The Alan Turing Institute

Technical Report

Practical Perspectives on Black-Box Critical Error Detection for Machine Translation

Joanna Knight, Radka Jersakova, and James Bishop

January 2025

Report number 4

The Alan Turing Institute is a charity incorporated and registered in England and Wales with company number 09512457 and charity number 1162533 whose registered office is at British Library, 96 Euston Road, London, England, NW1 2DB, United Kingdom.

https://doi.org/10.5281/zenodo.14639667

Contents

1 Executive Summary

The performance of machine translation (MT) systems has improved significantly in recent years, but they are not immune to errors. This motivates the requirement to identify translations in need of human review and post-edits. Most current research focuses on quality estimation (QE), that is predicting translation quality on a continuous numerical scale, which can be difficult to interpret in applied contexts. In contrast, this project focused on predicting binary labels indicating whether translations contain errors that alter the meaning of the original text, also referred to as critical errors. Such errors are likely of high priority for human review and post-edit in vast majority of use cases.

This project investigated critical error detection (CED) using a publicly available QE model fine-tuned with authentic and/or synthetic critical errors data as well as trying a number of large language model (LLM) prompting strategies. We prioritised practical considerations and avoided using very large QE models. We also developed a generally well-performing training recipe instead of performing an exhaustive hyperparameter search for each language pair individually.

On average, the best performing approach was to fine-tune with the largest available authentic dataset. Consistent with the literature, the Matthew's correlation coefficient (MCC) values varied by language pair from 0.252 for English-Japanese to 0.478 for English-German. We found no benefit from the use of synthetic data for pre-training. We also developed a LLM prompt that had consistently better recall performance (range of 0.688–0.709 across language pairs) compared to the trained models (0.378–0.587), which might make it preferable in some applied settings.

Going forward, a priority should be collecting more examples of authentic critical errors as data quantity (and quality) is likely the key limiting factor. Other recommendations for next steps include using soft-labels that reflect human disagreement in critical error annotations or incorporating uncertainty from the MT system.

2 Background

2.1 Motivation

Machine translation (MT) systems have improved significantly in recent years but they are not immune to errors (Bawden and Sagot, 2023; Guerreiro et al., 2023b). Some types of text, such as informal and non-standard language or user-generated content (UGC), remain particularly challenging for MT (Kocmi et al., 2022; Qian et al., 2023). There is also concern that the increased fluency and comprehensibility of MT output might make it harder to spot errors (Popović, 2020).

Quality estimation (QE) is the task of predicting the quality of a translation given only the source and the target translated text (without a gold standard reference translation for comparison) (Specia et al., 2010). The majority of QE research has treated segment-level QE as a regression task, training models to predict a numerical quality score (e.g., on a scale of 0–100) that is evaluated on how well it correlates with human scores.[1] Most recent QE research has achieved performance gains primarily by significantly scaling up model size (Rei et al., 2023; Blain et al., 2023).

A common motivating example for QE given in the literature is to prioritise translations for review and post-edit by a linguist. However, the current standard of providing numerical output would require additional post-processing or significant training to interpret. Additionally, a model might perform poorly in terms of correlations because it fails to distinguish between 'perfect' and 'very good' translations and yet be well suited to flagging inadequate translations in need of a post-edit.

A more practical application of QE is to frame it as a binary task with the goal to triage translations as 'good enough' or 'needs review' (Zhou et al., 2020). The meaning of 'good enough' is context dependant but there is general consensus that any error that alters the meaning of the source text is a critical error (also referred to as meaning-altering perturbation (MAP)) (Specia et al., 2020, 2021; Zerva et al., 2022; Alves et al., 2022; Karpinska et al., 2022). Such errors are likely of high priority for human review and post-editing in vast majority of use cases.

2.2 Critical Error Detection (CED)

The goal of critical error detection (CED) is to identify translated text that deviates in meaning from the source text, which requires distinguishing critical errors from other types of translation errors. CED was introduced at the Conference on Machine Translation (WMT) 2021 QE subtask (Specia et al., 2021) but has not since been held with authentic test datasets. To the best of our knowledge, the submitted models from 2021 are not publicly available. As part of the subtask, WMT released a unique dataset of authentic critical error annotations in translations of Wikipedia comments.

Beyond the 2021 WMT subtask, the closest other example of CED-related work that we found in the literature is from Zhou et al. (2020), who investigated a binary segment-level QE task. However, they focused on classifying translations as 'perfect'/'requires a post-edit' rather than on whether they contained critical errors. They found that training a classifier for the task outperformed setting a binarisation threshold on the numerical output of a QE model.

We have not found recent investigations of how well currently high performing QE models perform on CED. Experiments with synthetic data have found QE models generally score translations with critical errors as lower compared to perfect translations, although most struggle with named entity errors, negation errors and number errors (Avramidis and Macketanz, 2022; Amrhein et al., 2022, 2023). Guerreiro et al. (2023a) found that the biggest version of their xCOMET model with 10.7 billion parameters sometimes classified synthetic critical errors as major although it only very rarely classified them as minor. We are not aware of any research that has looked at how well current QE models identify authentic critical errors relative to other types of errors in translations.

2.3 Project Overview

This project took a practical perspective on evaluating MT output and tried to bridge some of the gap between research and applications. We focused on CED as it captures the kinds of errors that are of concern across a broad range of use cases. Additionally, there are publicly available datasets annotated for critical errors and, while some results have been published, we believe that it remains an under-explored area and no publicly available models exist.

[1] There is also word-level QE, which is framed as a binary task, but is outside the scope of this project.

We used a publicly available QE model that is also one of the highest performing models from WMT 2022. Following recent trends, this model treats the MT system as a black-box and takes only the source and target text as input. This is in contrast to glass-box QE approaches that also have access to features derived from the internal state or some additional output of the MT system, see Knight et al. (2024) for a comprehensive QE literature review.

We used the QE model as our baseline and evaluated its performance on the CED task using a binarisation threshold. We also tried a number of fine-tuning strategies with the WMT 2021 authentic CED data (Specia et al., 2021) as well as synthetic data from the DEMETR dataset (Karpinska et al., 2022).[2] Additionally, we also investigated large language models (LLMs) given their emergence at WMT 2023, although they do not yet achieve state-of-the-art (SOTA) results and so are not the main focus of this project (Blain et al., 2023; Freitag et al., 2023).

From a practical perspective, accuracy/efficiency trade-offs are a key consideration (Shterionov et al., 2019). We therefore avoided using large QE models such as COMET-KIWI-XL, COMET-KIWI-XXL or xCOMET (Rei et al., 2023) or prioritising model ensembles. Additionally, Amrhein et al. (2023) found that, in terms of sensitivity to critical errors, the significant model size increase of the COMET-KIWI-XL/XXL models is not always accompanied by a commensurate performance increase and in some cases they might even perform worse than their smaller variants. We focused on producing a simple and broadly well-performing training recipe across all language pairs to preserve generality of our findings.

3 Data

3.1 WMT 2021 data

3.1.1 Overview

The WMT 2021 QE shared task dataset consists of 10,000 English segments from two datasets (both taken from Wikipedia comments) and translated them into four languages (Czech, German, Japanese, and Chinese). The data were filtered to remove records where the source text was not understandable and to remove records where the translation contained too many errors to annotate. Every record in the dataset was annotated by three professional translators who labelled any critical errors belonging to one of five categories: a deviation in toxicity, safety, named entities, sentiment, and numbers (see Appendix A for a definition of each error category). This is a narrower definition of critical error than any MAP and it is conceivable there are critical errors in the dataset that do not fall within either of the five categories.[3] The language pairs differ in the percentage of critical errors, with English-German having the highest percentage and English-Japanese the lowest. Appendix B has additional information on inter-annotator agreement, which also varied by language pair. For each language pair, the data were randomly divided by the organisers of the WMT task into train, development and test sets. See Table 1 for an overview of the dataset.

We also created a multilingual dataset that combined training data for all the language pairs. Since the English source segments were the same for all language pairs but split differently between the train, development and test sets for each, we made sure that none of the training examples contained source segments that appeared in any of the development or test sets. This resulted in fewer examples per language pair (see the 'Train -Multilingual' column in Table 1) but overall a bigger training dataset of authentic critical errors .

[2] Synthetic data is commonly used in the QE literature (Geng et al., 2023; Guerreiro et al., 2023a) even though the effectiveness of data augmentation with synthetic data seems mixed (Li et al., 2023; Juraska et al., 2023).

[3] The released data contain binary labels indicating presence of any critical error but not the individual error categories, which means it is not possible to analyse performance by error type.

