the Sixth Conference on Machine Translation*, pages 1049–1052, Online. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

A WMT 2021 System Mapping

- ‡: cushLEPOR(LM) (Han et al., 2021);
- $\perp$: C-SPEC and C-SPECpn (Takahashi et al., 2021);
- $\wedge$: tgt-regEMT and tgt-regEMT-baseline (Stefanik et al., 2021);
- $\|$: COMET-MQM_2021 and COMET-QE-MQM_2021-src (Rei et al., 2021);
- $\vee$: TER (Snover et al., 2006);
- †: BLEU (Papineni et al., 2002);
- $\top$: MTEQA (Krubiński et al., 2021a,b).