MATESE: Machine Translation Evaluation as a Sequence Tagging Problem

Stefano Perrella\textsuperscript{1}, Lorenzo Proietti\textsuperscript{1}, Alessandro Scire\textsuperscript{1,2}
Niccolò Campolungo\textsuperscript{1} and Roberto Navigli\textsuperscript{1}

\textsuperscript{1}Sapienza NLP Group, Sapienza University of Rome
\textsuperscript{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 Multi-dimensional 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\textsubscript{PN} (Takahashi et al., 2021), BLEURT-20\textsuperscript{1} and COMET-MQM\textsubscript{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\textsubscript{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 Multi-dimensional Quality Metrics framework (Lommele 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-
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, 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, 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.

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
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;\footnote{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.} 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, as 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.

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 \( c \) 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 \( c \), e.g., \( \{c_1, c_2, c_3\} \), and an error annotation \( A \) as a set of disjoint error spans, i.e., that satisfies \( \bigcap_{A' \in A} s' = \emptyset \). Furthermore, we define the Span Hit Indicator as

\[
\text{SHI}(s, A) = \mathbb{I}(s \cap \sigma(A) \neq \emptyset)
\]

where \( \sigma(A) = \bigcup_{A' \in A} s' \), i.e., the set of all tokens in annotation \( A \). In simpler terms, \( \text{SHI}(s, A) = 1 \) 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 \( c \) was originally annotated with. We define the Hypothesis Span Hit and Target Span Hit metrics as follows:

\[
\text{HSH}(A_h, A_t) = \frac{\sum_{s_h \in A_h} \text{SHI}(s_h, A_t)}{|A_h|}
\]

\[
\text{TSH}(A_t, A_h) = \frac{\sum_{s_t \in A_t} \text{SHI}(s_t, A_h)}{|A_t|}
\]

**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).
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.

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 \(L = \{O, B\text{-Minor}, I\text{-Minor}, B\text{-Major}, I\text{-Major}\}\); 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

\(^5\)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.
We can see the results of comparing the aforemen-
tioned architectures in Table 2. The best perform-
ing architecture is XLM-R\textsubscript{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 pre-
dict only the errors that are easier to detect, while
assigning 0s more frequently. This is also con-
firmed by the TSH score which, ruling 0 labels out
of the computation, exacerbates the difference be-
tween different architectures, with XLM-R\textsubscript{BASE} and
mBART clearly failing to detect a higher number of
errors of the target annotation compared to XLM-
R\textsubscript{LARGE}. An additional interesting fact that emerges
from this comparison is that XLM-R architectures
perform better than mBART, with XLM-R\textsubscript{BASE} out-
performing it despite having less than half of its
parameters.

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.

<table>
<thead>
<tr>
<th>Language Pair</th>
<th>0</th>
<th>B-Minor</th>
<th>I-Minor</th>
<th>B-Major</th>
<th>I-Major</th>
</tr>
</thead>
<tbody>
<tr>
<td>EN→DE</td>
<td>818,945</td>
<td>32,667</td>
<td>37,897</td>
<td>8516</td>
<td>25,192</td>
</tr>
<tr>
<td>ZH→EN</td>
<td>1,053,663</td>
<td>33,633</td>
<td>48,333</td>
<td>33,996</td>
<td>76,984</td>
</tr>
<tr>
<td>EN→RU</td>
<td>343,449</td>
<td>614</td>
<td>1015</td>
<td>7271</td>
<td>3189</td>
</tr>
<tr>
<td>ALL</td>
<td>2,216,057</td>
<td>66,914</td>
<td>87,245</td>
<td>49,783</td>
<td>105,365</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>Architecture</th>
<th>P</th>
<th>R</th>
<th>F1</th>
<th>HSH</th>
<th>TSH</th>
</tr>
</thead>
<tbody>
<tr>
<td>XLM-R\textsubscript{LARGE}</td>
<td>47.38</td>
<td>38.40</td>
<td>41.72</td>
<td>57.73</td>
<td>46.08</td>
</tr>
<tr>
<td>XLM-R\textsubscript{BASE}</td>
<td>46.64</td>
<td>34.12</td>
<td>37.93</td>
<td>58.01</td>
<td>38.70</td>
</tr>
<tr>
<td>mBART</td>
<td>47.97</td>
<td>31.94</td>
<td>36.01</td>
<td>55.85</td>
<td>32.66</td>
</tr>
</tbody>
</table>

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 En-
lish language model like RoBERTa outperforms
XLM-R, when dealing with English-only data. Our
results show that training on the whole data is ben-
eficial to the task, with XLM-R\textsubscript{ALL}, obtaining a
higher Recall and Target Span Hit in both language
pairs, and an F1-score that is higher, or on par with,
it’s variants. Similarly to what happens with dif-
ferent 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\textsubscript{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\textsubscript{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 ref-
ences: 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.

\(^{6}\text{https://github.com/Lightning-AI/metrics}\)
we observe a sizeable drop in correlation on both
while on Chinese
we also report two additional baselines: #1 WMT
Table 4: Comparison of our reference-based
ALL
Table 3: Model performances on monolingual and multilingual settings. XLM-R
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
→
German (EN
→
DE ZH
→
EN datasets, while XLM-R
MONO
is trained and evaluated on
EN
→
DE
ZH
→
EN
variables 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.

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.

Table 5: System- and segment-level Kendall 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
valuation 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 manually-curated references, with MATESE-QE being slightly less accurate than its reference-based counterpart, but still presenting
courageous 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


Yinhua Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis,


A WMT 2021 System Mapping

• †: cushLEPOR(LM) (Han et al., 2021);

• ‰: C-SPEC and C-SPECpn (Takahashi et al., 2021);

• ∧: tgt-regEMT and tgt-regEMT-baseline (Stefanik et al., 2021);

• ‖: COMET-MQM_2021 and COMET-QE-MQM_2021-src (Rei et al., 2021);

• V: TER (Snover et al., 2006);

• †: BLEU (Papineni et al., 2002);

• ⊤: MTEQA (Krubínski et al., 2021a,b).