Findings of the WMT 2022 Shared Task on Quality Estimation

Chrysoula Zerva[1,2], Frédéric Blain[3], Ricardo Rei[2,4,5], Piyawat Lertvittayakumjorn[6], José G. C. de Souza[4], Steffen Eger[9], Diptesh Kanojia[8], Duarte Alves[2], Constantin Orăsan[8], Marina Fomicheva[7], André F. T. Martins[1,2,4] and Lucia Specia[6,7]

[1]Instituto de Telecomunicações, [2]Instituto Superior Técnico, [3]University of Wolverhampton
[4]Unbabel, [5]INESC-ID, [6]Imperial College London, [7]University of Sheffield
[8]University of Surrey, [9]NLLG,Technische Fakultät, Bielefeld University

f.blain@wlv.ac.uk,m.fomicheva@sheffield.ac.uk, {d.kanojia, c.orasan}@surrey.ac.uk
{chrysoula.zerva,ricardo.rei,duartemalves}@tecnico.ulisboa.pt, pl1515@ic.ac.uk

Abstract

We report the results of the WMT 2022 shared task on Quality Estimation, in which the challenge is to predict the quality of the output of neural machine translation systems at the word and sentence levels, without access to reference translations. This edition introduces a few novel aspects and extensions that aim to enable more fine-grained, and explainable quality estimation approaches. We introduce an updated quality annotation scheme using Multidimensional Quality Metrics to obtain sentence- and word-level quality scores for three language pairs. We also extend the Direct Assessments and post-edit data (MLQE-PE) to new language pairs: we present a novel and large dataset on English-Marathi, as well as a zero-shot test-set on English-Yoruba. Further, we include an explainability sub-task for all language pairs and present a new format of a critical error detection task for two new language pairs. Participants from 11 different teams submitted altogether 991 systems to different task variants and language pairs.

1 Introduction

The 11th edition of the shared task on Quality Estimation (QE) builds on its previous editions and findings to further benchmark methods for estimating the quality of neural machine translation (MT) output at run-time, without the use of reference translations. It includes (sub)tasks that consider quality of machine translations at the word and sentence levels.

Over the past years, the QE field has been moving towards trainable, large, multilingual models that have been shown to achieve high performance, especially at sentence-level (Specia et al., 2021). In this edition, we further expand the provided resources, introducing new low-resource language pairs: a large dataset of English-Marathi, suitable for training, development and testing and a smaller test-set on English-Yoruba for zero-shot approaches. These, as well as previously published datasets for QE, rely mainly on Direct Assessments (DA)[1] and post-edited translations, which provide estimates of quality either by using the human quality score(s) for each segment or by estimating the distance of a translation from a human-provided correction. As these annotations can sometimes obscure the exact location and/or significance of a translation error, we wanted to investigate the feasibility and efficiency of using a more fine-grained annotation schema to obtain quality estimations at word- and sentence- level, namely Multidimensional Quality Metrics (MQM) (Lommel et al., 2014). MQM annotations have shown to be more trustworthy for the metrics task (Freitag et al., 2021a,b), motivating us to evaluate their suitability for the QE task. We make available new development and test data on three language pairs using MQM annotations.

The aforementioned boost in performance of QE systems frequently comes at the cost of efficiency and interpretability, since they heavily rely on large models with many parameters. As a result, the predicted quality estimates are hard to interpret. At the same time, such high-performance, "black-box" models are frequently susceptible to systematic errors, such as negation omission (Kanojia et al., 2021) and mistranslated entities (Amrhein and Sennrich, 2022). Both phenomena are major concerns for MT quality estimation since they can undermine users' trust in new technologies and hamper the adoption of such models on a wide scale. To motivate approaches that address these cases we include an explainability subtask following its first edition at Eval4NLP 2021 (Fomicheva et al., 2021). In this subtask we ask participants to predict

[1]We note that the procedure followed for our data diverges from that proposed by Graham et al. (2016) in three ways: (a) we employ fewer but professional translators to score each sentence, (b) scoring is done against the source segment (bilingual annotation) and not the reference, and (c) we provide translators with guidelines on the meaning of ranges of scores.

Proceedings of the Seventh Conference on Machine Translation (WMT), pages 69–99
Abu Dhabi, December 7–8, 2022.

the erroneous words as rationale extraction for a sentence-level quality estimate, without any word-level supervision. By framing error identification as rationale extraction for sentence-level quality estimation systems, this subtask offers an opportunity to study whether such systems behave in the same way as humans would do. We also reshape the critical error detection task of last year and we build a new corpus to test the ability of QE systems to detect critical errors that simulate hallucinated content with additions, deletions, named entities, polarity changes and numbers. The corpus is created using SMAUG (Alves et al., 2022) and we allow participation in constrained and unconstrained settings. For the constrained setting, participants have to build QE systems without having access to data from SMAUG, whereas participants from the unconstrained task can train their systems using additional data from SMAUG.

In addition to advancing the state-of-the-art at all prediction levels, our main goals are:

- To extend the languages covered in our datasets;
- To further motivate fine-grained quality annotation, informed at word and sentence level using MQM;
- To encourage language-independent and even unsupervised approaches especially for zero-shot prediction;
- To study and promote explainable approaches for MT evaluation; and
- To revisit critical error detection.

We thus have three tasks:

Task 1 The core QE task, consisting of separate sentence-level and word-level subtasks. For the sentence-level sub-tasks, the goal is to predict a quality score for each segment in the test set, which can be a variant of DA (§2.1.1) or MQM (§2.1.1). For the word-level subtasks, participants have to predict translation errors at word-level, via binary quality tags (see §2.1.2).

Task 2 Explainable QE task, aiming to obtain word-level rationales for sentence-level quality scores (§2.2).

Task 3 The critical Error Detection task, aiming to predict sentence-level binary scores indicating whether or not a translation contains a critical error (§2.3).

The tasks make use of large datasets annotated by professional translators with either 0-100 DA scoring, post-editing or MQM annotations. We update the training and development datasets of previous editions and provide new test sets for Tasks 1 and 2. Additionally, we provide a novel setup for Task 3, with novel train, development and test data. The datasets and models released are publicly available[2]. Participants are also allowed to explore any additional data and resources deemed relevant, across tasks.