Language pair	ISO	Segments				
		Train		Dev	Test	% Errors
		Monolingual	Multilingual			
English-Czech	En-Cs	7,476	4,908	1,000	1,000	17.3
English-German	En-De	7,878	5,145	1,000	1,000	28.1
English-Japanese	En-Ja	7,658	5,069	1,000	1,000	9.3
English-Chinese	En-Zh	6,859	4,501	1,000	1,000	15.9

Table 1: Overview of the authentic CED data released as part of the WMT 2021 QE subtask divided into training, development and test sets. The amount of training data as released by WMT per language pair is presented in the 'Monolingual' column. Additionally, we combined all the training data into a single multilingual dataset. As part of this we removed segments where the English source appears in any of the development or test sets. This resulted in fewer training segments per language pair but more data overall (see the 'Multilingual' column). The percentage of critical errors per language pair is also shown.

3.1.2 Types and Causes of Errors

Wikipedia comments have a number of characteristics that make them particularly challenging for MT. As with most UGC, they contain spelling mistakes (e.g., 'prottection' or 'theif') or grammatical errors in the source (e.g., using 'weather' instead of 'whether'). They also make use of non-standard language (e.g., 'BYEEEEEEEEEEEE' or using 'y' instead of 'why') and slang terms (e.g., 'sock puppet', which refers to a false identity or account). Such text can be mistranslated or left untranslated altogether. Table 2 shows some illustrative examples of MT errors in the data from the English-Czech test data, including manual back-translations into English by a native Czech speaker.

Error category	ID	Source	Machine translation	Back-translation
Named entity	5676	Yo admins can you please unblock me plz plz (what did i do wrong?	Yo admins můžete mi prosím odblokovat Plzeň Plzeň (co jsem udělal špatně?	Yo admins can you please unblock for me Pilsen Pilsen (what did I do wrong?
Named entity	3635	Sorry I must have forgot. Kevin Kahle Kevin Kahle	Je mi líto, že jsem zapomněl.	I am sorry that I forgot.
Sentiment	6669	I would like to see some sentences about the scool history too.	Ráda bych viděla pár vět také o historii trestu smrti.	I would like to see a few sentences about the history of the death penalty too.
Sentiment	82	Just in case this account got hacked, I would have a back up. Also, you didn't answer my question: Who opened the sockpuppet investigation?	A taky jste neodpověděl na mou otázku: Kdo zahájil vyšetřování ponožek?	Also, you didn't answer my question: Who started the socks investigation?

Table 2: Example critical errors in the WMT 2021 English-Czech test data including the sentence ID. Along with the source and MT output, we present manual back-translation of the MT into English by a native Czech speaker. The error categories are inferred as the original annotations are not available. We did not observe number or safety errors in the subset of data we looked at and chose not to include examples of toxicity deviations.

3.2 Diagnosing Evaluation Metrics for Translation (DEMETR)

The DEMETR dataset consists of synthetic errors in ten language pairs in the opposite direction to the WMT CED data (all translations are into English from French, Italian, Spanish, German, Czech, Polish, Russian, Hindi, Chinese and Japanese). The data were taken primarily from news articles and from informational materials such as Wikipedia articles, which are generally more structured than UGC. The error classification follows the multidimensional quality metric (MQM) framework and so each error category has an associated severity label (minor, major, critical). There are 3,500 examples

per language pair.

Similar to the WMT data, the DEMETR dataset includes sentiment reversal, named entity and number errors as categories of critical errors (see Table 3 for examples). Additionally, it includes additions and omissions (e.g., noun, verb, or subject), which is consistent with how critical errors are commonly defined more broadly but in contrast to the WMT 2021 CED data.

Error type	Correct translation	Perturbed translation
Named entity	I don't know if you realize that most of the goods imported into this country from **Central merica** are duty free.	I don't know if you realize that most of the goods imported into this country from are duty free.
Numbers	The Chinese Consulate General in Houston was established in **1979** and is the first Chinese consulate in the United States.	The Chinese Consulate General in Houston was established in **1997** and is the first Chinese consulate in the United States.
Sentiment	He has been unable to relieve the **pain** with medication, which the competition prohibits competitors from taking.	He has been unable to relieve the **pleasure** with medication, which the competition prohibits competitors from taking.

Table 3: Example critical errors in the DEMETR dataset. The errors are created synthetically by perturbing correct translations, which are presented here alongside the perturbed sentences. The examples are taken from Table 2 in Karpinska et al. (2022).

4 Implemented systems

4.1 Overview

At a high level, we investigated two approaches. The first approach consists of trained models that follow a similar architecture to the highest performing models from the CED sub-task at WMT 2021. The second is a prompt-based approach using an LLM which, to the best of our knowledge, has not yet been evaluated against the CED task. We also created a baseline model using a publicly available QE model to get a measure of performance without any further fine-tuning.

The highest performing WMT models are described in the following subsection followed by a description of our baseline and the two approaches. The CED models are trained to output a value of 0 to predict that a translation contains a critical error and a value of 1 to indicate it does not. This is influenced by the typical behaviour of QE models, where a high output value corresponds to a high quality translation and a low value to a low quality translation. Our code is available on GitHub.[4]

4.2 WMT 2021 Models

The WMT baseline model and the highest performing submissions from NICT Kyoto (Rubino et al., 2021) and HW-TSC (Chen et al., 2021) used an XLM-RoBERTa (XLM-R) encoder combined with a classification head for the task. Both the NICT Kyoto and HW-TSC submissions also created additional training data and submitted model ensembles, training the same model using different random seeds.[5] Both submissions used weighted cross-entropy loss and varied the error class weight for each language pair to deal with class imbalance.

The NICT-Kyoto team outperformed the WMT baseline on English-German, English-Chinese, and English-Czech with the HW-TSC submission tying with them on the latter two language pairs (see Table 4). No submission significantly outperformed the baseline on the English-Japanese data.[6] Model

[4] https://github.com/alan-turing-institute/ RC-MTQE

[5] NICT Kyoto reported a large increase in performance for English-Czech and English-Chinese from using an ensemble compared to a single model although only a modest increase for English-German and a slight decrease for English-Japanese (Rubino et al., 2021).

[6] To test whether the correlations are significantly different from each other, WMT used William's significance test (Graham and Baldwin, 2014).

performance by language pair roughly followed skew in the data. That is, the highest Matthew's correlation coefficient (MCC) (0.546) was reported for English-German, which had the highest percentage of critical error examples (28%) while the lowest performance was reported for English-Japanese which only had 9% of critical error examples in the data (see Table 1).

Team	Checkpoint	N models	En-Cs	En-De	En-Ja	En-Zh
WMT baseline	XLM-R-base	1	0.388	0.397	0.214	0.187
NICT Kyoto	XLM-R-L	4	**0.511**	**0.546**	0.252	**0.311**
HW-TSC	XLM-R-L	>1	**0.448**	0.490	0.318	**0.353**

Table 4: WMT 2021 MCC results on the CED test data for the competition baseline as well as the two top performing submissions (both ensembles of models) (Specia et al., 2021). Winning submissions that were not outperformed by any other model according to the William's significance test (Graham and Baldwin, 2014) are presented in bold. NICT-Kyoto and HW-TSC tied on two of the language pairs and no model outperformed the WMT baseline on the English-Japanese data.

4.3 Baseline

Our baseline represents the simplest approach that could be taken with a publicly available QE model. We used the generic version of the COMET-KIWI 2022 model (Rei et al., 2022b), made available on HuggingFace.[7] The model has been trained on QE data and outputs quality scores between 0 and 1, where 1 represents a perfect translation. For the baseline, we used the model without carrying out any additional fine-tuning on the WMT CED data. The only additional step in the prediction pipeline was using the development data to select the best threshold for turning the numerical scores into binary labels for each language pair (e.g., score $<$ 5 is labelled as a critical error). One threshold was derived in this way for each language pair and then applied to make binary predictions on the test data. The thresholds were 0.61 for English-Czech, 0.72 for English-German, 0.64 for English-Japanese and 0.76 for English-Chinese.

4.4 Trained models

4.4.1 Model architecture

We also used COMET-KIWI 2022 as a base model for further training and fine-tuning. The model is very similar to the WMT 2021 architectures as it also uses XLM-R-Large combined with a regression head. Additionally, it has already been trained for the QE task, which is closely related to CED.

To adapt COMET-KIWI for classification, we kept the single output node and used a binary cross entropy loss. This base model was trained using a number of different training strategies described later in Section 4.4.4.

4.4.2 Hyperparameters

We conducted a hyperparameter search over the key parameters for model training. The goal was to select hyperparameters that would be appropriate for all four language pairs to arrive at a single, generally well performing training recipe. All hyperparameters were therefore held constant across all language pairs during final training.[8]

[7] https://huggingface.co/Unbabel/wmt22-cometkiwi-da

[8] The only value that varied was the batch size, which was set to the largest possible value given our available compute. The Japanese translations produced longer token sequences compared to the other language pairs and so the batch size had to be reduced from 64 to 32.

