Multi-Dimensional Machine Translation Evaluation: Model Evaluation and Resource for Korean

Dojun Park, Sebastian Padó
IMS, University of Stuttgart, Germany
{dojun.park,sebastian.pado}@ims.uni-stuttgart.de

Abstract

Almost all frameworks for the manual or automatic evaluation of machine translation characterize the quality of an MT output with a single number. An exception is the Multidimensional Quality Metrics (MQM) framework which offers a fine-grained ontology of quality dimensions for scoring (such as style, fluency, accuracy, and terminology). Previous studies have demonstrated the feasibility of MQM annotation but there are, to our knowledge, no computational models that predict MQM scores for novel texts, due to a lack of resources. In this paper, we address these shortcomings by (a) providing a 1200-sentence MQM evaluation benchmark for the language pair English–Korean and (b) reframing MT evaluation as the multi-task problem of simultaneously predicting several MQM scores using SOTA language models, both in a reference-based MT evaluation setup and a reference-free quality estimation (QE) setup. We find that reference-free setup outperforms its counterpart in the style dimension while reference-based models retain an edge regarding accuracy. Overall, RemBERT emerges as the most promising model. Through our evaluation, we offer an insight into the translation quality in a more fine-grained, interpretable manner.

Keywords: Corpus, Evaluation Methodologies, Explainability, Machine Translation, Multilinguality

1. Introduction

Machine Translation (MT) evaluation refers to assessing the quality of translations generated by MT systems. Since the early stages of machine translation, evaluating its output has been an integral concern, as it enables systems to be assessed, refined, and enhanced (Dorr et al., 2011). As recent transformative advances in MT technologies (Bahdanau et al., 2015; Wu et al., 2016; Vaswani et al., 2017) enabled higher-quality, more nuanced translations, the importance of MT evaluation has evolved correspondingly, requiring a more sophisticated ability to measure quality accurately.

While recent advancements by neural metrics (Zhang et al., 2020; Sellam et al., 2020; Rei et al., 2020) have significantly brought the evaluation of machine translation forwards, virtually all current work on automatic MT evaluation distills the complexity of translation quality into a single score to be annotated and predicted, respectively. A single-score approach, while greatly simplifying computational modeling, arguably falls short of capturing the inherent multidimensional concept of translation quality. This limitation underscores the importance of a fine-grained evaluation. As an illustration, consider the three types of translation errors in Table 1: *Accuracy* errors fail to convey the meaning of the input. *Fluency* errors arise from outputs that fail to be grammatical and natural. *Style* errors finally change the style of the input substantially.

Being able to distinguish different aspects of translation quality allows for a fairer comparison across different MT systems in a line and reveals their strengths and weaknesses across different aspects of evaluation criteria (Avramidis et al., 2018; Klubicka et al., 2018). Interestingly, early work on MT evaluation generally distinguished two major aspects of translation quality, namely adequacy and fluency (White and O'Connell, 1993). However, the field later saw a focus on single-score evaluations that arose both from simpler modeling (Papineni et al., 2002) and from concerns about annotation reliability (Chatzikoumi, 2020).

Error Type	Sentence
Original	*The cat chased the mouse.*
Accuracy	The cat chased the **ball**.
Fluency	The cat the mouse **chased**.
Style	The cat **found itself in pursuit of** the mouse.

Table 1: Different types of translation errors

An MT evaluation framework that takes these distinctions seriously is MQM, the *multidimensional quality metrics* framework (Lommel et al., 2014). It decomposes translation quality into a number of aspects and their subaspects. Errors are identified according to their error types and severity levels and they are converted into numerical scores by their pre-defined weights (see Section 3.1 for details). MQM is a robust scheme that corresponds well to judges' overall assessment of translation quality and provides nuanced insights into the properties of MT output (Freitag et al., 2021a). However, there are few corpora annotated with MQM scores, and correspondingly, little computational work that assesses the automatic prediction of MQM scores for the purposes of MT evaluation.

In this paper, we show that a suitably adapted

LREC-COLING 2024, pages 11723–11744
20-25 May, 2024.

version of MQM lends itself to comparatively easy modeling on top of current neural language models which makes use of three key error dimensions: accuracy, fluency, and style. This multidimensionality is not a huge concern any more, since these models straightforwardly support multi-task learning.

Our study provides three main contributions: First, we present a resource of MQM-annotated dataset for English-Korean translation evaluation, a language pair known to be challenging both for translation and for evaluation (Choi et al., 2018). Second, we train and evaluate models for automatic MQM score prediction. Third, we identify the optimal conditions for our method by scrutinizing an array of models on varied data scales and diverse inputs, showing that robust prediction is possible even with relatively limited training data. The corpus and the model code are publicly available on our GitHub repository at `https://github.com/DojunPark/multidimensional_MTE`.

Plan of the paper. Section 2 introduces related work. Section 3 describes our MQM resource for the the language pair of English–Korean. Section 4 discusses the model architectures and the experimental setup. Section 5 presents our experiments along with their results. Finally, Section 6 discusses our findings and directions for future research.

2. Related Work

2.1. Human Evaluation

Adequacy and fluency (White and O'Connell, 1993) are among the most traditional approaches to human evaluation. Adequacy assesses if translations by MT systems convey the meaning of the source text accurately, while fluency evaluates the naturalness and fluency of the target text, paying particular attention to grammar and idiomatic expressions.

Ranking (Duh, 2008) is another traditional method where translations from different systems are compared and ranked sentence by sentence. This relative ranking method often yields better inter-annotator agreement than evaluations based on adequacy and fluency (Koehn, 2010).

2.2. Automatic Evaluation

String-based Metrics. BLEU (Papineni et al., 2002) is the most common automatic metric for MT quality evaluation. It assesses machine-generated translations by computing n-gram precisions, which focus on precision, and imposing a brevity penalty, which serves to capture the aspect of recall. While being pointed out for its limitations repeatedly (Marie et al., 2021; Chauhan and Daniel, 2022), BLEU still remains widely used in a majority of MT publications (Marie et al., 2021; Chauhan and Daniel, 2022). METEOR (Banerjee and Lavie, 2005) addresses some limitations of BLEU by incorporating stemming and synonymy.

Instead of focusing on the word level n-gram, Character F-score (ChrF) (Popović, 2015) evaluates translations based on character-level n-gram overlaps. An enhanced version, ChrF++ (Popović, 2017), extends the original metric by also considering word-level n-grams in its evaluation and they are often considered as alternative metrics to BLEU.

Translation Error Rate (TER) (Snover et al., 2006) is another automatic metric based on the edit distance such as insertion, deletion, substitution, and shift. HTER (Human-mediated Translation Error Rate) integrates human intervention in the evaluation process by comparing the machine output to a version post-edited by a human translator.

Neural Metrics. Neural metrics represent an advanced approach for automatic evaluation of machine-generated translations. It can be categorized as either unsupervised or supervised (Lee et al., 2023). Unsupervised metrics such as BERTscore (Zhang et al., 2020) and YiSi (Lo, 2019) leverage the contextual embeddings of pre-trained models to measure the semantic similarity between words in candidate and reference translations, enabling a more informative evaluation.

Supervised metrics such as BLEURT (Sellam et al., 2020) and COMET (Rei et al., 2020) go a step further. They fine-tune pre-trained models on manually scored data sets utilizing specific human evaluation metrics, thereby producing assessments that more closely approximate human judgments. Recent studies have confirmed superior performance of supervised neural metrics in correlating with human assessments, outperforming other traditional and unsupervised neural metrics (Mathur et al., 2020; Freitag et al., 2021b).

However, there are potential pitfalls: These metrics are prone to a higher susceptibility to overfitting on the data they are trained on. Also, if the embeddings used in both the MT and evaluation models are too similar, there is a risk of bias, potentially skewing evaluation outcomes.

2.3. Quality Estimation

Quality Estimation (QE) is considered an alternative approach to MT evaluation. As a reference-free evaluation, it assesses the quality of a translation by taking only the source sentence and its machine-generated translation without a reference translation (Specia et al., 2018). This technique is advantageous as it provides an estimation of translation quality without requiring reference translations, which might not always be available.