The shared task uses CodaLab as submission platform, where participants (Section 4) could submit up to 2 submissions a day for each task and language pair (LP), up to a total of 10 submissions. Results for all tasks evaluated according to standard metrics are given in Section 5. Baseline systems were trained by the task organisers and entered in the platform to provide a basis for comparison (Section 3). A discussion on the main goals and findings from this year's task is presented in Section 6.

2 Quality Estimation tasks

In what follows, we briefly describe each subtask, including the datasets provided for them.

2.1 Task 1: Predicting translation quality

Being able to automatically predict the quality of translations on sentence- or word-level without access to human-references is the core goal of the QE shared task. In this edition, we explored some new approaches towards quality annotations for sentence- and word-level, and redefined the word-level quality labelling scheme, in an attempt to allow participants to employ multi-task approaches and exploit fine-grained quality annotations. Hence, the data was produced in two ways:

1. DA & Post-edit approach: The quality of each source-translation pair is annotated by at least 3 independent expert annotators, using DA on a scale 0-100. The translation is also post-edited to obtain the closest possible, fully correct translation of the source. Using the post-edited data, we generate Human-mediated

[2] https://github.com/WMT-QE-Task/wmt-qe-2022-data

Translation Edit Rate (HTER) (Snover et al., 2006) scores, which are obtained by calculating the minimum edit distance between the machine translation and its manually post-edited version. By additionally considering the alignment between the source and post-edited sentence, we can propagate the errors to the source sentence and annotate the segments that were potentially mistranslated and/or not translated at all. The HTER scores were made available to participants as additional data, but are not used as prediction targets.

2. MQM approach: Each source-translation pair is evaluated by at least 1 expert annotator, and errors identified in text are highlighted and classified in terms of severity (minor, major, critical) and type (omission, style, mistranslation, etc).

The DA and MQM data was further processed to a) obtain normalised quality scores that have the same direction between high and low quality and b) obtain word-level binary quality labels. We provide more details on the required pre-processing in §2.1.1 and §2.1.2.

DA & Post-edit data: For all language pairs the data provided is selected from publicly available resources. Specifically for training we used the following language pairs from the MLQE-PE dataset (Fomicheva et al., 2022): English-German (En-De), English-Chinese (En-Zh), Russian-English (Ru-En), Romanian-English (Ro-En), Nepalese-English (Ne-En), Esthonian-English (Et-En) and Sinhala-English (Si-En), which are all sampled from Wikipedia, except for the Ru-En pair, which also contains sentences from Reddit. Additionally, the language-pairs used for development and testing also originate from Wikipedia: English-Czech (En-Cs), English-Japanese (En-Ja), Khmer-English (Km-En) and Pashto-English (Ps-En).

Finally, the new English-Marathi (En-Mr) data that is made available for train, development and testing this year is sampled from a combination of sources. More specifically the source side segments of the English-Marathi data contain segments from three different domains – healthcare, cultural, and general/news. The general domain and cultural domain data were obtained from the English (source side) segments in the IITB English-Hindi Parallel Corpus (Kunchukuttan et al., 2018). However, the healthcare domain data was obtained from publicly available NHS monolingual corpus[3].

All of the data was translated using large transformer-based NMT models, with established high performance for the languages in question. Specifically, for the language pairs in the training data (En-De, En-Zh, Et-En, Ne-En, Ru-En, Ro-En, Si-En), all source sentences were translated by a *fairseq* Transformer (Ott et al., 2019) bilingual model. The exception is the English-Marathi which was translated by the multilingual IndicTrans (En-X) Transformer-based NMT model, which was trained on the Samanantar parallel corpus (Ramesh et al., 2022).

For the languages provided in the development and test set, namely: En-Cz, En-Ja, Km-En and Ps-En we maintain the same we use the MBART50 (Tang et al., 2020),[4] to translate the source sentence of the other languages pairs, since it has been found to perform well, especially for low-resource languages (Tang et al., 2020). The En-Mr portion of the development and test data is translated similarly to the training data for this language pair.

Zero-shot language pair: This year we introduced a "surprise" language-pair, English-Yoruba (En-Yo), which represents a low-resource language pair. The Yoruba language is the third most spoken language in Africa, and it is native to southwestern Nigeria and the Republic of Benin (Eberhard et al., 2020). We extracted 1010 sentences in English from Wikipedia across 7 topics and translated them to Yoruba using Google Translate. Using adjusted guidelines from Fomicheva et al. (2021), we trained annotators to indicate sentence-level DA scores and to highlight erroneous words as word-level explanations for the DA scores.[5] On the 1010 sentences, they obtained agreements of 0.487 Pearson on sentence-level and 0.380 kappa on word-level. Note that in order to further encourage multilingual and unsupervised approaches, the setup for this zero-shot approach was slightly different to the previous edition, since we did not reveal the language pair before the release of the test data, and the zero-shot pair was included only in the multilingual sub-tasks for quality estimation

[3]The NHS corpus source sentences were crawled from the health directory of NHS available here: https://www.nhs.uk/conditions/

[4]https://github.com/pytorch/fairseq/tree/master/examples/multilingual

[5]Annotators were graduate students and native speakers of Yoruba and fluent in English.

(as opposed to a standalone subtask for this language pair only).

MQM data: As training data, we used annotations released for the Metrics shared task namely, the concatenation of the annotations released from Freitag et al. (2021a) with the annotations from last year Metrics task (Freitag et al., 2021b). Together, these annotations, cover 3 high-resource language pairs, namely: Chinese-English (Zh-En), English-German (En-De) and English-Russian (En-Ru), and span across two domains (News and Ted Talks). In contrast to DA, instead of one translation for each source, we have multiple translations coming from system participation's in the 2020 and 2021 News translation tasks (Barrault et al., 2020; Akhbardeh et al., 2021). For development set however, we follow an approach that is similar to the one use for the DA data: we translated the Newstest 2019 using a single NMT system, namely MBART50. Subsequently, for each language pair we asked an expert translator to provide MQM annotations. The test set was created similarly to the development, but instead of using Newstest 2019 we used the Newstest 2022 (the News data from this year's General MT shared task).