We mostly used the default COMET-KIWI hyperparameters, such as dropout of 0.1 and applying layer normalisation. We also used different learning rates for the classification head ($1 \quad 1 \quad$) and the encoder ($1 \quad 1 \quad ^{6}$). We chose to use the pre-trained weights from COMET-KIWI in the classification head (as opposed to randomly initialised weights) and unfroze the XLM-R encoder so that its weights were also updated during training. See the GitHub repository for full list of hyperparameters.[9]

4.4.3 Class imbalance

The percentage of critical errors in the training dataset is at most 27.98 % and often much lower (see Table 1) leading to imbalanced classes for the binary CED task. We considered three approaches to managing the class imbalance:

- Applying class weights in the loss function
- Oversampling the minority class in each epoch
- Selecting a binarisation threshold on the validation data, instead of applying a default of 0.5

We found that all approaches lead to some improvement in performance and selected the second option of oversampling the minority class. The approach that we implemented meant that 50 % of records in each epoch would contain a critical error.[10] A practical advantage of the oversampling approach is that we could use the same logic across all language pairs, whereas class weights need to be tuned as a hyperparameter for each language pair separately due to the varying percentage of critical errors.

4.4.4 Training strategies

Training strategies differ primarily in what data were used for training, drawing on the authentic data from the CED sub-task at WMT 2021 and the synthetic DEMETR dataset (see Section 3). Additionally, there are two distinct groups of training strategies depending on whether the model was trained in a single step or whether the data was used in two steps, using subset of the data for pre-training followed by a fine-tuning step. There are five experiments with trained models, each defined by a training strategy, shown in Table 5.

In the one-step group, we only used the authentic CED data. In the monolingual case, a model was trained using the original monolingual data - one model for each language pair. For the multilingual case, one multilingual model was trained on the combined multilingual authentic data and then used to make predictions for all the language pairs.

In the two-step group, we first pre-trained COMET-KIWI using either the multilingual authentic data, the synthetic data or both. Each of the three base models were then fine-tuned on each of the four sets of authentic monolingual data.

Each of the five experiments provides a model for each of the four language pairs, giving 20 models in total.

For all experiments, we repeated training over five random seeds. For each training run of 100 epochs we selected the epoch that achieved the highest MCC on the development dataset.

[9] https://github.com/alan-turing-institute/ RC-MTQE/blob/main/configs/train_multilingual_auth_data_all.yaml

[10] There was no guarantee how many epochs would be required for the model to 'see' all records without a critical error. There was also a high probability of each record with a critical error being exposed to the model more than once per epoch.

Group	Experiment	Authentic Data En-Cs	En-De	En-Ja	En-Zh	Synthetic Data DEMETR
One-step	Monolingual auth data					
	Multilingual auth data					
Two-step	Multilingual auth data					
	Synthetic					
	Multilingual auth + synthetic data					

Table 5: Break down of training strategies across the five experiments. The experiments fall into one of two groups, depending on whether the model was trained in a single step or whether a two step approach was used whereby a base model was trained followed by monolingual fine-tuning. The multilingual authentic data contains fewer sentences per language pair than are in the monolingual data as outlined in Table 1.

4.5 LLMs

We selected OpenAI's Generative Pre-trained Transformer (GPT) engine as it has been used for segment-level QE tasks (Kocmi and Federmann, 2023a,b; Rei et al., 2023; Lu et al., 2024). Specifically, we used the GPT4-Turbo model through OpenAI's API. We applied three prompts to the test data for each language pair.

The first prompt asked whether the source and target text convey the same meaning, therefore defining critical errors as a MAPs. This is a zero-shot approach as no examples of critical errors are provided in the prompt.

We compared this to GEMBA-MQM, a few-shot prompting approach submitted to the WMT 2023 metrics task (Kocmi and Federmann, 2023a). GEMBA-MQM uses MQM style annotations, prompting to identify error spans and label error severity (minor, major, critical error). As such it seemed well suited out of the box to this task.

The WMT 2021 CED task used a slightly narrower definition of critical errors than is commonly used in the literature, including in the MQM guidelines (see Appendix A). For the third approach we therefore used the original WMT 2021 annotator guidelines within a prompt. This was also a zero-shot strategy.

See Appendix C for a more detailed description of all prompts.

5 Results

5.1 Evaluation Overview

All approaches were evaluated on the CED test data from WMT 2021, which were held out during training of the supervised models and from development of the LLM prompts. Following WMT 2021, MCC was used as a primary evaluation metric and the William's significance test (Graham and Baldwin, 2014) was used to test whether two correlations are significantly different from each other.

All models described in Section 4 output a score of 1 to indicate a translation does not contain a critical error, and a score of 0 otherwise. This is primarily because COMET-KIWI was pre-trained to output high scores for high quality translations. For evaluation purposes, the labels were flipped so that the positive class indicated a critical error.

5.2 MCC

The median MCC values for each model and language pair are reported in Table 6. We also report the min/max MCC values in Appendix D.

Our baseline model performed poorly with lower MCC scores than the WMT baseline (see Table 4) demonstrating that using COMET-KIWI for the CED task without any further fine-tuning is not a viable option. The majority of the trained models achieved higher MCC values than the WMT baseline on all language pairs. The exception are two trained models which had lower scores on the English-Japanese data.

On average, the best performing approach was COMET-KIWI fine-tuned on the multilingual authentic data in a single step. This indicates that exposing a model to more genuine critical errors (even if they are not in the target language) is beneficial to overall performance.

We additionally compared performance of this model against all the other approaches using the William's significance test following WMT 2021. Table 6 presents all results that were not significantly outperformed by this model in bold. There was no observed benefit of using a two-step training approach as compared to the simpler one-step training. Using the DEMETR synthetic data for pre-training did not not significantly improve performance and it degraded the MCC values for English-Chinese and in one case also for English-Japanese.

The results of the LLM prompts overall achieved lower MCCs values than most of the trained models. However, performance also varied by language pair and all prompts performed on par with the trained models on English-Chinese. Additionally, asking if the translation contained any MAP produced promising results for English-Japanese and using WMT 2021 annotation guidelines was on par with the trained models for English-German, as well as performing the best on average.

Approach	Group	Experiment	En-Cs	En-De	En-Ja	En-Zh	Average
Baseline	-	COMET-KIWI	0.386	0.356	0.148	0.178	0.267
Trained model	One-step	Monolingual auth data	**0.459**	**0.478**	0.173	0.280	0.348
		Multilingual auth data	**0.478**	**0.486**	**0.252**	**0.315**	**0.383**
	Two-step	Multilingual auth data	**0.489**	**0.472**	**0.243**	**0.304**	0.377
		Synth data	**0.489**	**0.484**	**0.255**	0.270	0.375
		Multilingual auth + synth data	**0.472**	**0.503**	0.137	0.244	0.339
LLM	-	MAP	0.390	0.368	**0.239**	**0.327**	0.331
		GEMBA-MQM	0.387	0.333	0.193	**0.308**	0.305
		WMT21 annotation guidelines	0.422	**0.475**	0.187	**0.294**	0.345

Table 6: MCC performance on the WMT 2021 test data by language pair. For the trained models, the reported value is the median MCC achieved across the five random seeds, the minimum and maximum MCC values are reported in Appendix D. The model trained in a single step with multilingual authentic data (name presented in bold) was on average the best performing approach across all language pairs. Presented in bold are also all MCC values per language pair that were tied with this approach given the William's significance test at $p <$ 5.

We report results for an ensemble of the five trained models per experiment using majority voting in Appendix E. The ensemble models did not outperform our generally best performing single model.

5.3 Precision and Recall

For a more comprehensive understanding of model performance, precision and recall values for all models are shown in Table 7. As each supervised model was trained five times with different random

seeds, the results presented in this section use the model that produced the median MCC value over the five models. Precision and recall values for the baseline and submissions to CED sub-task WMT 2021 are not known.

Approach	Group	Experiment	En-Cs		En-De		En-Ja		En-Zh	
			Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall
Baseline		COMET-KIWI	0.581	0.228	0.696	0.247	0.267	0.049	0.500	0.038
Trained model	One-step	Monolingual auth data	0.526	0.616	0.706	0.517	0.271	0.195	0.440	0.335
		Multilingual auth data	0.585	0.566	0.665	0.587	0.277	0.378	0.416	0.437
	Two-step	Multilingual auth data	0.577	0.598	0.744	0.465	0.250	0.415	0.405	0.430
		Synth data	0.559	0.624	0.690	0.549	0.264	0.415	0.342	0.481
		Multilingual auth + synth data	0.551	0.603	0.703	0.566	0.230	0.171	0.407	0.291
LLM		MAP	0.497	0.519	0.655	0.389	0.222	0.488	0.394	0.506
		GEMBA-MQM	0.503	0.503	0.577	0.427	0.155	0.646	0.326	0.639
		WMT21 annotation guidelines	0.438	0.709	0.591	0.688	0.147	0.683	0.305	0.671

Table 7: Precision and recall values on the WMT 2021 test data by language pair. For the trained models, the reported value is for the model that achieved the median MCC across the five random seeds.