Major Cat.	Minor Cat.	Description
Accuracy	Addition	Translation includes information not present in the source.
	Omission	Translation is missing content from the source.
	Mistranslation	Translation does not accurately represent the source.
	Untranslated text	Source text has been left untranslated.
Fluency	Punctuation	Incorrect punctuation (for locale or style).
	Spelling	Incorrect spelling or capitalization.
	Grammar	Problems with grammar, other than orthography.
	Register	Wrong grammatical register (eg, inappropriately informal pronouns).
	Inconsistency	Internal inconsistency (not related to terminology).
	Character encoding	Characters are garbled due to incorrect encoding.
Terminology	Inappropriate for context	Terminology is non-standard or does not fit context.
	Inconsistent use	Terminology is used inconsistently.
Style	Awkward	Translation has stylistic problems.
Locale convention	Address format	Wrong format for addresses.
	Currency format	Wrong format for currency.
	Date format	Wrong format for dates.
	Name format	Wrong format for names.
	Telephone format	Wrong format for telephone numbers.
	Time format	Wrong format for time expressions.
Other		Any other issues.
Source error		An error in the source.
Non-translation		Impossible to reliably characterize distinct errors.

Table 2: MQM hierarchy from Freitag et al. (2021a)

QuEst (Specia et al., 2013) and QuEst++ (Specia et al., 2015) are statistical QE systems which operate by leveraging linguistic features from both the source sentence and its translation and learning to predict translation quality based on these features.

Following the general trend, neural methods have recently become central in QE systems. OpenKiwi (Kepler et al., 2019) incorporates leading QE systems from the WMT 2015–18 tasks into its framework. TransQuest(Ranasinghe et al., 2020), building on XLM-RoBERTa embeddings (Conneau et al., 2019), has surpassed other state-of-the-art systems (Specia et al., 2020). Most recently, COMETKiwi (Rei et al., 2022b) achieved a new state-of-the-art in the WMT 2022 shared task on QE (Zerva et al., 2022). It has brought QE performance forward by integrating the COMET framework (Rei et al., 2020) with the predictor-estimator architecture (Kim et al., 2017), where a predictor processes both source and target sentences to predict target words, and an estimator uses these feature vectors to estimate translation quality. This advancement has further blurred the lines between traditional MT evaluation and Quality Estimation.

3. An MQM Resource for the Language Pair English–Korean

3.1. Multidimensional Quality Metrics

MQM or Multidimensional Quality Metrics (Lommel et al., 2014) is a framework specifically designed to identify translation issues and transform them into quantifiable scores. It features a hierarchical structure of translation error types as shown in Table 2. In addition, MQM proposes to weigh each error by its severity. It offers three weights: minor error (weight 1), major error (weight 5), and critical error (weight 25). At the segment level, we can aggregate errors to obtain first category-specific scores, and then aggregate category scores to obtain overall scores.

MQM sees itself as a general framework whose users should select the most pertinent error categories and severities for each particular translation context. We now describe our use case and then proceed to describing our adaptations to MQM.

3.2. Constructing an English–Korean Parallel Dataset

To construct our dataset, we start with parallel corpora, then generate translations and quality assessments using a two-step approach: paraphrasing and quality evaluation.

From the OPUS (Open Parallel Corpus) project (Tiedemann, 2012), we chose two English-Korean parallel corpora to capture diverse linguistic styles: Global Voices (Tiedemann, 2012) and TED Talks 2020 (Reimers and Gurevych, 2020). Global Voices consists of news articles crawled from its website in 46 languages, while TED Talks 2020 features around 4,000 transcripts from TED and TED-X events as of July 2020, covering 108 languages. From each, we randomly sampled 600 translation pairs, summing up to 1200. Out of these, 1000 pairs form our training set, with the remaining

	Global Voices	TED Talks 2020
Genre	News	Presentation Transcript
Style	Formal	Conversational
Total Pairs	9,381	399,413
Sampled Pairs	600	600
Avg. Len. (src.)	17.74	15.99
Avg. Len. (ref.)	12.90	10.98
Avg. Len. (tgt.)	13.02	11.23

Table 3: Corpus summary statistics

200 equally divided as validation and test set. This makes our dataset comparable in size to COMET, which employed 1000 translation pairs for each language pair (Rei et al., 2020).

Paraphrasing. Parallel corpora are often used as gold standard in translation. When such corpora only contain a single manual translation, however, only capture a small part of the space of translations, and towards correctness. We believe that we can enhance the robustness of our evaluation (and evaluation models) by considering a larger sample of potential translations, including somewhat flawed ones which exhibit a range of natural errors.

To do so, we automatically paraphrase our parallel corpus. We opt for proprietary online services for both paraphrasing the English source sentences and subsequently translating the paraphrased content into Korean, given their notable superiority over open-source alternatives. We obtained paraphrases from ChatGPT gpt-3.5-turbo[1] using the following prompt:

```
Please rewrite the given sentence in
English while maintaining the same
meaning, using different vocabulary
or sentence structures:
[English sentence]
```

The placeholder “[English sentence]” is replaced with the actual sentence for paraphrasing, then translated into Korean using Google Translate.

Table 3 shows statistics of the selected parallel corpora and the generated sentences. Global Voices is news-based and formal, while TED Talks 2020 contains conversational presentation transcripts. Notably, the generated Korean sentences are slightly longer than their corresponding reference sentences. The increase in sentence length suggests that the paraphrasing step introduces structural and stylistic variation, which could impact the quality of translation.

3.3. Annotation of Annotation Quality

For our annotation, we make the following adaptations to the general MQM framework:

- Error Dimensions: We select three major error dimensions of accuracy, fluency and style. We leave out the major category of terminology since our selected corpora comprise general texts where terminology is not a major factor. We also exclude the dimensions of audience appropriateness and design and markup since our corpora neither target a specific audience nor include graphical presentations.
- Sub-error Type Integration: Given the infrequency of formatting issues in our evaluation, we chose to integrate them as a sub-error type under the fluency dimension, deeming it unnecessary to maintain them as a primary error dimensions. Furthermore, since untranslated text — initially a sub-error type under accuracy — impedes both the original meaning and the readability, we include it under both the accuracy and fluency dimensions.
- Error Severity Classification: We distinguish two severity levels: major errors and minor errors. We omit the critical error category due to its inherent subjectivity, as supported by extant literature (Freitag et al., 2021a).

We calculate the scores for the three dimensions accuracy S_a, fluency S_f, and style S_s as follows:

$$S_a = 5 \times E_{a,m} + 1 \times E_{a,i}, \quad (1)$$

$$S_f = 5 \times E_{f,m} + 1 \times E_{f,i}, \quad (2)$$

$$S_s = 5 \times E_{s,m} + 1 \times E_{s,i}, \quad (3)$$

where each score is the sum of the products of their respective major errors (weighted by 5) and minor errors (weighted by 1). Specifically, for each dimension d (accuracy, fluency, style), the score S_d is determined by the major $E_{d,m}$ and minor $E_{d,i}$ errors of that dimension. The total score, S_{total}, is:

$$S_{\text{total}} = S_a + S_f + S_s, \quad (4)$$

We created a set of guidelines for our EN-to-KO translation evaluation based on the MQM guidelines[2], given in Appendix D. The annotation was carried out primarily by an evaluator proficient in English at a level above CEFR C1 [3] and native in Korean with a background in computational linguistics. The five most frequent error types annotated

[1] https://platform.openai.com/docs/models

[2] https://themqm.org/

[3] https://www.coe.int/en/web/common-european-framework-reference-languages/table-1-cefr-3.3-common-reference-levels-global-scale

Accuracy	Fluency	Style
0.54	0.57	0.34

Table 4: Correlations (Kendall's τ) between annotators by dimension.

	Accuracy	Fluency	Style	BLEU
Accuracy	1			0.17***
Fluency	0.29***	1		0.15***
Style	0.10***	0.01	1	0.08***

Table 5: Correlations (Kendall's τ) among MQM dimensions and between MQM dimensions and BLEU score (stars indicate statistical significance).

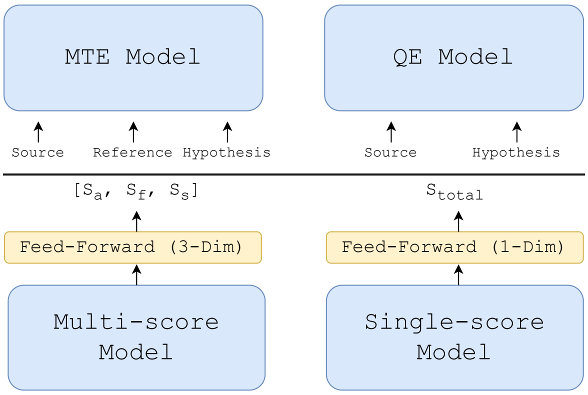

Figure 1: Top: Input configurations for MTE and QE Models. Bottom: Output layer setups for multi-score and single-score models

are: mistranslation; unnaturalness; structure; untranslated text; omission. Details on error types and score distributions are given in Appendix B.