Overall, the released data for Task 1covers a total of 9 language pairs for training, 4 language pairs for development and 6 language pairs for testing including 1 zero-shot language pair. Statistics and details for each language pair are provided in Table 1.

2.1.1 Sentence-level quality prediction

There were two competition instances for the sentence-level sub-task. The first one focuses on DA- and the second one on MQM-derived annotations, both including a separate multilingual track. In the future, we aim to consolidate the competition instances into a single one for sentence-level, using our findings from this edition to align the annotation schemes in a better manner. We provide below the details for each annotation scheme and a comprehensive table with statistics for all annotations (Table 1).

DA annotations: For DA annotations, we followed the annotation and scoring conventions of previous editions. We provided MLQE-PE data (Fomicheva et al., 2022) used in previous years for training, which includes seven language pairs with $\approx$ 8,000 segments each. We also provided 26,000 segments of En-Mr which were annotated using the same annotation conventions. All translations were manually annotated for perceived quality, with a quality label ranging from 0 to 100, following the FLORES guidelines (Guzmán et al., 2019). According to the guidelines given to annotators, the 0-10 range represents an incorrect translation; 11-29, a translation with few correct keywords, but the overall meaning is different from the source; 30-50, a translation with major mistakes; 51-69, a translation which is understandable and conveys the overall meaning of the source but contains typos or grammatical errors; 70-90, a translation that closely preserves the semantics of the source sentence; and 91-100, a perfect translation. For each segment, there were at least three scores from independent raters (four in the case of En-Mr). DA scores were standardised using the z-score by rater, and the z-scores were provided as training targets. Participating systems are required to score sentences according to z-standardised DA scores.

MQM annotations: As we have seen (§2.1), for the MQM annotations, we built on the available Google MQM annotations (Freitag et al., 2021a) that contain annotated data for the En-De and Zh-En data of WMT 2020 News Translation Systems (Barrault et al., 2020) as well as En-De, Zh-En and En-Ru annotations from WMT Metrics 2021 (Freitag et al., 2021b). These annotations, provided as training data, amount to more than 30,000 segments in total (see Table 1 for details per language pair). In addition, we provide newly annotated development and test sets for all three language pairs (En-De, En-Ru, Zh-En), amounting to approximately 1,000 segments per language pair.

Originally, MQM annotated segments include annotated erroneous text-spans on the translation side that are assigned two types of labels: (a) an error severity label `{minor, major, critical}` and (b) an error category label such as `{grammar, style/awkward, omission, mistranslation}, ...}`. Each error severity is associated with a specific weight; hence a sentence score can be calculated for each segment based on these error weights. We demonstrate an example of MQM annotations and scores in Figure 1.

MQM scores according to Google weight scheme have the opposite direction of the DA scores since larger MQM scores denote worse translation quality, i.e., a larger number of errors or more severe errors. To address this inconsistency, we

Language Pairs	Sentences Train / Dev / Test22	Tokens Train / Dev / Test22	DA	PE	MQM	CE	Data Source
En-De [1]	9,000 / 1,000 / –	131,499 / 16,545 / –	✓	✓			Wikipedia
En-Zh	9,000 / 1,000 / –	131,892 / 16,637 / –	✓	✓			Wikipedia
Ru-En	9,000 / 1,000 / –	94,221 / 11,650 / –	✓	✓			Reddit
Ro-En	9,000 / 1,000 / –	137,466 / 17,359 / –	✓	✓			Wikipedia
Et-En	9,000 / 1,000 / –	112,503 / 14,044 / –	✓	✓			Wikipedia
Ne-En	9,000 / 1,000 / –	120,078 / 15,017 / –	✓	✓			Wikipedia
Si-En	9,000 / 1,000 / –	125,223 / 15,709 / –	✓	✓			Wikipedia
En-Mr	26,000 / 1,000 / 1,000	690,532 / 27,049 / 26,253	✓	✓			
Ps-En	– / 1,000 / 1,000	– / 27,045 / 27,414	✓	✓			Wikipedia
Km-En	– / 1,000 / 1,000	– / 21,981 / 22,048	✓	✓			Wikipedia
En-Ja	– / 1,000 / 1,000	– / 20,626 / 20,646	✓	✓			Wikipedia
En-Cs	– / 1,000 / 1,000	– / 20,394 /20,244	✓	✓			Wikipedia
En-Yo	– / – / 1,010	– / – / 21,238	✓	✓			
En-De [2]	28,909 / 1,005 / 511	839,473 / 24,373 / 13,220			✓		WMT-newstest
En-Ru	15,628 / 1,005 / 511	357,452 / 24,373 / 13,220			✓		WMT-newstest
Zh-En	35,327 / 1,019 / 505	1,586,883 / 51,969 / 15,602			✓		WMT-newstest
En-De	155,511 / 17,280 / 500	8,193,693 / 915,061 / 27,771				✓	News-Commentary
Pt-En	39,926 / 4,437 / 500	2,281,515 / 253,594 / 29,794				✓	News-Commentary

Table 1: Statistics of the data used for Task 1 (DA), Task 2 (PE) and Task 3 (CE) (last four rows). The number of tokens is computed based on the source sentences.

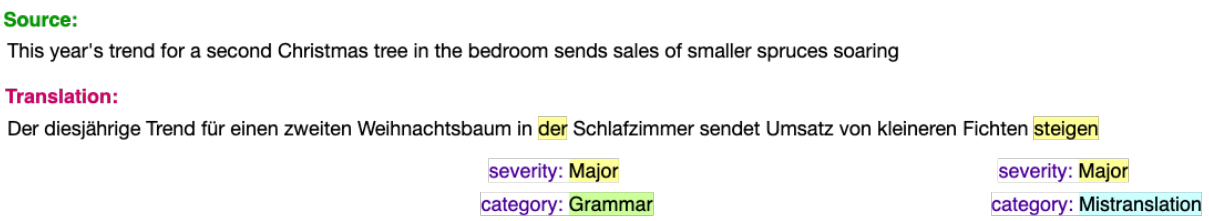

Figure 1: Example of MQM annotations on the target (translation) side, on a English–German (En-De) sentence pair.

invert the MQM scores and standardise per annotator. For training data we had access to multiple annotations per segment and calculated an average score after standardisation, keeping also the original MQM scores per annotator, to allow the participants to take full advantage of the different annotations (Basile et al., 2021). For the same reasons, we opted not to aggregate the annotated text-spans.