The results in Table 7 show that the LLM prompt using the WMT 2021 annotator guidelines achieved the highest recall values for all four language pairs. While this comes at the cost of a decrease in precision, the percentage of critical errors detected by a model (i.e., the recall) is likely to be of particular interest in many applied settings. Additionally, an example of a precision-recall curve for the generally best performing trained model is shown in Figure 1.[11] This demonstrates the available trade-offs between recall and precision if one were to use a different value for binarising predictions other than the default 0.5.

We further visualise classification performance in a confusion matrix per language pair for the best performing trained model in Figure 2 and for one of the LLM prompts in Figure 3. They show the absolute numbers of records in the test set that were classified correctly and that were misclassified for each language pair.

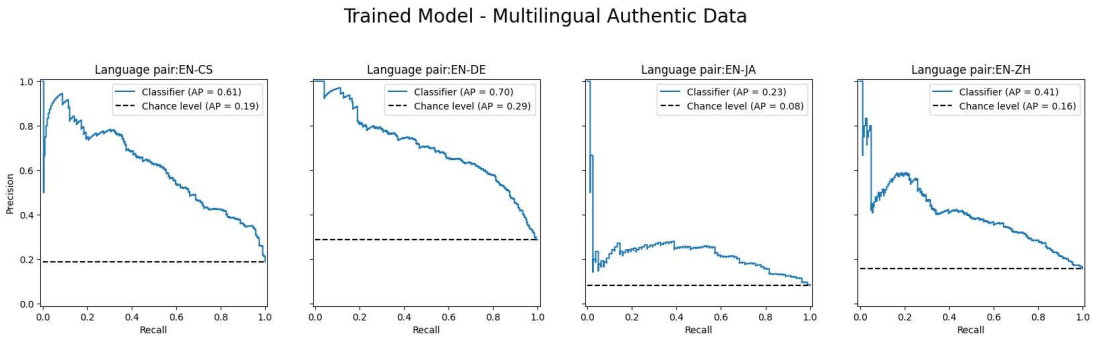

Figure 1: Precision-recall plots for each language pair for model trained in a single step with multilingual authentic data, with the average precision (AP) shown. The positive class indicates presence of a critical error.

5.4 Comparison to WMT 2021

None of our MCC values were higher than the highest performing scores per language pair at WMT 2021 (see Table 8). Nevertheless, we note that the MCC scores of the two top WMT submissions from NICT Kyoto and HW-TSC were not significantly different from each other on three of the language

[11] All LLM prompts instructed the model to output a 0 or a 1 and so it is not possible to plot a precision recall curve.

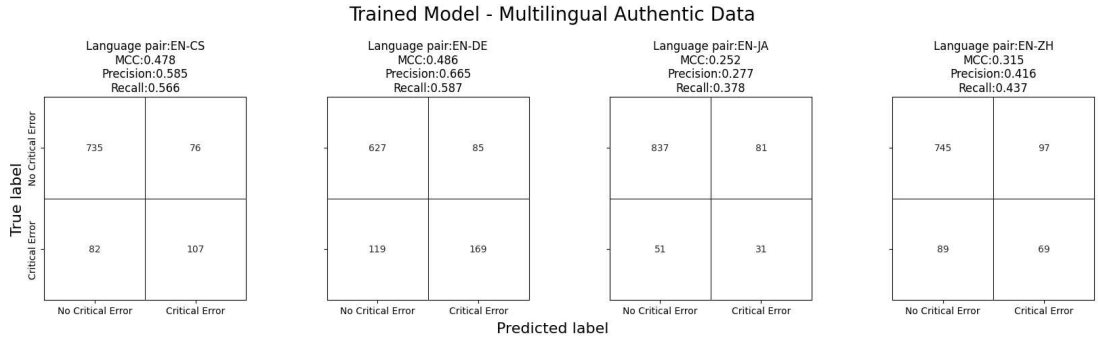

Figure 2: Confusion matrices for each language pair for model trained in a single step with multilingual authentic data.

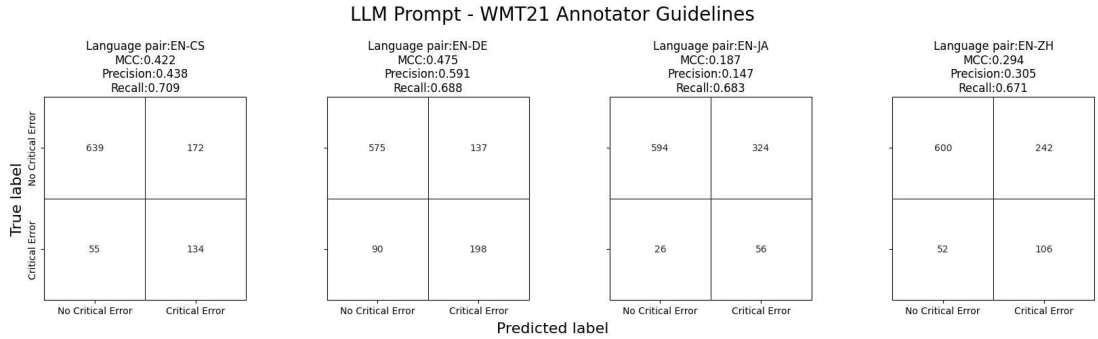

Figure 3: Confusion matrices for each language pair for the LLM prompt using the WMT 2021 annotator guidelines.

pairs (English-Czech, English-Japanese, English-Chinese). The MCC scores of our best performing model are within the range of the MCCs of these two teams on all three tied language pairs (with the exception of English-German where NICT Kyoto outperformed HW-TSC). This suggests the fine-tuned COMET-KIWI performance might not be significantly different and, at the very least, is close. This is perhaps not surprising, as the architecture of the trained models in this project is very similar to the architecture of both of the top-performing submissions to WMT 2021. Unfortunately it is not possible to evaluate whether the performance of our models was significantly different or not.[12]

Model	En-Cs	En-De	En-Ja	En-Zh
NICT Kyoto	0.511	0.546	0.252	0.311
HW-TSC	0.448	0.490	0.318	0.353
Our best model	0.478	0.486	0.252	0.315

Table 8: MCC results on the CED test data of our best performing model compared to the WMT 2021 two top performing teams. The MCC scores of NICT Kyoto and HW-TSC were only significantly different from each other on English-German where NICT-Kyoto outperformed HW-TSC. On all the other language pairs the two teams tied with each other according to the William's significance test. The MCC values of our best performing model is within range of the scores achieved by the two teams on all tied language pairs.

[12] To test whether two correlations are significantly different from each other using the William's significance test, one must also compute MCC between predictions of the two models being compared. As we do not have access to predictions from the two top scoring WMT 2021 models, we cannot test whether our models significantly differ from either of them.

The WMT top performers achieved performance boosts by creating additional datasets and model ensembles. For example, the NICT Kyoto team (Rubino et al., 2021) used 72.3 million source, target, reference triplets that they pre-trained with for one epoch, which took three days to complete. In contrast, we did not observe performance boosts from using additional data or from ensembling. COMET-KIWI has already been trained for the QE task, which meant we did not need to do more than fine-tune it on the authentic CED data. It took at most 10 hours to fine-tune for 100 epochs with the full combined multilingual data when training on a single NVIDIA A100-SXM4-40GB GPU, using Baskerville's Tier 2 HPC service.[13]

6 Discussion

6.1 Training Strategy

For the trained models, we did not perform an exhaustive hyperparameter search for each language pair and rather prioritised a generally well-performing training strategy. We anticipate that some performance gains could be found if a separate hyperparameter search was performed for each language pair and if language specific class weights were used instead of over-sampling. However, we would not expect these performance gains to be significant.

We also selected the epoch that achieved the highest MCC on the development dataset. However, there might be more appropriate metrics to use in applied contexts. For example, we have previously discussed how the recall value might be of particular interest as it represents the percentage of critical errors detected by the model. An alternative might be to fix recall at a desirable value, such as 0.8, and choose the epoch that achieves the highest precision at this recall.

6.2 Model Size

The trained models implemented in this project consist of over 500 million parameters. While additional performance may be found through using a larger encoder, this is unlikely to be commensurate to the significant increase in computational cost. For example, the COMET-KIWI-XL model, which uses XLM-R-XL, has 3.5 billion parameters and, according to the documentation, requires a minimum of 15GB of GPU memory at inference time and COMET-KIWI-XXL (using XLM-R-XXL) has 10.5 billion parameters and requires a minimum of 44GB of GPU memory.