The annotation was timed at approximately 5 min/unit. While this may sound long, this is comparable to a previous annotation study by (Mariana et al., 2015) who report an annotation speed of around 10 min/unit, albeit for longer sentences.

3.4. Validation of MQM Scores

We can now ask two questions: (a), are the MQM scores that we have obtained *reliable*? (b), do they provide us with *additional* information compared to single-score metrics, as we claimed above?

To address point (a), we employ cross-validation with two independent annotators, namely two undergraduates who satisfied the same prerequisites as our primary annotator. They independently annotated a subset (100 translation units) of our primary data using the guidelines we established. We use Kendall's Tau correlation as our evaluation metric (see Section 4.4 for details).

Table 4 shows the correlation between our primary dataset scores and average scores from two cross-validators. Fluency correlates highest at 0.57, with accuracy at 0.54, and style at 0.34, suggesting that evaluating translation style may be more subjective. Yet, the cross-validation confirms a robust level of agreement with our primary evaluation.

Regarding point (b), Table 5 shows that there are some correlations among the three MQM dimensions, but they are sufficiently mild to warrant the conclusion that the scores indeed measure distinct aspects of quality. While the correlations for style are generally low, we find a correlation of about 0.3 between accuracy and fluency, which we take to reflect a general cline between 'bad' and 'good' translations and which is presumably also related to the presence of untranslated text, the 4-th most frequent error category. This interpretation is further supported by the pattern of correlations between MQM scores and (inverse) BLEU scores we obtain from the SacreBLEU (Post, 2018) implementation, shown in the right column: BLEU correlates to a similar extent with accuracy and fluency, but the correlation overall remains weak (0.15-0.20). We take this to mean that the MQM scores offer a fine-grained evaluation, capturing nuances potentially missed by single-score metrics such as BLEU.

4. Modeling Multidimensional Translation Quality

We carry out a series of experiments in predicting the manual MQM translation quality judgments with various language models: Our main experiment (Experiment 1) assesses the performance of a set of selected pre-trained models for multi-score prediction. Experiment 2 assesses the impact of training data size on model performance. Experiment 3 compares multi-score and single-score prediction of overall translation quality, and Experiment 4 compares our models against the COMET model family.

4.1. Base Model Selection

We compare six Transformer-based pre-trained LMs with multilingual capabilities that cover English and Korean.[4] We consider both encoder-only and encoder-decoder models. In the latter case, we only use the encoder part which, given its extensive pre-training, can potentially function similarly to an encoder-only model in creating inputs for MT evaluation, a regression task (Kocmi et al., 2022). Concretely, we use Multilingual BERT base cased (Devlin et al., 2019), XLM-RoBERTa large (Conneau et al., 2020), and RemBERT (Chung et al., 2021) (encoder-only) and Multilingual BART large-50 (Liu et al., 2020), Multilingual T5 large (Xue

[4] We use the Huggingface `transformers` library.

et al., 2021) and M2M100 1.2B (Fan et al., 2021) (encoder-decoder). For details see Appendix A.

4.2. Input: With vs. Without Reference

The top part of Figure 1 shows the input configurations for our two main setups, namely MTE (Machine Translation Evaluation) models that take the source, reference, and hypothesis sentences; and QE (Quality Estimation) models that only take the source and hypothesis, but no reference. In either case, the input is structured using special tokens similar to BERT's use of `[CLS]` and `[SEP]`. See Table 10 (Appendix A) for details.

4.3. Output: Single- vs. Multi-Score Prediction

The bottom part of Figure 1 shows that we further distinguish between a multi-score setup (predicting several scores) and single-score setup (predicting one score). For both model types, we extract embeddings of the initial token from the output of base model taking advantage of the input encoding described in the previous paragraph. The embedding of the initial token is passed to a simple feed-forward regression layer. For multi-score models, we predict a vector of length three, representing scores for *accuracy*, *fluency*, and *style* of the translation. Single-score models output a single scalar value representing the *overall* quality of the translation.

4.4. Experimental Setup

Dataset. We use our created MQM-annotated dataset for the English-Korean language pair, consisting of 1,200 translation units (cf. Section 3). The dataset is split into training (1,000 units), validation (100 units), and test (100 units) sets, each with an equal distribution between the Global Voices and Ted Talks 2020 corpora. In Experiment 2, training data was reduced to 200, 400, 600, and 800 units respectively, maintaining the corpus distribution.

Training Regimen. All our models are regression models, trained to minimize a mean squared error objective. Optimization is carried out with stochastic gradient descent, using the AdamW algorithm (Loshchilov and Hutter, 2019). Our learning rate is 2e-6 for all models except mT5, which is adjusted to 2e-5 to accelerate its relatively slower training compared to the others. We train for 100 epochs with a batch size of 8. To ensure robustness in our evaluation, we train each model across three separate runs and report averages.

Evaluation Metric. For evaluation, we use Kendall's Tau. It offers a more robust measure of general rank correlation compared to Pearson, which can be sensitive to outliers. Compared to Spearman, it is more sensitive to changes between ranks providing a clearer interpretation of pairs that are in agreement (concordant) and those that are not (discordant). This choice aligns with the WMT Metrics Shared Task, which have adopted Kendall's Tau for evaluating segment-level MQM scores since 2021 (Freitag et al., 2021b, 2022):

$$\tau = \frac{\#\text{concordant pairs} - \#\text{discordant pairs}}{\#\text{concordant pairs} + \#\text{discordant pairs}} \quad (5)$$

A τ value of 1 indicates perfect agreement, -1 complete disagreement, and 0 signifies no correlation.

5. Results and Discussion

5.1. Experiment 1: Model Comparison

Table 6 shows the results for our main experiment, evaluating a range of language models in both an MLE and a QE setting (cf. Section 4.4). Overall, RemBERT stands out in both MTE and QE setups: RemBERT (QE) achieves the highest performance overall with an average score of 0.37 and leads in two error dimensions: fluency and style. Following closely, RemBERT (MTE) yields an average of 0.35 and shows the top performance in the accuracy dimension. XLM-R (QE) is another strong performer, with an average score of 0.33. In contrast, mBART (MTE), XLM-R (MTE) and mBERT (QE) record the lowest average score with 0.24.

Results by Dimension. Three of the five best models for accuracy use the MTE setup. This is to be expected, given that accuracy concerns the relationship between input and the output of the MT process. Rather, it is surprising that the best reference-free model in the QE setting, M2M100 with a score of 0.36, is not far removed from the best score of any model (0.40). In contrast, we see the best results for both fluency and style in the reference-free QE setting. For fluency, this might be expected since this is primarily a target language property. It is again surprising for style, though, in particular given that style shows the largest margin for QE among the three dimensions. This indicates that the model may currently learn how well the MT output matches the "typical" style of our corpus, rather than an actual match between the styles of input and output.

Encoder-only Models vs. Encoder Component Models. While the top spots are dominated by encoder-only models, they are followed by four models based on the encoder components of

		MTE (with reference)				QE (reference-free)			
	Model	**Accuracy**	**Fluency**	**Style**	**Overall**	**Accuracy**	**Fluency**	**Style**	**Overall**
Enc. only	mBERT	0.36	0.34	0.12	0.27	0.31	0.24	0.18	0.24
	XLM-R	0.32	0.32	0.09	0.24	0.35	0.37	0.27	0.33
	RemBERT	**0.40**	0.38	0.26	0.35	0.35	**0.43**	**0.33**	**0.37**
Enc. of Enc.-Dec.	mBART	0.25	0.32	0.14	0.24	0.24	0.41	0.18	0.27
	mT5	0.36	0.40	0.14	0.30	0.35	0.33	0.14	0.27
	M2M100	0.34	0.39	0.11	0.28	0.36	0.36	0.18	0.29

Table 6: Experiment 1: Kendall's Tau between model predictions and human scores across models and evaluation settings. Best result per dimension bolded.