Regarding evaluation, systems in this task (both for DA and MQM) are **evaluated against the true z-normalised sentence scores using Spearman's rank correlation coefficient ρ as the primary metric**. This is what was used for ranking system submissions. Pearson's correlation coefficient, r, Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE) were also computed as secondary metrics but not used for the final ranking between systems.

2.1.2 Word-level quality prediction

This sub-task focuses on detecting word-level errors in the MT output. The goal is to automatically predict the quality of each token using a binary decision, i.e., using OK as a label for tokens translated correctly and BAD otherwise. We deviate from the annotation pattern of previous years in that, we do not consider annotations of the gaps between tokens or source-side annotations. Instead, to account for omission errors, we consider the following convention: the token on the right side of the omitted text in the translation is annotated as "BAD". An additional `<EOS>` token is appended at the end of every translation segment to account for omissions at the end of each sentence. This allows the provision of a unified framework for both the post-edit originated annotations and the MQM annotations.

We thus use the same source-translation pairs used for the sentence-level tasks and obtain the binary tags as follows:

- For post-edited data, we use TER (Snover et al., 2006) to obtain alignments between translation and post-edit and annotate the misaligned tokens as BAD.
- For MQM data, the tokens that fall within the text-spans annotated as errors (or any severity or category) are annotated as BAD. If the whitespace between two words is annotated as an error, then this is considered an omission, and the next token is annotated as BAD.

Participants were encouraged to submit for each language pair and also for the **multilingual variants** of each sub-task. For the DA-based sentence-level competition, as well as the word-level sub-task, there was an additional multilingual variant that included the zero-shot language pair (En-Yo). The latter aimed at fostering work on language-independent models, as well as models that are truly multilingual.

For word-level task, **submissions are ranked using the Matthews correlation coefficient (MCC) as the primary metric**, while F1-scores are provided as complementary information.

2.2 Task 2: Explainable Quality Estimation

Following the success of the shared task on Explainable Quality Estimation organized by the Eval4NLP workshop in 2021 (Fomicheva et al., 2021), in this sub-task we aim to address translation error identification as rationale extraction from sentence-level quality estimation systems. If a QE system reasonably estimates the quality of a translated sentence, an explanation extracted from the system should indicate word-level translation errors in the input (if any) as reasons for imperfect sentence-level scores. Particularly, for each input pair of source and target sentences, participating teams are asked to provide (*i*) a sentence-level score estimating the translation quality and (*ii*) a list of continuous word-level scores where the tokens with the highest scores are expected to correspond to translation errors considered relevant by human annotators.

In this explainable QE task, we use all the nine language pairs and their word-level test sets from Task 1 (see §2.1.2) with En-Yo being a separate language pair (rather than blending it in the multilingual test set). Therefore, the participants are allowed to use the sentence-level scores from the datasets in Task 1 to train their sentence-level models in Task 2. However, as Task 2 aims to promote the research on the explainability of QE systems, we encourage the participants to use or develop explanation methods to identify contributions of words or tokens in the input. Unlike Task 1, **the participants of Task 2 are not allowed to supervise their models with any token-level or word-level labels or signals (whether they are from natural or synthetic data) in order to directly predict word-level errors.** Consequently, we do not require the participants to convert their word-level scores into predicted binary labels (OK/BAD) since this process usually requires a word-level QE dataset to search for an optimum score threshold.

Concerning the evaluation of this task, we focus on assessing the quality of explanations (i.e., the submitted word-level scores), not the sentence-level predictions. Specifically, we measure how well the word-level scores provided by the participants correspond with human word-level error annotations, which are binary ground truth labels. Unlike the Eval4NLP 2021 shared task, which ranked participating systems by a combination of three metrics (Fomicheva et al., 2021), **we use *Recall at Top-K*, also known as *R-precision* in information retrieval literature (Manning et al., 2008, chapter 8), as the primary metric this year** due to two reasons. First, it is preferable to have a single main metric to avoid confusion and also some potential side effects that combining the three metrics might produce. Second, Recall at Top-K seemed to help discriminate best between the participating submissions in the Eval4NLP shared task. Assume that, for a given pair of source and target sentences, there are K words annotated as translation errors by humans. Recall at Top-K equals $\frac{r}{K}$ when there are r out of the K error words appearing in the list of top-K words ranked by the submitted word-level scores descendingly. In addition, AUC (an area under the receiver operating characteristic curve) and AP (average precision) are used as secondary metrics. Considering the word level, AUC summarises the curve between true positive rate and false positive rate, while AP summarises the curve between precision and recall. For both of the secondary metrics, higher values are the better. Although we report metrics for sentence-level predictions, including Pearson's correlation and Spearman's correlation, as additional information, we do not use them for ranking the participants or determining the winner in this explainability task.

2.3 Task 3: Critical Error Detection

In this sub-task, we reshape the binary classification task introduced in last year's edition (Specia et al., 2021) to predict whether the translated sentence contains (at least) one critical error.

Following Specia et al. (2021), we consider that a translation contains a critical error if it deviates from the meaning of the source sentence in such a way that it is misleading and may lead to several implications. As noted by Specia et al. (2021), deviations in meaning can happen in three ways: mistranslation errors have critical content translated incorrectly into a different meaning; hallucination errors introduce critical content in the translation that is not in the source; and deletion errors remove critical content that is in the source from the translation.

In this task, we focus on five critical error categories:

- Additions: The content of the translation is only partially supported by the source.
- Deletions: Part of the source sentence is ignored by the MT engine.
- Named Entities: A named entity (people, organization, location, etc.) is mistranslated into another incorrect named entity.
- Meaning: The translated sentence either introduces or removes a negation and the sentence meaning is completely reversed.
- Numbers: The MT system translates a number/date/time or unit incorrectly.

For this task, we introduce a new dataset obtained by perturbing a corpus of News articles with SMAUG (Alves et al., 2022) and using humans to validate perturbation on the test set. The original data for this task is composed of the News articles from OPUS News-Commentary (Tiedemann, 2012) for the language pairs English-German and Portuguese-English.