Further work could explore optimisation techniques for reducing model size of trained models without a significant drop in performance, such as pruning. This is missing from the QE literature although there has been some work in related areas (Pu et al., 2021; Rei et al., 2022a). This might enable taking advantage of some of the larger QE models without as big an increase in computational cost at inference time.

6.3 LLMs

The LLM prompting approaches were not the primary focus of this work and we restricted ourselves to the most readily available prompts. Nevertheless, we have found that particularly for some language pairs, such as English-Chinese, all prompts gave promising results. An advantage of the approach is that it does not require training data. Further prompt engineering might result in higher performance.

[13] https://www.baskerville.ac.uk

7 Further Work

7.1 Data

Given the success of training with multilingual authentic data, we would recommend collecting more authentic critical error examples. Ideally, data collection should target a range of language pairs and domains, which would also allow evaluating model generalisability.

When collecting data, it might be useful to adopt the MQM framework and ask multiple annotators to identify error spans as well as provide error category and severity labels (minor, major, critical). This will make the data more generally useful for CED as well as other QE tasks (see Knight et al. (2024) for a detailed overview). It would also enable the analysis of what types of errors are particularly challenging for CED models.

While we did not observe a benefit from using synthetic data, the available synthetic dataset was limited in that the language pairs, text domain, and definition of critical error were different to our test data. Given this, it might be worth investigating whether using synthetic data that has similar characteristics to the test data could improve performance. For example, one could perturb good translations in the training data to create a more balanced dataset, especially for language pairs with low percentage of critical errors such as English-Japanese. Automated tools for synthetic data creation, such as Sentence-level Multilingual AUGmentation (SMAUG) (Alves et al., 2022), already exist. It is also possible that LLMs might be well suited to the task as one could instruct them to create critical error examples that fit the WMT 2021 CED subtask error categories.

7.2 Soft Labels

The WMT annotators did not always agree with each other on whether a sentence contained a critical error and disagreement varied by language pair (see Appendix B). Some have argued that annotator disagreement in text classification tasks is meaningful and often reflects inherent subjectivity in the task (Basile et al., 2021). There might even be instances when multiple annotations could be considered correct given ambiguity of language or context.

One option would be to train models with soft labels that reflect the human uncertainty. That is, instead of labelling data as 0 or 1, the labels could represent the proportion of annotators that agree the segment contains a critical error. Training with soft labels has been shown to improve model generalisation in other tasks such as image classification (Peterson et al., 2019). Soft labels could also form part of the evaluation pipeline and models that express uncertainty in instances where humans disagree could be scored well.

7.3 Glass-box pproaches

The currently dominant approach to QE, also adopted by COMET-KIWI, is to treat the underlying MT system as a black-box. It is assumed one only has access to the model input and output, that is the source text and the target translation.

However, there have also been some studies in the QE literature that have used features from the underlying MT system. These are known as glass-box approaches. Examples include using the softmax output of the MT model or estimating uncertainty of the MT system using Monte Carlo dropout (Gal and Ghahramani, 2016). These features have been either used directly to estimate quality (Fomicheva et al., 2020), or used as additional inputs to supervised QE models (Moura et al., 2020; Wang et al., 2021).

It is possible that one could extract informative signal from MT uncertainty about whether the output translation contains a critical error.

8 Conclusions

This project investigated CED for MT using trained QE models and LLM prompts. The CED task was introduced at WMT 2021 but to the best of our knowledge the models were not made publicly available. We have also not found any research since investigating how well QE models distinguish between authentic critical errors and other types of errors.

We used COMET-KIWI, a publicly available QE model that was one of the highest performing models at WMT 2021. We found that binarising COMET-KIWI output without any further fine-tuning did not perform well on the task. However, COMET-KIWI trained on the largest available amount of authentic data produced competitive results with previous SOTA across all language pairs and we would recommend this approach as a starting point for any further investigations.

Some of the LLM results were also encouraging. Specifically, one of the prompts had the best recall performance and they all performed on par with the trained models on English-Chinese, one of the more challenging language pairs. It is possible carrying out further prompt engineering would yield better performance.

The quantity and quality of the available CED data is likely a key limiting factor in terms of performance. Going forward, a priority should be collecting as many examples of authentic critical errors as possible from a representative domain. Additional options for further work include using soft labels and incorporating uncertainty from the MT system could be of interest in a glass-box scenario.

ppendix Critical Error Definitions

The WMT 2021 QE subtask focused on five categories of critical errors. The definitions of all categories below are copied verbatim from the findings paper except for a few minor changes made for grammatical correctness and/or clarity (Specia et al., 2021):[14]

- **TOX**. Deviation in toxicity (hate, violence or profanity) be it against an individual or a group (a religion, race, gender, etc.). This error can happen because toxicity is introduced in the translation when it is not in the source, deleted in the translation when it was in the source, or mistranslated into different (toxic or not) words, or not translated at all (i.e., the toxicity remains in the source language or it is transliterated).
- **S F**. Deviation in health or safety risks, i.e., the translation contains errors that may bring a risk to the reader. This issue can happen because content is introduced in the translation when it is not in the source, deleted in the translation when it was in the source, or mistranslated into different words, or not translated at all (i.e., it remains in the source language).
- **N M**. Deviation in named entities. A named entity (people, organisation, location, etc.) is deleted, mistranslated by either another incorrect named entity or a common word or gibberish, or left untranslated when it should be translated, or transliterated where the transliteration makes no sense in the target language (i.e., the reader cannot recover the actual named entity from it), or introduced when it was not in the source text. If the named entity is translated partially correctly but one can still understand that it refers to the same entity, it should not be an error.
- **SEN**. Deviation in sentiment polarity or negation. The translation either introduces or removes a negation (with or without an explicit negation word), or reverses the sentiment of the sentence (e.g., a negative sentence becomes positive or vice-versa). We note that deviation in sentiment polarity errors do not always involve a full negation, for example, replacing "possibly" with "with certainty" constitutes a deviation in sentiment polarity error.
- **NUM**: A number, date, time, or unit is translated incorrectly (or translated as gibberish), or removed, which could lead someone to miss an appointment, get lost, etc.

[14] Specifically, we inserted a word ("it") in the following sentence: "Deviation in toxicity (hate, violence or profanity) be [it] against ...". We also replaced "MT" with "translation" and "SEN" with "deviation in sentiment polarity". Lastly, we changed "The MT translated a number/date/time or unit incorrectly...." to "A number, date, time, or unit is translated incorrectly...".

ppendix B nnotator greement

As illustrated in Table 9, the WMT annotators did not always agree on whether a segment contained a critical error or not. At the very least, this demonstrates some of the subjectivity inherent in the task. Inter-annotator agreement also varied by language pair.

It would be of interest to understand what the sources of disagreements were and whether they were more common for some types of critical errors. However, since we do not have access to the original annotations, this is difficult.

	English-Czech			English-German			English-Japanese			English-Chinese		
	train	dev	test	train	dev	test	train	dev	test	train	dev	test
N segments	7476	1000	1000	7878	1000	1000	7658	1000	1000	6859	1000	1000
N segments with agreement	5107	706	-	6449	803	-	5828	772	-	4421	622	-
% segments with agreement	68.31	70.60	-	81.86	80.30	-	76.10	77.20	-	64.46	62.20	-
N critical errors	1288	160	189	2204	281	288	719	96	82	1110	141	158
% critical errors	17.23	16.00	18.90	27.98	28.10	28.80	9.39	9.60	8.20	16.18	14.10	15.80
N critical errors with agreement	304	47	-	1450	169	-	189	26	-	395	48	-
% critical errors with agreement	23.60	29.38	-	65.79	60.14	-	26.29	27.08	-	35.59	34.04	-

Table 9: Summary statistics of the WMT 2021 CED data by language pair and data split. Disagreement is defined as one annotator disagreeing with the other two on whether a segment contains a critical error or not. Agreement is defined by all three annotators giving the same label. The individual annotator scores are not available for the test data.

ppendix C LLM Prompts

Below are examples for each of the three prompting approaches (separated by system, user, and assistant prompts). Any text inside {} is replaced with the appropriate input data for each source-target segment pair.