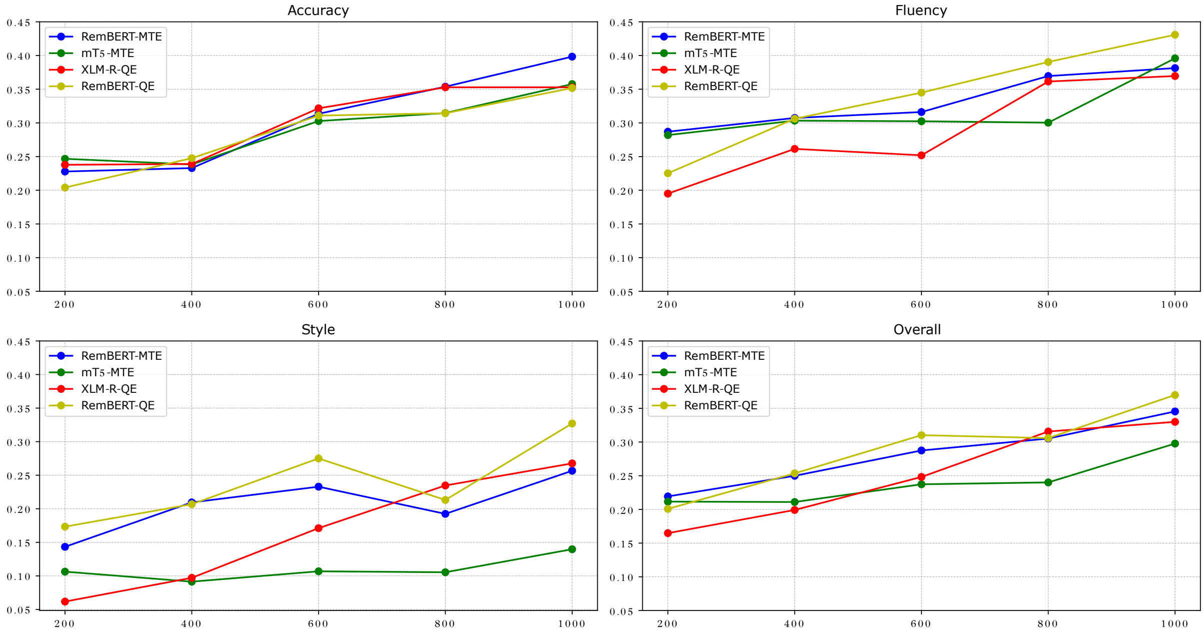

Figure 2: Experiment 2: Kendall's Tau for different amounts of training data (Accuracy, Fluency, Style, and Overall scores).

encoder-decoder models: mT5 (MTE), M2M100 (both MTE and QE) and mBART (QE) record average scores of 0.30, 0.29, 0.28, and 0.28, respectively, outperforming three more encoder-only models. This is particularly true for the accuracy dimension, where M2M100 (QE) and mT5 (MTE) are close to the best model, RemBERT (MTE), and for the fluency dimension, where mBART (QE) and mT5 (MTE) are close to RemBERT (QE). In contrast, the encoder component-based models generally rank lower in the style dimension. A potential explanation derives from their distinct training methodologies. Encoder-only models, designed to predict masked tokens, excel at capturing nuances of word meaning, while the encoders of encoder-decoder models, optimized to create a representation of the essence of the source text, might overlook some nuances of style (Lin et al., 2021).

Error Analysis. We focus on the two top-performing models, RemBERT (MTE) and RemBERT (QE). We observe in general that predicted scores align closely with human assessments when a single type of error is present and sentences are short. In *mistranslation*, the most commonly identified error, the content domain of the text and the depth of contextual understanding required play significant roles in prediction accuracy. Models struggle terms specific to certain domains such as *color revolution* or *blogosphere*. Regarding *omission* errors, sentences that leave out frequent terms like *if* and *the* seem to align more closely with human assessments, while deictic expressions such as *below* are not only challenging to translate but also to evaluate. We also found that quality was often overestimated as a result of minor errors especially when *untranslated* text was involved – our models are evidently not always able to detect subtle discrepancies. See Appendix B and C for details.

	Model	Accuracy	Fluency	Style	Overall
MTE	RemBERT	0.40	0.38	0.26	0.35
	COMET-22	**0.30**	0.10	0.02	0.28
QE	RemBERT	0.35	0.43	0.33	**0.37**
	CometKiwi	**0.40**	0.09	0.17	**0.37**

Table 7: Experiment 4: Kendall's Tau correlations of different models (COMET and CometWiki, our best RemBERT-based multi-score model) against different translation quality dimensions on the test set.

Model	Single	Multi	Δ Score
RemBERT (MTE)	0.39	0.42	+0.03
mT5 (MTE)	0.38	0.37	-0.01
XLM-R (QE)	0.35	0.38	+0.03
RemBERT (QE)	0.39	0.41	+0.02

Table 8: Experiment 3: Correlations (Kendall's τ) for multi-score and single-score models of overall translation quality

5.2. Experiment 2: Impact of Training Data Size on Model Performance

Experiment 2 varies the amounts of training data between 200 and 1000 training data points. Figure 2 shows the performance of the four best models, per dimension and averaged. The rightmost results (for 1000 training datapoints) correspond to the values in Table 6. We see that even though the models obtain reasonable level of performance for 1000 training datapoints, there still appears to be a fairly linear improvement with increasing data size, especially in the accuracy and fluency dimensions, indicating that further improvements are possible with more training data. Overall, the four models behave comparatively similar to one another.

5.3. Experiment 3: Predicting Overall Translation Quality

Experiment 3 focuses on the prediction of overall translation quality. It compares the results obtained when predicting overall translation quality directly ("single-score") vs. predicting the three individual quality dimensions ("multi-task") first and then accumulating them to obtain the overall quality. Table 8 shows the results, again for the four best models. For three of our four models, the multi-task setting improves over the single-task setting, and the positive effect in the three cases is more substantial than the single negative case. These findings highlight the benefits of multi-score models in that they not only offer fine-grained performance insights but also have the potential of outperforming simpler models — even if one is ultimately only interested in a single overall quality assessment.

5.4. Experiment 4: Comparison against COMET

While our models, trained as multi-output regressors, diverge from the conventional approach of directly predicting overall quality scores, we conducted a comparative analysis, contrasting our standout RemBERT models with COMET-22 and CometKiwi, which are recognized as state-of-the-art evaluators within the MTE and QE frameworks, respectively. To ensure a fair comparison, we derive overall quality predictions from COMET-22 and CometKiwi and evaluate them as before, using Tau correlation, against both overall quality scores and the MQM-derived specific quality dimensions. The results are detailed in Table 7, alongside the findings from our RemBERT models from above.

A notable observation from our analysis is that COMET's scores, particularly COMET-22, are skewed towards accuracy. This contrasts with our multi-output regressor models, which demonstrate a more balanced correlation for accuracy and fluency, with style slightly behind. This disparity highlights that our approach to MT evaluation as multi-task learning can provide a reliable evaluation by integrating diverse aspects without losing focus on individual dimensions.

Both COMET-22 and CometKiwi show a notably lower correlation for fluency, approximately around 0.10, in contrast to our models. While COMET-22's correlation for style diminishes further from its fluency score, CometKiwi demonstrates an enhanced correlation with human-evaluated scores for style. This suggests CometKiwi might better capture the nuances of style, aligning with our findings in Experiment 1, where QE models generally outperformed MTE models in evaluating style. Such an observation leads us to speculate about the underlying mechanisms of QE models' performance suggesting that QE models may act similarly to language models in distinguishing between styles.

Finally, CometKiwi matches the performance of RemBERT (QE) when predicting overall translation quality. This observation is noteworthy since CometKiwi operates in a zero-shot setting for Korean, without specific training on Korean MT data. Both CometKiwi and RemBERT have undergone Korean pre-training. While CometKiwi was extensively fine-tuned on a vast MT multlingual evaluation dataset (657k sentence pairs), our RemBERT-based multi-score prediction model was fine-tuned on 1k Korean sentences. This demonstrates an interesting trade-off between focused language-specific and broad language-agnostic tuning.

6. Discussion and Conclusions

This paper tackles a bottleneck for using the Multidimensional Quality Metric (MQM) framework for automatic fine-grained MT, namely the scarcity of suitably annotated resources. We introduce a benchmark dataset for English–Korean, demonstrating the feasibility of an annotation setup for three major error categories (accuracy, fluency, style) and show its utility to train automatic models for MQM-based three-dimensional MT evaluation models, reframing MT evaluation as multi-task learning.