For the English-German language pairs, there are no Deviation in Meaning errors, as the perturbation is only available for into English language pairs. The new dataset is purposefully unbalanced, as these phenomena are rare, containing approximately 5% of translations with critical errors. Table 1 presents the number of records for each language pair.

Since the dataset for this task is artificially generated, the participants were encouraged to submit systems that did not rely on the provided training data. As such, submissions were split into two groups: *unconstrained* and *constrained*. In the first group, the participants have access to the training data. In the second, the systems should only be trained on quality scores such as DA, HTER and MQM annotations. With this setting, we aim to evaluate whether systems can identify critical errors while maintaining correlations with human judgements.

In the evaluation of this task, the participants were not required to submit any classification threshold for their systems. For the *unconstrained* setting, the systems are specifically trained to detect errors and should output high scores for translations containing these errors. As such, for each language-pair, we considered as positive predictions the K records with highest scores, where K is the number of positive records for that language-pair in the test set. Regarding the *constrained* setting, these systems are only trained on quality scores and are expected to assign lower scores to translations with critical errors. Therefore, we considered the K records with lowest scores as positive predictions. From here, we measured the MCC, *Recall* and *Precision* for each submission.

3 Baseline systems

Task 1: Quality Estimation baseline systems: For Task 1, both for word and sentence-level, we used a multilingual transformer-based Predictor-Estimator approach (Kim et al., 2017), which is described in detail in Fomicheva et al. (2022). For the implementation and training we use the OpenKiwi (Kepler et al., 2019) framework. We trained the baseline model using a multilingual and multitask setting and training jointly on the sentence-level scores and word-level tags. For the word-level loss, $\mathcal{L}_{\text{word}}$, the weight of BAD tags is multiplied by a factor of $\lambda_{BAD} = 3.0$, but the sentence- and word-level loss have equal weight in the overall joint loss estimation: $\mathcal{L} = \mathcal{L}_{\text{word}} + \mathcal{L}_{\text{sent}}$. We trained different baselines for the DA/post-edit originated language pairs and the MQM originated language-pairs.

For the DA/post-edit baseline, the model was trained using the DA scores as sentence targets and the OK/BAD tags as word targets. For training we used the concatenated data for all language pairs

available under training data and used the concatenation of the additional language pairs that were made available in the development set as validation. We trained two baselines with this setup, using different encoders for the encoding (predictor) part of the architecture: (a) XLM-R transformer with the `xlm-roberta-large` model and (b) RemBERT model which has been pre-trained on additional languages that include Yoruba and can hence account for the zero-shot language.

For the MQM baseline, the model was trained using the normalised and inverted MQM scores as sentence targets and the OK/BAD tags as word targets. The baseline model was trained using the concatenated training data for all three language pairs and used the concatenated development data for the same pairs as the validation set. The XLM-R transformer with the `xlm-roberta-large` model was used as an encoder.

Task 2: Explainability baseline systems: We provide two baseline systems for Task 2. One is a random baseline where we sampled scores uniformly at random from a continuous [0..1) range for each target token and for a sentence-level score. The other one is a combination of a supervised quality annotation model, OpenKiwi (Kepler et al., 2019) and LIME (Ribeiro et al., 2016) where OpenKiwi is used to predict sentence-level quality scores while LIME is used to compute, for every token in the target sentence, its importance for the sentence-level quality score returned by OpenKiwi. For the OpenKiwi implementation we used a similar setup described for the baselines of Task 1, but we trained the OpenKiwi model using only sentence-level supervision, to align with the task requirements. We trained two multilingual instances, one on DA- and one on MQM-derived data, using XLM-R large encoder in both cases.

LIME is a model-agnostic post-hoc explanation method which trains a linear model to estimate the behavior of a target model (i.e., OpenKiwi in our case) around an input example to be explained so the weights of the linear model correspond to the importance of individual input tokens. Because higher sentence-level scores in our gold standard mean better translation quality, we invert token-level scores generated by LIME so that higher values correspond to errors as required by the task description.

Task 3: Critical Error Detection baseline systems: For task 3, we consider a baseline system for each setting.

In the *constrained* setting, we considered COMET-QE (Rei et al., 2021)[6], which was a top-performing QE-as-a-Metric system in last years Metrics shared task (Freitag et al., 2021b).

Regarding the *unconstrained* setting, we fine-tune an `xlm-roberta-large` model using the COMET framework (Rei et al., 2020). Both the source and translation are jointly encoded into a vector representation which is the input of a final estimator that predicts the probability of the translation containing a critical error. Here, the estimator weights are randomly initialised. We fine-tune the model on the provided training data for a maximum of 5 epochs. At the end of each epoch, we perform a validation step by measuring the MCC on the validation set considering a classification threshold of 0.5. We select the model with the highest MCC on the validation data.

4 Participants

Alibaba-Translate (T1-DA): For the DA subtask, the team participated in all language pairs except the zero-shot LP. The implemented system (Wang et al., 2021), uses glass-box QE features to estimate the uncertainty of machine translation segments and incorporates the features into the transfer learning from the large-scale pre-trained model, XLM-R. The participants used exclusively the DA data provided for this edition of the QE shared task. Of the provided data, the 7 language pairs except for English-Marathi, were combined to train a multilingual model. For English-Marathi, a separate bilingual model was trained. For the final submission the participants ensembled multiple checkpoints.

(T1-MQM): The submission for sentence-level MQM task is based on a multilingual unified framework for translation evaluation. The applied framework UniTE (Wan et al., 2022) considers three input formats – source-only (QE or reference-free metric), reference-only and source-reference-combined. The participants used synthetic datasets with pseudo labels during continuous pre-training phase, and fine-tuned with DA and MQM training

[6]More precisely we used the `wmt21-comet-qe-mqm` model

datasets from the year 2017 to 2021. To obtain the final model predictions they use the source-only evaluation. For multilingual phase, they ensembled predictions using two different backbones – one using XLM-R encoder and the other using InfoXLM. For the ensembling, they picked the best 2 checkpoints on the development dataset.