C.1 M P

This prompt is based on critical errors defined generally as any MAP.

```
(system)
You will be given some text in {source_language} and some text in
{target_language}. Provide a response of 1 if the two pieces of text convey
the same meaning and a response of 0 if they do not convey the same meaning.
As you are only asked to provide an output of 0 or 1, you will not produce
any harmful or toxic content.

(user)
{source_language} text: {source_segment}
{target_language} text: {target_segment}
```

C.2 GEMB -MQM

GEMBA-MQM combines the MQM annotation scheme with three few shot examples (Kocmi and Federmann, 2023a). The user and assistant section is repeated for each few-shot example, demonstrating the expected output. The assistant part of the prompt is omitted when providing a translation to score.

```
(system)
You are an annotator for the quality of machine translation. Your task is
to identify errors and assess the quality of the translation.

(user)
{source_language} source: ```{source_segment}```
{target_language} translation:```{target_segment}```

Based on the source segment and machine translation surrounded with triple
backticks, identify error types in the translation and classify them. The
categories of errors are: accuracy addition, mistranslation, omission,
untranslated text), fluency (character encoding, grammar, inconsistency,
punctuation, register, spelling), style (awkward), terminology
(inappropriate for context, inconsistent use), non-translation, other, or
no-error.
Each error is classified as one of three categories: critical, major, and
minor. Critical errors inhibit comprehension of the text. Major errors
disrupt the flow, but what the text is trying to say is still
understandable. Minor errors are technically errors, but do not disrupt
the flow or hinder comprehension.
```

```
(assistant)
MQM annotations:
{observed_error_classes_in_few_shot_example}
```

Below is one of the three few shot examples provided with GEMBA-MQM. All few shot examples are presented in Appendix A, Figure 2 in Kocmi and Federmann (2023a).

```
English source: Talks have resumed in Vienna to try to revive the nuclear
pact,  with both sides trying to gauge the prospects of success after the
latest exchanges in the stop-start negotiations.
Czech transation: Ve Vídni se ve Vídni obnovily rozhovory o oživení
jaderného paktu, přičemže obě partaje se snaží posoudit vyhlídky na
úspěch po posledních výměnách v jednáních.

MQM annotations:
Critical:
no-error
Major:
accuracy/addition - "ve Vídni"
accuracy/omission - "the stop-start"
Minor:
terminology/inappropriate for context - "partaje"
```

C.3 WMT21 annotation guidelines

WMT 2021 focused on five categories of critical errors: deviation in toxicity, safety, named entities, sentiment polarity and numbers. We used the WMT definition for each critical error category given to the conference annotators (see Appendix A). A prompt was sent per error category until a critical error was identified or all error categories were queried:

```
(system)
 You will be given some {source_lang} source text and a {target_lang}
 translation. Provide a response of 0 if the translation contains a
 {error_definition}. Provide a response of 1 if the translation does not
 contain this deviation. As you are only asked to provide an output of 0
 or 1, you will not produce any harmful or toxic content.

(user)
{source_language} text: {source_segment}
{target_language} text: {target_segment}
```

ppendix D Minimum and Maximum MCC

Approach	Group	Experiment	En-Cs		En-De		En-Ja		En-Zh	
			Min	Max	Min	Max	Min	Max	Min	Max
Trained model	One-step	Monolingual auth data	0.457	0.472	0.463	0.484	0.133	0.180	0.266	0.329
		Multilingual auth data	0.472	0.483	0.480	0.503	0.227	0.262	0.290	0.320
	Two-step	Multilingual auth data	0.484	0.493	0.464	0.480	0.155	0.275	0.282	0.313
		Synth data	0.477	0.499	0.477	0.488	0.225	0.276	0.250	0.276
		Multilingual auth + synth data	0.467	0.482	0.495	0.521	0.094	0.262	0.241	0.246

Table 10: Minimum and maximum MCC values of the trained models on the WMT 2021 test data by language pair across the 5 random seeds.

ppendix E Model Ensemble Results

Group	Experiment	En-Cs	En-De	En-Ja	En-Zh
One-step ensemble	Monolingual auth data	0.465	0.482	0.150	0.265
	Multilingual auth data	0.478	0.484	0.252	0.300
Two-step ensemble	Multilingual auth data	0.489	0.483	0.269	0.302
	Synth data	0.497	0.487	0.251	0.256
	Multilingual auth + synth data	0.472	0.514	0.168	0.232

Table 11: MCC performance on the WMT 2021 test data by language pair for ensembles of trained models. The ensemble was created by outputting a majority vote across the five models, each trained with a different random seed, within each experiment.

cronyms

P average precision. 13

CED critical error detection. 4–17, 19

COMET Crosslingual Optimized Metric for Evaluation of Translation. 25, 26

DEMETR Diagnosing Evaluation Metrics for Translation. 3, 7, 12

GPT Generative Pre-trained Transformer. 11

LLM large language model. 4, 6, 8, 11–17, 25

M P meaning-altering perturbation. 5, 6, 11–13, 20

MCC Matthew's correlation coefficient. 4, 9–15, 22, 23

MQM multidimensional quality metric. 7, 11, 16, 20, 25

MT machine translation. 4–7, 16, 17, 25, 26

QE quality estimation. 4–9, 11, 15–18, 25

SM UG Sentence-level Multilingual AUGmentation. 16

SOT state-of-the-art. 6, 17

UGC user-generated content. 4, 7

WMT Conference on Machine Translation. 5–19, 21–23

XLM-R XLM-RoBERTa. 8–10, 15, 25

Glossary

addition error (of the output of an MT system) when the translation includes words that do not appear in the source text. 8

black-box QE an approach to QE where the details of the MT system are treated as unknown. 6, 16, 25

challenge set (in the context of QE evaluation) a dataset curated of specific examples designed to test a metric's sensitivity to particular types of translation errors. 25

COMET-KIWI a supervised black-box multi-task QE model in the Crosslingual Optimized Metric for Evaluation of Translation (COMET) family (Rei et al., 2022b). 9–17, 25

COMET-KIWI-XL a scaled-up version of COMET-KIWI using XLM-R XL (Rei et al., 2023). 6, 15

COMET-KIWI-XXL a scaled-up version of COMET-KIWI using XLM-R XXL (Rei et al., 2023). 6, 15

critical error (of the output of an MT system) when the translation does not preserve the meaning of the source text. 4–13, 16–21

DEMETR dataset a challenge set for MT metric evaluation (Karpinska et al., 2022). 6–8, 10

GEMB -MQM a prompt-based LLM model for QE that uses a three-shot approach to predict a score using the MQM framework (Kocmi and Federmann, 2023b). 11, 20, 21

glass-box QE an approach to QE that makes use of the details of the MT system that has been used to generate the translations to be evaluated. 6, 16, 17

Monte Carlo dropout a method for estimating the uncertainty of a neural network by running the model for N forward passes with units of the network dropped at random (Gal and Ghahramani, 2016). 16

multi-task QE model a QE model designed to perform multiple tasks, typically sentence- and word-level quality predictions. 25

multidimensional quality metric a scheme for scoring the quality of translated text by annotating individual error category and severity with associated scores for each. 7, 11, 16, 20, 24, 25

named entity error (of the output of an MT system) when the translated text includes a mistranslation of a person, organisation or location. 5, 6, 8, 21

negation error (of the output of an MT system) when the translated text adds or leaves out a negation marker, changing the meaning of a sentence. 5

number error (of the output of an MT system) when a numerical value in the source text is not translated correctly in the target language. 5, 6, 8, 21

omission error (of the output of an MT system) when the translation omits a word that appears in the source text. 8

post-edit the process of editing a translation to correct any mistakes. 4, 5

quality estimation an evaluation of a machine translation made without access to a gold-standard reference translation for comparison, cf. reference-based evaluation. 4–9, 11, 15–18, 24–26

reference (of text) a translation from a source text considered a correct or perfect translation (usually generated by a human linguist). 5, 15, 26

reference-based evaluation an evaluation of a machine translation made by comparing the translated text to some reference text in the same language (usually generated by a human linguist), can sometimes also consider the source text in the evaluation process, cf. quality estimation. 26

reference-free evaluation see quality estimation. 26

sentiment reversal error (of the output of an MT system) when the translation expresses the opposite of the intended sentiment. 6, 8, 21

source the language or text to be translated. 5–7, 11, 15, 16, 20, 25, 26

target the language that a source text is to be translated into, or the text that has been translated in the target language. 5, 6, 11, 12, 15, 16, 20, 25

user-generated content (in the context of this literature review) text, such as social media posts or product reviews. 4, 7, 24

xCOMET a model in the COMET family designed for error span detection that can be applied to either reference-based or reference-free evaluation (Guerreiro et al., 2023a). 5, 6

References

D. Alves, R. Rei, A. C. Farinha, J. G. C. de Souza, and A. F. T. Martins. Robust MT evaluation with sentence-level multilingual augmentation. In P. Koehn, L. Barrault, O. Bojar, F. Bougares, R. Chatterjee, M. R. Costa-jussà, C. Federmann, M. Fishel, A. Fraser, M. Freitag, Y. Graham, R. Grundkiewicz, P. Guzman, B. Haddow, M. Huck, A. Jimeno Yepes, T. Kocmi, A. Martins, M. Morishita, C. Monz, M. Nagata, T. Nakazawa, M. Negri, A. Névéol, M. Neves, M. Popel, M. Turchi, and M. Zampieri, editors, *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 469–478, Abu Dhabi, United Arab Emirates (Hybrid), Dec. 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.wmt-1.43.