Empirically, we find that RemBERT consistently emerges as the standout among other SOTA models in both MTE (reference-based) and QE (reference-free) setups.Our finding that reference-free models outperform their counterparts in the style dimension, while reference-based models excel in the accuracy dimension is in line with earlier findings that the evaluation of style is more influenced by fidelity to the original source than to the reference. In contrast, reference translations still hold importance in determining factual correctness and completeness (Fomicheva et al., 2020; Sun et al., 2020). Furthermore, we demonstrate that the multi-dimensional evaluation approach not only enhances the interpretability of MT evaluation systems at the level of fine-grained dimensions but also rivals or even surpasses single-score models in terms of overall quality score. Our approach is also competitive with the MT evaluation metrics of the Comet family, arguably striking a better balance between different dimensions of translation quality.

Multi-dimensional evaluation is a strong desideratum for MT practitioners to navigate trade-offs in translation (Alves and Jakobsen, 2020; Lim et al., 2024). For instance, scientific reports prioritize adequacy, whereas literary works like novels or poems require more fluent and idiomatic expressions, sometimes at the expense of literal adequacy. By giving users access to evaluation at these different dimensions, our approach yields enhanced explainability of translation quality assessments and potentially contributes to better user acceptance.

Limitations

A major limitation of our study is that we only consider a single language pair, namely English-Korean. At the same time, this language pair is known to be challenging due to the typological differences between the two languages (Hong et al., 2005; Choi et al., 2018). Therefore, we take the good results we obtain regarding annotation reliability and regarding automatic prediction quality to be promising, also with respect to generalizing our approach to other language pairs.

While the weighting of individual dimensions in determining an overall quality score may vary based on the specific translation objectives, we adopted a straightforward method employing averages for general-purpose evaluation, primarily to establish a simple evaluation schema comparable to other MT evaluation setups (Rei et al., 2022a,b). Our method however straightforwardly supports adjusting the weights to reflect the relative significance of different translation quality aspects for specific contexts of translation.

Another direction that we did not explore is the use of cross-lingual transfer learning methods to address the need for manual annotation. Future work can build on studies leveraging existing MT evaluation datasets from other languages (Freitag et al., 2021a) to address this challenge, extending them from the overall quality case to the multidimensional MQM case.

Ethics Statement

In this study, we introduce a benchmark dataset for English-Korean translation evaluation. We have ensured that all data utilized is publicly available and does not contain any personal or identifiable information. We involved several annotators in our primary annotation and cross-validation process. All were informed about the purpose of the research and the methods employed. The dataset and the model implementation are publicly available.

We believe that our fine-grained automatic MT evaluation holds the potential to enhance trust in neural MT systems by presenting results in a more interpretable manner. As research, including our own, continues to improve the interpretability of these neural systems, we believe that such detailed evaluations can better position them to handle translations that are sensitive to cultural and societal contexts.

7. References

Fábio Alves and Arnt Lykke Jakobsen. 2020. *The Routledge Handbook of Translation and Cognition*. Routledge.

Eleftherios Avramidis, Vivien Macketanz, Arle Lommel, and Hans Uszkoreit. 2018. Fine-grained evaluation of quality estimation for machine translation based on a linguistically motivated test suite. In *Proceedings of the AMTA 2018 Workshop on Translation Quality Estimation and Automatic Post-Editing*, pages 243–248, Boston, MA. Association for Machine Translation in the Americas.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of the International Conference on Learning Representations*, San Diego, CA.

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.

Eirini Chatzikoumi. 2020. How to evaluate machine translation: A review of automated and human metrics. *Natural Language Engineering*, 26(2):137–161.

Shweta Singh Chauhan and Philemon Daniel. 2022. A comprehensive survey on various fully automatic machine translation evaluation metrics. *Neural Processing Letters*, 55:12663–12717.

Sung-Kwon Choi, Gyu-Hyeun Choi, and Youngkil Kim. 2018. Automatic evaluation of English-to-Korean and Korean-to-English neural machine translation systems by linguistic test points. In *Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation*, Hong Kong. Association for Computational Linguistics.

Hyung Won Chung, Thibault Févry, Henry Tsai, Melvin Johnson, and Sebastian Ruder. 2021. Rethinking embedding coupling in pre-trained language models. In *Proceedings of the 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. 2019. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

B. Dorr, Joseph P. Olive, John McCary, Caitlin Christianson, Matthew G. Snover, and Nitin Madnani. 2011. Part 5: Machine translation evaluation and optimization. In *Handbook of Natural Language Processing and Machine Translation*. Springer.

Kevin Duh. 2008. Ranking vs. regression in machine translation evaluation. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 191–194, Columbus, Ohio. Association for Computational Linguistics.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2021. Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research*, 22(1).

Marina Fomicheva, Shuo Sun, Lisa Yankovskaya, Frédéric Blain, Francisco Guzmán, Mark Fishel, Nikolaos Aletras, Vishrav Chaudhary, and Lucia Specia. 2020. Unsupervised Quality Estimation for Neural Machine Translation. *Transactions of the Association for Computational Linguistics*, 8:539–555.

Markus Freitag, George F. 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, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André F. T. Martins. 2022. Results of WMT22 metrics shared task: Stop using BLEU – neural metrics are better and more robust. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 46–68, Abu Dhabi, United Arab

Emirates (Hybrid). Association for Computational Linguistics.

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.

Munpyo Hong, Young-Gil Kim, Chang-Hyun Kim, Seong-Il Yang, Young-Ae Seo, Cheol Ryu, and Sang-Kyu Park. 2005. Customizing a Korean-English MT system for patent translation. In *Proceedings of Machine Translation Summit X: Papers*, pages 181–187, Phuket, Thailand.

Fabio Kepler, Jonay Trénous, Marcos Treviso, Miguel Vera, and André F. T. Martins. 2019. OpenKiwi: An open source framework for quality estimation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 117–122, Florence, Italy. Association for Computational Linguistics.

Hyun Kim, Jong-Hyeok Lee, and Seung-Hoon Na. 2017. Predictor-estimator using multilevel task learning with stack propagation for neural quality estimation. In *Proceedings of the Second Conference on Machine Translation*, pages 562–568, Copenhagen, Denmark. Association for Computational Linguistics.

Filip Klubicka, Antonio Toral, and Víctor M. Sánchez-Cartagena. 2018. Quantitative fine-grained human evaluation of machine translation systems: a case study on english to croatian. *Machine Translation*, 32:195–215.

Tom Kocmi, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Thamme Gowda, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Rebecca Knowles, Philipp Koehn, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Michal Novák, Martin Popel, Maja Popović, and Mariya Shmatova. 2022. Findings of the 2022 conference on machine translation (WMT22). In *Proceedings of the Seventh Conference on Machine Translation*, pages 1–45, Abu Dhabi. Association for Computational Linguistics.

Philipp Koehn. 2010. *Statistical Machine Translation*. Cambridge University Press.

Seungjun Lee, Jungseob Lee, Hyeonseok Moon, Chanjun Park, Jaehyung Seo, Sugyeong Eo, Seonmin Koo, and Heuiseok Lim. 2023. A survey on evaluation metrics for machine translation. *Mathematics*, 11(4).

Zheng Wei Lim, Ekaterina Vylomova, Trevor Cohn, and Charles Kemp. 2024. Simpson's paradox and the accuracy-fluency tradeoff in translation. *arXiv:2402.12690*.

Tianyang Lin, Yuxin Wang, Xiangyang Liu, and Xipeng Qiu. 2021. A survey of transformers. *AI Open*, 3:111–132.

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

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, Florence, Italy. Association for Computational Linguistics.

Arle Lommel, Hans Uszkoreit, and Aljoscha Burchardt. 2014. Multidimensional quality metrics (mqm): A framework for declaring and describing translation quality metrics. *Tradumàtica*, 12.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *Proceedings of the International Conference on Learning Representations*.

Valerie Mariana, Troy Cox, and Alan Melby. 2015. The multidimensional quality metric (mqm) framework: A new framework for translation quality assessment. *The Journal of Specialised Translation*, 23.

Benjamin Marie, Atsushi Fujita, and Raphael Rubino. 2021. Scientific credibility of machine translation research: A meta-evaluation of 769 papers. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7297–7306, Online. Association for Computational Linguistics.