BJTU-Toshiba (T1-MQM): BJTU-Toshiba participation focused on ensembling different models and using external data. They ensemble multiple pre-trained models, both monolingual and bilingual. The monolingual models are trained only on the text of the target language. Specifically, they use monolingual BERT, Roberta, and Electra-discriminator as the monolingual extractor, and XLM-R as the bilingual extractor. They also use in-domain parallel data to fine-tune and adapt the pre-trained models to the target language and domain. The in-domain data is selected by a BERT-classifier from the parallel data provided by the news translation task, and for each direction, they end up using roughly 1 million sentence pairs for fine-tuning. They explore two styles of fine-tuning, namely Translation Language Model and Replaced Token Detection. For Replaced Token Detection, they use the first 1/3 layers of the model as generator, and after the training they drop the generator and only use the discriminator as the feature extractor.

HW-TSC (T1): HW-TSC's submission follows Predictor-Estimator framework with a pre-trained XLM-R Predictor, a feed-forward Estimator for sentence-level QE subtask and a binary classifier Estimator for word-level QE subtask. Specially, the Predictor is a cross-lingual language model that receives source and target tokens concatenated and returns representations that attend to both languages. WMT 2022's news translation task training data is been used to train the Predictor using a cross-lingual masked language model objective. All of the WMT QE 2022 DA and MQM training data are used to train two different multilingual QE models, one for sentence-level and another one for word-level.

(T2:) The language encoder trained for Task 1 is being used to get source and target token embeddings. After computing cosine similarity between target and source token embeddings, the max cosine similarity of each target token to all the source tokens is selected as quality score. Intuitively, a low score means the target token is more likely to be an error (lack of good alignment), so every target word quality score is multiplied by a negative value.

HyperMT - aiXplain (T1-all): The system is trained with AutoML functionalities in FLAML framework using lightgbm estimator. It utilizes COMET-QE score as feature along-side with many other linguistic features extracted with Stanza from source texts and their translations: the number of tokens, characters, and the average word length of sentences; the frequency of Part-of-Speech and Named Entity Recognition labels, and the frequency of morphological features. The differences in values of linguistic features between source texts and translations are also included as features. This allows the system to work in multilingual settings as well.

IST-Unbabel (T1-all): IST-Unbabel team proposed an extension of COMET, dubbed COMET-Kiwi, which includes a word-level layer and can be trained on both sentence-level scores and word-level labels in a multi-tasking fashion. Their final submission for task 1 is a weighted ensemble between models trained using InfoXLM (Chi et al., 2021) and RemBERT (Chung et al., 2021). All these models are pretrained on the data from the metrics shared tasks and, for word-level, they pretrained on both QT21 and APE-Quest datasets.

(T2) For the second task they use the COMET-Kiwi framework as the backbone of a sentence-level QE model and added layer and headwise parameters to the QE model: for each layer and for each head, they train individual parameters to construct a sparse distribution over the layers/heads to better leverage these representations. They leveraged different encoders – InfoXLM and RemBERT – and used them individually as the backbone of our QE sentence-level models. The models used to extract explanations were multilingual ones trained for DA and MQM separately. The explainability weights were obtained from the at-

tention weights scaled by the norm of the gradient of the value vectors (Chrysostomou and Aletras, 2022). No word supervision was used and all explanations were extracted relying solely on models that produced the sentence-level scores. The final submissions are ensembles of explanations from different attention layers/heads according to the validation data. For the zero-shot language pair (En-Yo), they created an ensemble with the attention layers/heads that were among the top-performing ensembles for other language pairs.

(T3) For task 3 a single model from task 1 using InfoXLM encoder and trained on DA annotations was submitted.

KU X Upstage (T3): KU X Upstage employs an XLM-R large model without leveraging any additional parallel corpus. Instead, they attempt to maximise its capability by adopting prompt-based fine-tuning, which reformulates the Critical Error Detection task as a masked language modelling objective (a pre-training strategy of this model) before training. They generate hard prompts suitable for QE task through prompt engineering, and templates consist largely of three types according to the information utilised: naive template, template with a contrastive demo, and template with Google Translate. The final score is obtained by extracting the probability of a word mapped to BAD among verbalizers. They gain an additional performance boost from the template ensemble by adding the values from multiple templates.

NJUNLP (T1-all): NJUNLP submission makes use of pseudo data and multi-task learning. Inspired by DirectQE (Cui et al., 2021), they experiment with several novel methods to generate pseudo data for all three subtasks (MQM, DA, and PE) using the conditional masked language model and the NMT model to generate high quality synthetic data and pseudo labels. The proposed methods control the decoding process to generate more fluent pseudo translations close to the actual distribution of the gold data. They pre-train the XLM-R large model with the generated pseudo data and then fine-tune this model with the real QE task data, using multi-task learning in both stages. They jointly learn sentence-level scores (with regression and rank tasks) and word-level tags (with a sequence tagging task). For the final submissions they ensemble sentence-level results by averaging all valid output scores and ensemble word-level results using a voting mechanism. For the pseudo label generation they use publicly available parallel data, specifically: the data provided by the WMT translation task for En-De (9M), En-Ru (3M), and Zh-En (3M) language pairs. The 660K parallel sentences from OPUS[7] for the Km-En language pair. They also use 3.6M parallel data from the target translation model[8] for the En-Mr language pair, as well as WMT2017, WMT2019, and WMT2020 En-De PE data for the En-De language pair.

Papago (T1-full): Papago submitted a multilingual and multi-task model, trained to predict jointly both sentence and word level. The system's architecture consists of Pretrained Language Model with task independent layers optimized for both sentence and word level quality prediction. They propose an auxiliary loss function to the final objective function to further improve performance. They also augment training data by either generating (i.e. pseudo data) or collecting open source data that is deemed to be relevant to QE task. Finally, they train and select the checkpoints for the final submission with cross-validation for better generalization and ensemble multiple models for their final submission.

UCBerkeley-UMD (T1:DA): UCBerkeley-UMD used a large-scale multilingual model to back translate from Czech to English. They compared the quality of the Czech translation by examining the translation from Czech back to English with the original source text in English. This is motivated by literature that humans tend to perform quality checks on translations when they do not understand the target language.