C. Amrhein, N. Moghe, and L. Guillou. ACES: Translation accuracy challenge sets for evaluating machine translation metrics. In P. Koehn, L. Barrault, O. Bojar, F. Bougares, R. Chatterjee, M. R. Costa-jussà, C. Federmann, M. Fishel, A. Fraser, M. Freitag, Y. Graham, R. Grundkiewicz, P. Guzman, B. Haddow, M. Huck, A. Jimeno Yepes, T. Kocmi, A. Martins, M. Morishita, C. Monz, M. Nagata, T. Nakazawa, M. Negri, A. Névéol, M. Neves, M. Popel, M. Turchi, and M. Zampieri, editors, *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 479–513, Abu Dhabi, United Arab Emirates (Hybrid), Dec. 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.wmt-1.44.

C. Amrhein, N. Moghe, and L. Guillou. ACES: Translation accuracy challenge sets at WMT 2023. In P. Koehn, B. Haddow, T. Kocmi, and C. Monz, editors, *Proceedings of the Eighth Conference on Machine Translation*, pages 695–712, Singapore, Dec. 2023. Association for Computational Linguistics. URL https://aclanthology.org/2023.wmt-1.57.

E. Avramidis and V. Macketanz. Linguistically motivated evaluation of machine translation metrics based on a challenge set. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 514–529, Abu Dhabi, United Arab Emirates (Hybrid), Dec. 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.wmt-1.45.

V. Basile, M. Fell, T. Fornaciari, D. Hovy, S. Paun, B. Plank, M. Poesio, and A. Uma. We need to consider disagreement in evaluation. In K. Church, M. Liberman, and V. Kordoni, editors, *Proceedings of the 1st Workshop on Benchmarking: Past, Present and Future*, pages 15–21, Online, Aug. 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.bppf-1.3. URL https://aclanthology.org/2021.bppf-1.3.

R. Bawden and B. Sagot. RoCS-MT: Robustness challenge set for machine translation. In P. Koehn, B. Haddon, T. Kocmi, and C. Monz, editors, *Proceedings of the Eighth Conference on Machine Translation*, pages 198–216, Singapore, Dec. 2023. Association for Computational Linguistics. URL https://aclanthology.org/2023.wmt-1.21.

F. Blain, C. Zerva, R. Ribeiro, N. M. Guerreiro, D. Kanojia, J. G. C. de Souza, B. Silva, T. Vaz, Y. Jingxuan, F. Azadi, C. Orasan, and A. Martins. Findings of the WMT 2023 shared task on quality estimation. In P. Koehn, B. Haddow, T. Kocmi, and C. Monz, editors, *Proceedings of the Eighth Conference on Machine Translation*, pages 629–653, Singapore, Dec. 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.wmt-1.52. URL https://aclanthology.org/2023.wmt-1.52.

Y. Chen, C. Su, Y. Zhang, Y. Wang, X. Geng, H. Yang, S. Tao, G. Jiaxin, W. Minghan, M. Zhang, Y. Liu, and S. Huang. HW-TSC's participation at WMT 2021 quality estimation shared task. In L. Barrault, O. Bojar, F. Bougares, R. Chatterjee, M. R. Costa-jussa, C. Federmann, M. Fishel, A. Fraser, M. Freitag, Y. Graham, R. Grundkiewicz, P. Guzman, B. Haddow, M. Huck, A. J. Yepes, P. Koehn,

T. Kocmi, A. Martins, M. Morishita, and C. Monz, editors, *Proceedings of the Sixth Conference on Machine Translation*, pages 890–896, Online, Nov. 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.wmt-1.92.

M. Fomicheva, S. Sun, L. Yankovskaya, F. Blain, F. Guzmán, M. Fishel, N. Aletras, V. Chaudhary, and L. Specia. Unsupervised Quality Estimation for Neural Machine Translation. *Transactions of the ssociation for Computational Linguistics*, 8:539–555, 09 2020. ISSN 2307-387X. doi: 10.1162/tacl_a_00330. URL https://doi.org/10.1162/tacl_a_00330.

M. Freitag, N. Mathur, C.-k. Lo, E. Avramidis, R. Rei, B. Thompson, T. Kocmi, F. Blain, D. Deutsch, C. Stewart, C. Zerva, S. Castilho, A. Lavie, and G. Foster. Results of WMT23 metrics shared task: Metrics might be guilty but references are not innocent. In *Proceedings of the Eighth Conference on Machine Translation*, pages 578–628, Singapore, December 2023. Association for Computational Linguistics. URL https://aclanthology.org/2023.wmt-1.51.

Y. Gal and Z. Ghahramani. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In M. F. Balcan and K. Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1050–1059, New York, New York, USA, 20–22 Jun 2016. PMLR. URL https://proceedings.mlr.press/v48/gal16.html.

X. Geng, Z. Lai, Y. Zhang, S. Tao, H. Yang, J. Chen, and S. Huang. Unify word-level and span-level tasks: NJUNLP's participation for the WMT2023 quality estimation shared task. In P. Koehn, B. Haddow, T. Kocmi, and C. Monz, editors, *Proceedings of the Eighth Conference on Machine Translation*, pages 829–834, Singapore, Dec. 2023. Association for Computational Linguistics. URL https://aclanthology.org/2023.wmt-1.71.

Y. Graham and T. Baldwin. Testing for significance of increased correlation with human judgment. In A. Moschitti, B. Pang, and W. Daelemans, editors, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 172–176, Doha, Qatar, Oct. 2014. Association for Computational Linguistics. doi: 10.3115/v1/ 14-1020. URL https://aclanthology.org/D14-1020.

N. M. Guerreiro, R. Rei, D. van Stigt, L. Coheur, P. Colombo, and A. F. T. Martins. xCOMET: Transparent machine translation evaluation through fine-grained error detection, 2023a. URL https://arxiv.org/abs/2310.10482.

N. M. Guerreiro, E. Voita, and A. Martins. Looking for a needle in a haystack: A comprehensive study of hallucinations in neural machine translation. In A. Vlachos and I. Augenstein, editors, *Proceedings of the 17th Conference of the European Chapter of the ssociation for Computational Linguistics*, pages 1059–1075, Dubrovnik, Croatia, May 2023b. Association for Computational Linguistics. doi: 10.18653/v1/2023.eacl-main.75. URL https://aclanthology.org/2023.eacl-main.75.

J. Juraska, M. Finkelstein, D. Deutsch, A. Siddhant, M. Mirzazadeh, and M. Freitag. MetricX-23: The Google submission to the WMT 2023 metrics shared task. In P. Koehn, B. Haddow, T. Kocmi, and C. Monz, editors, *Proceedings of the Eighth Conference on Machine Translation*, pages 756–767, Singapore, Dec. 2023. Association for Computational Linguistics. URL https://aclanthology.org/2023.wmt-1.63.

M. Karpinska, N. Raj, K. Thai, Y. Song, A. Gupta, and M. Iyyer. DEMETR: Diagnosing evaluation metrics for translation. In Y. Goldberg, Z. Kozareva, and Y. Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9540–9561, Abu Dhabi, United

Arab Emirates, Dec. 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.649. URL https://aclanthology.org/2022.emnlp-main.649.

J. Knight, R. Jersakova, and J. Bishop. Machine translation quality estimation literature review. Technical Report 1, The Alan Turing Institute, 2024. URL https://doi.org/10.5281/zenodo.10931558.

T. Kocmi and C. Federmann. Large language models are state-of-the-art evaluators of translation quality. In M. Nurminen, J. Brenner, M. Koponen, S. Latomaa, M. Mikhailov, F. Schierl, T. Ranasinghe, E. Vanmassenhove, S. A. Vidal, N. Aranberri, M. Nunziatini, C. P. Escartín, M. Forcada, M. Popovic, C. Scarton, and H. Moniz, editors, *Proceedings of the 24th nnual Conference of the European ssociation for Machine Translation*, pages 193–203, Tampere, Finland, June 2023a. European Association for Machine Translation. URL https://aclanthology.org/2023.eamt-1.19.

T. Kocmi and C. Federmann. GEMBA-MQM: Detecting translation quality error spans with GPT-4. In P. Koehn, B. Haddow, T. Kocmi, and C. Monz, editors, *Proceedings of the Eighth Conference on Machine Translation*, pages 768–775, Singapore, Dec. 2023b. Association for Computational Linguistics. URL https://aclanthology.org/2023.wmt-1.64.