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.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

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, Brussels, Belgium. Association for Computational Linguistics.

Tharindu Ranasinghe, Constantin Orasan, and Ruslan Mitkov. 2020. TransQuest: Translation quality estimation with cross-lingual transformers. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5070–5081, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022a. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). 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.

Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. 2022b. CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid). 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, 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.

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

Lucia Specia, Gustavo Paetzold, and Carolina Scarton. 2015. Multi-level translation quality prediction with QuEst++. In *Proceedings of ACL-IJCNLP 2015 System Demonstrations*, pages 115–120, Beijing, China. Association for Computational Linguistics and The Asian Federation of Natural Language Processing.

Lucia Specia, Carolina Scarton, and Gustavo Henrique Paetzold. 2018. *Quality Estimation for Machine Translation*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

Lucia Specia, Kashif Shah, Jose G.C. 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, Sofia, Bulgaria. Association for Computational Linguistics.

Shuo Sun, Francisco Guzmán, and Lucia Specia. 2020. Are we estimating or guesstimating translation quality? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6262–6267, Online. Association for Computational Linguistics.

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).

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of Advances in Neural Information Processing Systems*, pages 5998–6008.

John S. White and Theresa A. O'Connell. 1993. Evaluation of machine translation. In *Proceedings of the Human Language Technology Workshop*, Plainsboro, NJ.

Yonghui Wu, Mike Schuster, Z. Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason R. Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Gregory S. Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv:1609.08144*.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mt5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

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

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

8. Language Resource References

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4512–4525, Online. Association for Computational Linguistics.

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).

A. Model Details

Model	**Params**	**Blocks**	**Heads**	**Emb Size**	**FFN Dim**	**Max Len**
mBERT	178M	12	12	768	3072	512
XLM-R	660M	24	16	1024	4096	514
RemBERT	576M	32	18	1152	4608	512
mBART (Enc)	408M	12	16	1024	4096	1024
mT5 (Enc)	564M	24	16	1024	2816	1024
M2M100 (Enc)	635M	24	16	1024	8192	1024

Table 9: Architectural specifics of the selected models in terms of model size, depth, and inherent complexity.

Setup	Model Type	Token Format
MTE	Encoder Models	[CLS] *source* [SEP] *reference* [SEP] *hypothesis* [SEP]
	mBART	en_XX *source* ko_KR *reference* ko_KR *hypothesis* </s>
	mT5	<init> *source* <sep> *reference* <sep> *hypothesis* </s>
	M2M100	__en__ *source* __ko__ *reference* __ko__ *hypothesis* </s>
QE	Encoder Models	[CLS] *source* [SEP] *hypothesis* [SEP]
	mBART	en_XX *source* ko_KR *hypothesis* </s>
	mT5	<init> *source* <sep> *hypothesis* </s>
	M2M100	__en__ *source* __ko__ *hypothesis* </s>

Table 10: Token Formats for Different Models Across Tasks.

B. Score and Error Distribution of the MQM Dataset

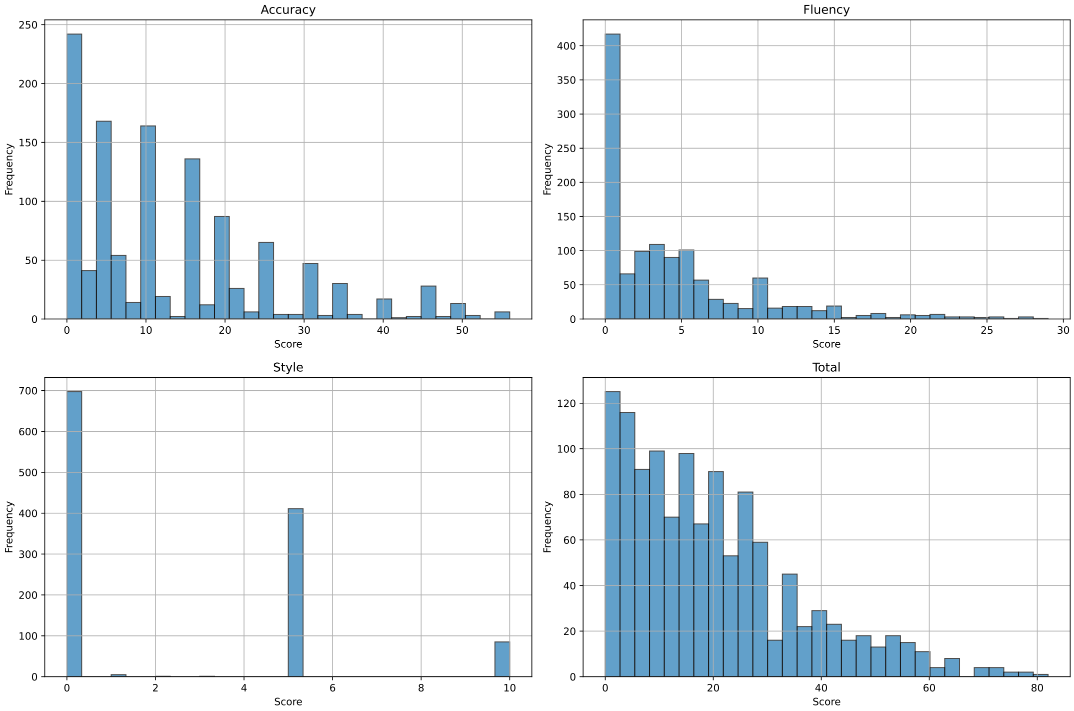

Figure 3: MQM Score Distribution by Dimension.

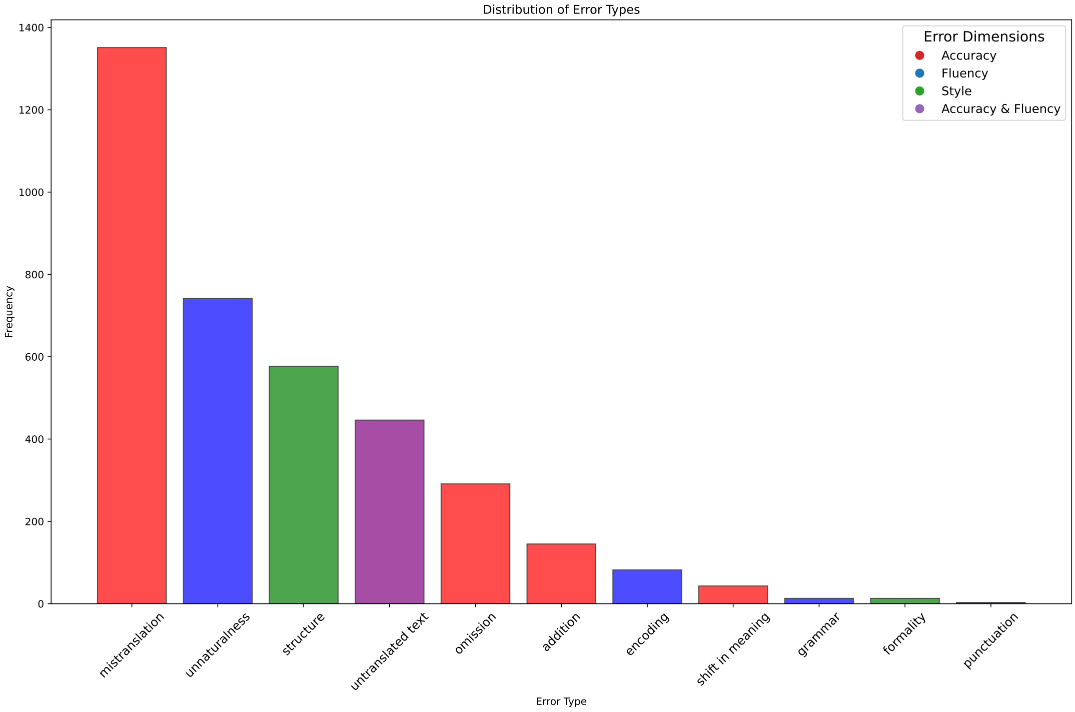

Figure 4: MQM Error Distribution by Dimension.