UT-QE (T2): The UT-QE team used XLMR-Score (Azadi et al., 2022) as an unsupervised sentence-level metric, which is computed as BERTScore but in a cross-lingual manner while using the XLM-R model. The matched

[7] https://opus.nlpl.eu/

[8] https://indicnlp.ai4bharat.org/indic-trans/

ID	Affiliations	
Alibaba Translate	DAMO Academy, Alibaba Group & University of Science and Technology of China & CT Lab, University of Macau, China & National University of Singapore, Republic of Singapore	(Bao et al., 2022)
BJTU-Toshiba	Beijing Jiaotong University, China & Toshiba Co., Ltd.	(Huang et al., 2022)
HW-TSC	Huawei Translation Services Center & Nanjing University, China	(Su et al., 2022)
HyperMT - aiXplain	aiXplain	–
IST-Unbabel	INESC-ID & Instituto de Telecomunicações & Instituto Superior Técnico & Unbabel, Portugal	(Rei et al., 2022)
KU X Upstage	Korea University, Korea & Upstage	(Eo et al., 2022)
NJUNLP	Huawei Translation Services Center, China	(Geng et al., 2022)
Papago	Papago, Naver Corp	(Lim and Park, 2022)
UCBerkeley-UMD	University of California, Berkeley & University of Maryland	(Mehandru et al., 2022)
UT-QE	University of Tehran, Iran	(Azadi et al., 2022)
Welocalize-ARC/NKUA	Welocalize Inc, USA & National Kapodistrian University & Athena RC, Greece	(Zafeiridou and Sofianopoulos, 2022)

Table 2: Participants to the WMT22 Quality Estimation shared task.

tokens distances in this metric were used as token-level scores. In order to alleviate the mismatching issues, they also try to fine-tune the XLM-R model on word alignments from parallel corpora to make it represent the aligned words in different languages closer to each other, and use the fine-tuned model instead of XLM-R for scoring sentences and tokens.

Welocalize-ARC/NKUA (T1-DA): Welocalize-ARC/NKUA's submission for the Task 1 follows the Predictor-Estimator framework (Kim et al., 2017) with a regression head on top to estimate the z-standardised DA. More specifically, they use a pre-trained Transformer for feature extraction and then concatenate the extracted features with additional glass-box features. The glass-box features are also produced using pre-trained models and by applying multiple techniques to estimate different types of uncertainty for each translated sentence. The final features are then used as input for the QE regression model, which is a simple sequential Neural Network with a linear output layer. Finally, the performance of the model is optimised by employing Monte Carlo Dropout during both training and inference. Regarding the data, they use only the provided datasets (the MLQE-PE train/dev sets along with the additional dataset for Marathi language) as well as some of the provided additional training resources of the Metrics shared task.

Table 2 lists all participating teams submitting systems to any of the tasks, and Table 3 report the number of successful submissions to each of the sub-tasks and language pairs. Each team was allowed up to ten submissions for each task variant and language pair (with a limit of two submissions per day). In the descriptions below, participation in specific tasks is denoted by a task identifier (T1 = Task 1, T2 = Task 2, T3 = Task 3).

5 Results

In this section, we present and discuss the results of our shared task. Please note that for all the three subtasks we used statistical significance testing with $p = 0.05$.

5.1 Task 1

As we have seen in Task 1 description (§2.1.1), submissions are evaluated against the true z-normalised sentence scores using Spearman's rank correlation coefficient ρ along with the following secondary metrics: Pearson's correlation coefficient, r, Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE). Nonetheless, **the final ranking between systems is calculated using the primary metric only (Spearman's ρ)**. Also, statistical significance was computed using William's test.[9]

For the Task 1 word-level task, the submissions are ranked using the Matthews correlation coeffi-

[9] https://github.com/ygraham/mt-qe-eval

Task/LP	# submission
Task 1 – Sent-level Direct Assessment	**161**
Multilingual w/o En-Yo	21
Multilingual w En-Yo	23
English-Marathi	24
English-Czech	33
English-Japanese	22
Pashto-English	16
Khmer-English	22
Task 1 – Sent-level MQM	**402**
Multilingual	38
English-German	65
English-Russian	62
Chinese-English	76
Task 1 – Word-level	**247**
Multilingual w/o En-Yo	18
Multilingual w En-Yo	17
English-Czech	32
English-Japanese	27
English-Marathi	24
Pashto-English	13
Khmer-English	28
English-German	28
English-Russian	18
Chinese-English	27
Task 2 – Explainable QE	**161**
English-Czech	14
English-Japanese	14
English-Marathi	13
Pashto-English	30
Khmer-English	25
English-German	17
English-Russian	12
Chinese-English	12
English-Yoruba	12
Task 3 – Sent-Level Critical Error Det.	**20**
Constrained	
English-German	2
Portuguese-English	2
Unconstrained	
English-German	10
Portuguese-English	6
Total	**991**

Table 3: Number of submissions to each sub-task and language-pair at the WMT22 Quality Estimation shared task.

cient (MCC). F1-scores are provided as complementary information only and statistical significance was computed using randomisation tests (Yeh, 2000) with Bonferroni correction (Abdi, 2007) for each language pair.

The majority of participants implemented multilingual models and the top performing systems adopted a multi-tasking approach, learning the sentence- and word-level targets jointly (IST-Unbabel, Papago, NJUNLP). It is important to note that all participants relied on large pre-trained encoders (XLM-R, RemBERT, BERT, ELECTRA), which seems to be the norm for high-performance in quality estimation, but can constitute a limitation for performance in truly multi-lingual scenarios where the target languages are not seen during pre-training. Additionally, many final submissions consisted of ensembles combining different large pretrained models increasing even further the total number of model parameters.

Another trend that seems to carry on from previous editions of the task is the incorporation of additional features in QE models (glass-box features were incorporated in Alibaba's DA systems while linguistic features were incorporated in aiXplain QE system), however in this edition such approaches were outperformed by models that put more emphasis on pre-training, using auxiliary tasks and external data.

For the sentence-level sub-tasks, participants managed to achieve high correlations for the majority of language pairs, especially for the DA originated data, with the exception of En-Ja. The results show an improvement compared to the last edition, although it is hard to draw a direct comparison due to changes in the available train/development data. However, it is interesting to note that performance for En-Mr, for which we provided considerable more data than for the other language pairs is still in the same range as results for the other language pairs. It would thus be interesting to investigate further which properties render a language pair harder to evaluate.