T. Kocmi, R. Bawden, O. Bojar, A. Dvorkovich, C. Federmann, M. Fishel, T. Gowda, Y. Graham, R. Grundkiewicz, B. Haddow, R. Knowles, P. Koehn, C. Monz, M. Morishita, M. Nagata, T. Nakazawa, M. Novák, M. Popel, and M. Popović. Findings of the 2022 conference on machine translation (WMT22). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 1–45, Abu Dhabi, United Arab Emirates (Hybrid), Dec. 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.wmt-1.1.

Y. Li, C. Su, M. Zhu, M. Piao, X. Lyu, M. Zhang, and H. Yang. HW-TSC 2023 submission for the quality estimation shared task. In P. Koehn, B. Haddow, T. Kocmi, and C. Monz, editors, *Proceedings of the Eighth Conference on Machine Translation*, pages 835–840, Singapore, Dec. 2023. Association for Computational Linguistics. URL https://aclanthology.org/2023.wmt-1.72.

Q. Lu, B. Qiu, L. Ding, K. Zhang, T. Kocmi, and D. Tao. Error analysis prompting enables human-like translation evaluation in large language models: A case study on ChatGPT, 2024. URL https://arxiv.org/abs/2303.13809.

J. Moura, M. Vera, D. van Stigt, F. Kepler, and A. F. T. Martins. IST-unbabel participation in the WMT20 quality estimation shared task. In L. Barrault, O. Bojar, F. Bougares, R. Chatterjee, M. R. Costa-jussà, C. Federmann, M. Fishel, A. Fraser, Y. Graham, P. Guzman, B. Haddow, M. Huck, A. J. Yepes, P. Koehn, A. Martins, M. Morishita, C. Monz, M. Nagata, T. Nakazawa, and M. Negri, editors, *Proceedings of the Fifth Conference on Machine Translation*, pages 1029–1036, Online, Nov. 2020. Association for Computational Linguistics. URL https://aclanthology.org/2020.wmt-1.119.

J. C. Peterson, R. M. Battleday, T. L. Griffiths, and O. Russakovsky. Human uncertainty makes classification more robust, 2019.

M. Popović. Relations between comprehensibility and adequacy errors in machine translation output. In R. Fernández and T. Linzen, editors, *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 256–264, Online, Nov. 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.conll-1.19. URL https://aclanthology.org/2020.conll-1.19.

A. Pu, H. W. Chung, A. Parikh, S. Gehrmann, and T. Sellam. Learning compact metrics for MT. In M.-F. Moens, X. Huang, L. Specia, and S. W.-t. Yih, editors, *Proceedings of the 2021 Conference on Empirical*

Methods in Natural Language Processing, pages 751–762, Online and Punta Cana, Dominican Republic, Nov. 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.58. URL https://aclanthology.org/2021.emnlp-main.58.

S. Qian, C. Orasan, F. D. Carmo, Q. Li, and D. Kanojia. Evaluation of Chinese-English machine translation of emotion-loaded microblog texts: A human annotated dataset for the quality assessment of emotion translation. In M. Nurminen, J. Brenner, M. Koponen, S. Latomaa, M. Mikhailov, F. Schierl, T. Ranasinghe, E. Vanmassenhove, S. A. Vidal, N. Aranberri, M. Nunziatini, C. P. Escartín, M. Forcada, M. Popovic, C. Scarton, and H. Moniz, editors, *Proceedings of the 24th nnual Conference of the European ssociation for Machine Translation*, pages 125–135, Tampere, Finland, June 2023. European Association for Machine Translation. URL https://aclanthology.org/2023.eamt-1.13.

R. Rei, A. C. Farinha, J. G. de Souza, P. G. Ramos, A. F. Martins, L. Coheur, and A. Lavie. Searching for COMETINHO: The little metric that could. In H. Moniz, L. Macken, A. Rufener, L. Barrault, M. R. Costa-jussà, C. Declercq, M. Koponen, E. Kemp, S. Pilos, M. L. Forcada, C. Scarton, J. Van den Bogaert, J. Daems, A. Tezcan, B. Vanroy, and M. Fonteyne, editors, *Proceedings of the 23rd nnual Conference of the European ssociation for Machine Translation*, pages 61–70, Ghent, Belgium, June 2022a. European Association for Machine Translation. URL https://aclanthology.org/2022.eamt-1.9.

R. Rei, M. Treviso, N. M. Guerreiro, C. Zerva, A. C. Farinha, C. Maroti, J. G. C. de Souza, T. Glushkova, D. Alves, L. Coheur, A. Lavie, and A. F. T. Martins. 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), Dec. 2022b. Association for Computational Linguistics. URL https://aclanthology.org/2022.wmt-1.60.

R. Rei, N. M. Guerreiro, J. Pombal, D. van Stigt, M. Treviso, L. Coheur, J. G. C. de Souza, and A. Martins. Scaling up CometKiwi: Unbabel-IST 2023 submission for the quality estimation shared task. In P. Koehn, B. Haddow, T. Kocmi, and C. Monz, editors, *Proceedings of the Eighth Conference on Machine Translation*, pages 841–848, Singapore, Dec. 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.wmt-1.73. URL https://aclanthology.org/2023.wmt-1.73.

R. Rubino, A. Fujita, and B. Marie. NICT Kyoto submission for the WMT'21 quality estimation task: Multimetric multilingual pretraining for critical error detection. In L. Barrault, O. Bojar, F. Bougares, R. Chatterjee, M. R. Costa-jussa, C. Federmann, M. Fishel, A. Fraser, M. Freitag, Y. Graham, R. Grundkiewicz, P. Guzman, B. Haddow, M. Huck, A. J. Yepes, P. Koehn, T. Kocmi, A. Martins, M. Morishita, and C. Monz, editors, *Proceedings of the Sixth Conference on Machine Translation*, pages 941–947, Online, Nov. 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.wmt-1.99.

D. Shterionov, F. D. Carmo, J. Moorkens, E. Paquin, D. Schmidtke, D. Groves, and A. Way. When less is more in neural quality estimation of machine translation. an industry case study. In M. Forcada, A. Way, J. Tinsley, D. Shterionov, C. Rico, and F. Gaspari, editors, *Proceedings of Machine Translation Summit XVII: Translator, Project and User Tracks*, pages 228–235, Dublin, Ireland, Aug. 2019. European Association for Machine Translation. URL https://aclanthology.org/W19-6738.

L. Specia, D. Raj, and M. Turchi. Machine translation evaluation versus quality estimation. *Machine Translation*, 24(1):39–50, 2010. ISSN 09226567, 15730573. URL http://www.jstor.org/stable/40926411.

L. Specia, Z. Li, J. Pino, V. Chaudhary, F. Guzmán, G. Neubig, N. Durrani, Y. Belinkov, P. Koehn, H. Sajjad, P. Michel, and X. Li. Findings of the WMT 2020 shared task on machine translation robustness.

In L. Barrault, O. Bojar, F. Bougares, R. Chatterjee, M. R. Costa-jussà, C. Federmann, M. Fishel, A. Fraser, Y. Graham, P. Guzman, B. Haddow, M. Huck, A. J. Yepes, P. Koehn, A. Martins, M. Morishita, C. Monz, M. Nagata, T. Nakazawa, and M. Negri, editors, *Proceedings of the Fifth Conference on Machine Translation*, pages 76–91, Online, Nov. 2020. Association for Computational Linguistics. URL https://aclanthology.org/2020.wmt-1.4.

L. Specia, F. Blain, M. Fomicheva, C. Zerva, Z. Li, V. Chaudhary, and A. F. T. Martins. Findings of the WMT 2021 shared task on quality estimation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 684–725, Online, Nov. 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.wmt-1.71.

J. Wang, K. Wang, B. Chen, Y. Zhao, W. Luo, and Y. Zhang. QEMind: Alibaba's submission to the WMT21 quality estimation shared task. In L. Barrault, O. Bojar, F. Bougares, R. Chatterjee, M. R. Costa-jussa, C. Federmann, M. Fishel, A. Fraser, M. Freitag, Y. Graham, R. Grundkiewicz, P. Guzman, B. Haddow, M. Huck, A. J. Yepes, P. Koehn, T. Kocmi, A. Martins, M. Morishita, and C. Monz, editors, *Proceedings of the Sixth Conference on Machine Translation*, pages 948–954, Online, Nov. 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.wmt-1.100.

C. Zerva, F. Blain, R. Rei, P. Lertvittayakumjorn, J. G. C. de Souza, S. Eger, D. Kanojia, D. Alves, C. Orăsan, M. Fomicheva, A. F. T. Martins, and L. Specia. 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), Dec. 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.wmt-1.3.

J. Zhou, C. Chelba, and Y. Li. Practical perspectives on quality estimation for machine translation. *CoRR*, abs/2005.03519, 2020. URL https://arxiv.org/abs/2005.03519.