C. Error Cases along with the Scores Predicted by our MTE/QE Models

Source Text (EN)	Target Text (KO)	Annotation	Golden Score	RemBERT (MTE)	RemBERT (QE)
His report matches this photograph shared by Kareem Fahim on Twitter, which shows Morsi supporters, carrying sticks and shields, and wearing helmets.	그가 제시한 보고서는 Kareem Fahim이 Twitter에서 공유한 사진과 일치합니다. 이미지는 헬멧을 쓰고 방패와 막대기를 들고 있는 무르시 지지자들을 묘사합니다.	Accuracy: Kareem Fahim이 Twitter (untranslated text/major) Fluency: Kareem Fahim이 Twitter (untranslated text/major) Style: -	30	**34.16**	**30.04**
And when I got myself together and I looked at her, I realized, this isn't about me.	정신을 차리고 그녀를 관찰한 결과 상황이 내 중심으로 돌아가는 것이 아니라는 사실을 깨달았다.	Accuracy: 관찰한 (mistranslation/minor), 상황이 내 중심으로 돌아가는 (mistranslation/major) Fluency: - Style: -	21	**21.14**	**18.83**

Table 11: Cases of overall score predictions; it demonstrates that predicted scores align closely with golden scores when only a single type of error is present.

Source Text (EN)	Target Text (KO)	Annotation	Golden Score	RemBERT (MTE)	RemBERT (QE)
It's a coming of age for bringing data into the humanitarian world.	인도주의 분야는 데이터 통합의 중추적 순간을 경험하고 있습니다.	Accuracy: 분야는 (mistranslation/major), 중추적 순간을 (mistranslation/major) Fluency: - Style: 인도주의 분야는 .. 경험하고 있습니다. (structure/major)	20	**21.77**	**22.03**
We have babies.	우리 가족에게는 신생아가 있습니다.	Accuracy: 가족에게는 (addition/major) Fluency: - Style: 우리 가족에게는 신생아가 있습니다. (structure/major)	10	**10.90**	**8.49**

Table 12: Cases of overall score predictions; it showcases that predictions match closely with golden scores, especially in cases where sentences are notably short.

Source Text (EN)	**Target Text (KO)**	**Annotation**	**Golden Score**	**RemBERT (MTE)**	**RemBERT (QE)**
The occupy protests are not a color revolution.	점거 시위는 색깔에 기반한 혁명을 구성하지 않습니다.	Accuracy: 색깔에 기반한 (mistranslation/minor), 구성하지 않습니다. (mistranslation/major) Fluency: 색깔에 기반한 혁명을 구성하지 않습니다. (unnaturalness/minor) Style: -	12	7.77	6.08
Does Guinea even have a blogosphere to speak of?	기니에 중요한 블로깅 커뮤니티가 있습니까?	Accuracy: 중요한 (addtion/major), 블로깅 (mistranslation/minor), to speak of (omission/major) Fluency: - Style: -	21	12.48	13.69

Table 13: Cases of accuracy score predictions; models often face challenges in accurately predicting scores when encountering domain-specific terminology like 'color revolution' or 'blogosphere'.

Source Text (EN)	**Target Text (KO)**	**Annotation**	**Golden Score**	**RemBERT (MTE)**	**RemBERT (QE)**
His report matches this photograph shared by Kareem Fahim on Twitter, which shows Morsi supporters, carrying sticks and shields, and wearing helmets.	그가 제시한 보고서는 Kareem Fahim이 Twitter에서 공유한 사진과 일치합니다. 이미지는 헬멧을 쓰고 방패와 막대기를 들고 있는 무르시 지지자들을 묘사합니다.	Accuracy: Kareem Fahim이 Twitter (untranslated text/major) Fluency: Kareem Fahim이 Twitter (untranslated text/major) Style: -	15, 15	**16.68, 14.15**	**13.1, 14.58**
The authorities have declared a state of emergency while the Prime Minister Mohammed Ghannouchi announced on state television that he was taking over as interim President.	모하메드 간누치(Mohammed Ghannouchi) 총리가 국영 TV에서 그가 임시 대통령의 역할을 맡을 것이라고 선언한 것과 함께 비상사태가 관리들에 의해 발표되었습니다.	Accuracy: (Mohammed Ghannouchi) (untranslated text/minor) Fluency: (Mohammed Ghannouchi) (untranslated text/minor) Style: 비상사태가 관리들에 의해 발표되었습니다. (structure/major)	2, 2	12.2, 8.76	12.01, 10.27

Table 14: Cases of accuracy and fluency score predictions; minor errors tend to be overly penalized, especially in untranslated texts (bottom), contrastive to how major errors (top) are assessed.

Source Text (EN)	**Target Text (KO)**	**Annotation**	**Golden Score**	**RemBERT (MTE)**	**RemBERT (QE)**
According to the source, such a reading of the law was supported by mainland Chinese legal experts.	소식통에 따르면 중국 본토의 법률 전문가들은 이러한 법률 해석을 지지한 것으로 알려졌습니다.	Accuracy: the (omission/major) Fluency: - Style: 중국 본토의 법률 전문가들은 이러한 법률 해석을 지지한 것으로 알려졌습니다. (structure/major)	5	**6.03**	**2.94**
If you think about what being a great parent is, what do you want? What makes a great parent?	뛰어난 부모의 자질을 고려하십시오. 귀하의 이상적인 기준은 무엇입니까? 육아에서 위대함을 구성하는 것은 무엇입니까?	Accuracy: If (omission/major), 육아에서 위대함을 구성하는 (mistranslation/major) Fluency: 고려하십시오. (unnaturalness/minor), 육아에서 위대함을 구성하는 것은 (unnaturalness/minor) Style: 뛰어난 부모의 자질을 고려하십시오. (structure/major), 육아에서 위대함을 구성하는 것은 무엇입니까? (structure/major)	20	**20.53**	**26.94**
Below is a brief explanation of the landmarks with photos taken by Au Kalun, a former journalist and a famous blogger.	저명한 블로거이자 전 저널리스트인 Au Kalun은 주목할만한 랜드마크에 대한 간결한 설명과 함께 시각 자료를 제공했습니다.	Accuracy: Below (omission/major), 주목할만한 (addition/major) Fluency: Au Kalun (untranslated text /major) Style: 시각 자료를 제공했습니다. (structure/major)	10	22.94	18.56

Table 15: Cases of accuracy score predictions; predictions tend to match closely with the golden scores when frequent words such as ‘the’ or ‘if’ (top) are excluded, unlike the case with ‘below’, a deictic term (bottom).

D. MQM Annotation Guidelines for English-Korean Translation

D.1. Introduction

These guidelines have been created to streamline the evaluation process of English-to-Korean translation quality, aligning with the Multidimensional Quality Metrics (MQM) framework. The MQM framework provides a robust structure enabling evaluators to judge the quality of translation under uniform and clear-cut criteria, thereby converting the abstract concept of translation quality into measurable values.

Evaluation of translation quality is intrinsically prone to subjectivity, with individual evaluators often holding different standards for what constitutes a good or poor translation. This variability can lead to significant discrepancies in the assessment results across different evaluators. The MQM framework, therefore, is utilized to provide explicit error classification criteria to mitigate such inconsistencies. It facilitates the conversion of these classifications into scores based on a uniform set of standards, leading to a more objective and comparable translation quality assessment.

D.2. Overview of the MQM Evaluation Process

The evaluation of translation quality consists of two distinct phases: “error annotation” and “score conversion”. In this task, your primary focus will be on “error annotation”. The “Score Conversion” phase, while not part of this task, is explained here for a more comprehensive understanding.

During the “error annotation” task, your job is to evaluate the quality of translation at the level of individual translation units, following the MQM framework. Each unit consists of original English text and its Korean translation. Your role is to critically compare and analyze both the original and translated texts in each unit, identifying and annotating any evident errors. Words that are deemed erroneous should be annotated according to their sub-error type and severity level under the respective error dimensions. This process is based on three main error dimensions: accuracy, fluency, and style. Please note that the error dimension of terminology, outlined in the official MQM, is not included in this task due to the non-domain-specific nature of the corpora used for our evaluation (news and presentations). More details on the annotation will be provided in section 4, “Annotation Process”.

The subsequent “score conversion” phase involves turning the annotated errors into quantifiable scores that reflect the translation quality. Major errors are assigned a weight of 5 points, while minor errors are given 1 point. The total scores for all errors within each error dimension are then added up to create the MQM dimension score. The MQM total score is obtained by aggregating these MQM dimension scores by these three error dimensions.

To maintain the integrity of this process, it is crucial that you adhere strictly to the definitions and instructions provided in the following sections during the annotation task. The definitions are refined based on the official MQM specifications [5].

D.3. Error Dimensions

D.3.1. Accuracy

This dimension evaluates whether the original meaning is well conveyed in the translated text. Words that change the original meaning are considered accuracy errors. The sub-error types under accuracy include:

- Addition: Inserting words that are not present in the original text.
- Omission: Leaving out words from the original text.
- Shift in meaning: Placing words in a different part of the text than originally intended, resulting in a change in meaning.
- Mistranslation: Translating words from the original text into words with different meaning.
- Untranslated text: Leaving words from the original text untranslated.

The severity of errors within the accuracy dimension is classified as major or minor. Major errors significantly distort the overall meaning of the text, whereas minor errors do not considerably impede comprehension. For example, if the English word “cell phone” is translated as “전자 기기”(electric device) and the main idea of the original message can still be conveyed, it belongs to minor errors. However, if

[5] https://themqm.org/

it is translated as “사무 용품”(office supplies) and the original message is significantly impacted, then it should be categorized as a major error.

For the error type “untranslated text”, any untranslated English words in the translation are considered major errors. However, untranslated words in parentheses alongside their Korean translation, such as “페이스북”(Facebook), are deemed minor errors. Hashtags, such as #MeToo, are viewed as loanwords and should be left untranslated. Therefore, translated hashtags are considered mistranslation errors.

After analyzing the original and translated text, errors are generally defined within the translated text. However, for omission errors, errors should be identified in the original text since they cannot be located in the translation.

D.3.2. Fluency

This dimension evaluates whether the translated text reads smoothly and naturally. Any words that disrupt the readability of the translated text are considered as fluency errors. Unlike other error dimensions, fluency is evaluated solely based on the translated text, not the original text. The sub-error types under fluency include:

- Grammar: Violating the grammar rules of the target language.
- Spelling: Words that are misspelled.
- Punctuation: Using punctuation incorrectly (e.g., commas, periods, question marks, exclamation marks, quotation marks, etc.).
- Encoding: Misrepresentations due to incorrect encoding processes (e.g., “&quot;” appearing where a quotation mark ” should be).
- Formatting: Violating conventional formats required in a particular region.
- Unnaturalness: Words that are awkward or unnatural.
- Untranslated text: Leaving words from the original text untranslated.

The severity of errors within the fluency dimension is also categorized as major and minor. Major errors significantly hinder the readability of the text, while minor errors, though not obstructive to understanding, can reduce the text’s overall quality.

Taking into account that ’unnaturalness’ errors have generally a minor impact on translation quality compared to other sub-error types such as grammar or encoding, they are typically classified as minor errors.

Untranslated text is considered an error in both fluency and accuracy since these errors not only impact the meaning of the text but also disrupt its readability. These errors are identified using the same criteria as explained in accuracy: untranslated words in parentheses alongside their Korean translation are considered minor errors, while words left entirely untranslated are considered major errors.

In the official MQM specification, formatting is classified under a separate error dimension, ’local convention’. However, given the rarity of such errors in this evaluation, formatting errors have been incorporated as a sub-error type under the fluency dimension. Formatting errors include incorrect representation of time, date, and currency formats. For instance, while the date format in the United States is MM-DD-YYYY, it is YYYY-MM-DD in Korea. Translations should mirror these local conventions.

Please note that fluency errors do not consider the meaning of the original text but do take into account the semantic relationship within the translated text. Hence, if a text’s flow feels unnatural due to incorrect semantic connections between words, this is deemed an ’unnaturalness’ error.

D.3.3. Style

This dimension evaluates whether the original writing style is adequately preserved in the translated text. Any words or phrases that deviate from the style of the original text are considered style errors. The sub-error types under style include:

- Formality: Having differences in formality between the original and the translated text, or inconsistent formality within the translated text.
- Structure: Having structural changes that affect the nuances of the original text (e.g., order of writing, passive/active voice, or word/sentence conversion).

The severity of errors within the style dimension is categorized as major or minor, depending on the degree of influence the error has on the overall tone of the text. Major errors are those that have a clear impact on the overall tone of the text. For example, shifting from a formal to an informal tone in the translated text, or modifying the text structure in a way that changes the original nuances. Minor errors, on the other hand, are those that have a minimal impact on the overall tone of the text. These may include slight modifications in sentence structure or minor fluctuations in formality.

Style errors can span across the entire text and are often not limited to specific words. Therefore, while other error types are calculated based on word count, style errors are calculated per text unit. In this context, a text unit refers to a sequence of words that is found to be erroneous. For instance, if a sequence like "어떤 영향을 미칠지 불확실했습니다" (translated as "it was uncertain what influence it would have") is deemed to cause a style shift, this text unit is counted as a single error, not four. This distinguishes style errors from other error dimensions, as they focus on the overall style of the text, which can affect an entire sentence or multiple phrases.

D.4. Error Annotation Process

This section describes the annotation process. During this task, it's important to stay strictly within the context of the given text pairs. Words or phrases that exceed the provided context should be annotated as errors. The annotation will be performed according to the three error dimensions, accuracy, fluency, and style which are explained in section 3.

Error annotation is to be carried out sequentially, one dimension at a time. If no errors are found within the dimension under examination, type a hyphen "-". This practice differentiates between error dimensions that are yet to be annotated and those without errors.

It's important to remember that a single word can potentially fall into multiple error dimensions simultaneously. For example, a word could be annotated as an error for both accuracy and fluency. However, it's not possible for the same word to be categorized under different sub-error types within the same dimension, such as grammar and unnaturalness in the fluency dimension. In cases where this might occur, the more severe error is chosen for annotation. The layout for each translation unit to be evaluated is as follows:

[n-th translation unit]
Original English Text
Translated Korean Text

Accuracy:
Fluency:
Style:

Your task is to identify errors within the original text and its translation and annotate them under their respective accuracy, fluency, and style dimensions. Each error should be annotated in the following format: 'Error_Word_or_Phrase(Sub_Error_ Type/Severity_Level)'. Here's an example:

[n-th translation unit]
And demonstrations also occurred in Ni'lin.
Ni'lin은 또한 시위가 일어나는 것을 목격했습니다.

Accuracy: Ni'lin(untranslated text/major), And(omission/minor), 목격했습니다.(mistranslation/major)
Fluency: Ni'lin(untranslated text/major), 또한(unnaturalness/minor)
Style: Ni'lin은 또한 시위가 일어나는 것을 목격했습니다.(structure/major)

The Korean translation "Ni'lin은 또한 시위가 일어나는 것을 목격했습니다." translates to "Ni'lin also witnessed the occurrence of demonstrations". Firstly, in terms of accuracy, "Ni'lin" remains untranslated in the translation, significantly impacting the translation quality. Hence, it's annotated as a major error under "untranslated text". The word "And", which should have been translated to "그리고" in Korean, is missing from the translated text, but this omission is considered to have a minor impact on the overall meaning of the translation. Therefore, it's annotated as a minor error under "omission". "목격했습니다 "(witnessed) is an incorrect translation that distorts the original meaning. If it were translated as "경험했습니다"(experienced), it could have preserved the original text's meaning. For that reason, it's annotated as a major error under "mistranslation".

Next, in terms of fluency, the untranslated word “Ni’lin” significantly affects the text’s readability, so it’s annotated as a major error under the “untranslated text”. While “또한” is the correct Korean word for its counterpart “also”, the Korean translation doesn’t sound natural with it, so it’s annotated as a minor error under the “unnaturalness”.

Lastly, the sentence structure has changed from active to passive voice, and this shift noticeably impacts the overall tone of the translation. Therefore, the text unit “Ni’lin은 또한 시위가 일어나는 것을 목격했습니다.” where this structural change occurs is annotated under the style dimension as a major error.

After your annotation work is done, the MQM scores can be calculated in the next score conversion phase. The annotation example above would yield MQM dimension scores of 11, 6, and 5 for accuracy, fluency, and style, respectively, resulting in an MQM total score of 22.

D.5. Final Remarks

Whenever you face uncertainty, don’t hesitate to revisit these guidelines. Your deep understanding and correct implementation of the MQM framework directly influence the quality and consistency of the evaluations. Remember, your meticulous work is crucial in maintaining high standards of translation quality evaluation. Thank you for taking the time to read these guidelines and good luck with your annotation work!