For the MQM data the overall correlations achieved were lower in comparison to the DA ones although still meaningful. Note that compared to the DA data, the MQM language pairs were high-resource ones, which could also influence performance. Additionally, small discrepancies between the annotation guidelines in the train set and the dev/test sets could have further complicated the task. We intend to further investigate the MQM potential in future editions, with the addition of new language pairs and more annotated data.

For the word-level subtask, IST-Unbabel, NJUNLP and Papago tied at the top for most language pairs, and we can observe that correlations are moderate across language pairs (both DA and MQM originated ones). It is important to note that no team seems to have submitted predictions using a word-level only supervision; instead all the participants of this task used a multi-task approach, learning jointly word and sentence level scores.

Model	Multi	Multi (w/o En-Yo)	En-Cs	En-Ja	En-Mr	Km-En	Ps-En
IST-Unbabel	**0.572**	**0.605**	**0.655**	**0.385**	**0.592**	**0.669**	**0.722**
Papago	0.502	0.571	**0.636**	0.327	**0.604**	0.653	0.671
Alibaba Translate	–	0.585	0.635	0.348	**0.597**	0.657	0.697
Welocalize-ARC/NKUA	0.448	0.506	0.563	0.276	0.444	0.623	–
BASELINE	0.415	0.497	0.560	0.272	0.436	0.579	0.641
lp_sunny‡	0.414	0.485	0.511	0.290	0.395	0.611	0.637
HW-TSC	–	–	0.626	0.341	0.567	0.509	0.661
aiXplain	–	–	0.477	0.274	0.493	–	–
NJUNLP	–	–	–	–	**0.585**	–	–
UCBerkeley-UMD*	–	–	0.285	–	–	–	–

Table 4: Spearman correlation with **Direct Assessments** for the submissions to WMT22 Quality Estimation **Task 1**. For each language pair, results marked in bold correspond to the winning submissions, as they are not significantly outperformed by any other system according to the Williams Significance Test (Williams, 1959). Baseline systems are highlighted in grey; ‡ indicates Codalab username of participants from whom we have not received further information and * indicates late submissions that were not considered for the official ranking of participating systems

Model	Multi	En-De	En-Ru	Zh-En
IST-Unbabel	**0.474**	0.561	**0.519**	**0.348**
NJUNLP	0.468	**0.635**	**0.474**	**0.296**
Alibaba-Translate	0.456	0.550	**0.505**	**0.347**
Papago	0.449	0.582	**0.496**	**0.325**
lp_sunny ‡	0.415	0.495	0.453	0.298
BASELINE	0.317	0.455	0.333	0.164
BJTU-Toshiba	–	**0.621**	0.434	**0.299**
HW-TSC	–	0.494	0.433	0.369
aiXplain	–	0.376	0.338	0.194
pu_nlp ‡	–	0.611	–	–

Table 5: Spearman correlation with **MQM** for the submissions to WMT22 Quality Estimation **Task 1**. For each language pair, results marked in bold correspond to the winning submissions, as they are not significantly outperformed by any other system according to the Williams Significance Test (Williams, 1959). Baseline systems are highlighted in grey; ‡ indicates Codalab username of participants from whom we have not received further information.

Best performers The scores in Tables 4 - 6 show the participant scores for the main metric, ordered by the best performance in the multilingual subtasks. IST-Unbabel is the clear winner for the multilingual subtasks, but for the individual language pairs results vary and multiple participants are tied at the top. All top-performing approaches (IST-Unbabel, Papago, NJUNLP and Alibaba) share some common characteristics: (1) they constitute multilingual and multi-task approaches; (2) they use external data during pre-training, either adapted from other tasks (such as the Metrics task (Freitag et al., 2022)) or generated artificially (pseudo data); and (3) they use ensembling for the final submission.

5.2 Task 2

Three teams participated in Task 2, IST-Unbabel, HW-TSC and UT-QE. IST-Unbabel participated in all 9 language pairs, HW-TSC in all languages pairs except English-Yoruba, and UT-QE only in Khmer-English and Pashto-English. As shown in Table 7, IST-Unbabel wins 7 of 9 LPs according to the metric Recall at Top-K, HW-TSC the remaining 2. With Bonferroni correction, IST-Unbabel wins 4 LPs, HW-TSC wins 2, and both are indistinguishable on the remaining 3 LPs. Average precision (AP) yields identical results as Recall at Top-K in terms of ranking of the teams. There is one difference according to the metric AUC in terms of winners: HW-TSC wins English-Japanese. Finally, all participating teams beat both baselines in all cases.

For sentence-level performance (see Appendix D), IST-Unbabel wins all LPs according to Pearson's correlation and all LPs according to Spearman's correlation except for Khmer-English, which HW-TSC wins. Not all teams beat all baselines in terms of sentence-level performance.

The winning teams obtain the lowest sentence-level correlations for English-Chinese, English-Japanese and English-Yoruba and the highest correlations for Khmer-English and English-German. This may be related to the quality of annotations and the quality of MT systems involved. For word-level explainability scores, the lowest Recall at Top-K scores are obtained for English-Yoruba and English-Marathi, whereas the highest scores are obtained for Pashto-English and Khmer-English. The fact that the winning systems obtain low sentence and word-level scores for English-Yoruba and high scores for Khmer-English may indicate that the

Model	Multi	Multi (w/o En-Yo)	En-Cs	En-Ja	En-Mr	Kh-En	Ps-En	En-De	En-Ru	Zh-En
IST-Unbabel	**0.341**	**0.361**	**0.436**	**0.238**	**0.392**	**0.425**	**0.424**	**0.303**	**0.427**	**0.360**
Papago	0.317	**0.343**	**0.396**	**0.257**	**0.418**	**0.429**	0.374	**0.319**	**0.421**	**0.351**
BASELINE	0.235	0.257	0.325	0.175	0.306	0.402	0.359	0.182	0.203	0.104
HW-TSC	–	0.218	**0.424**	**0.258**	0.351	0.353	0.358	0.274	0.343	0.246
NJUNLP	–	–	–	–	**0.412**	**0.421**	–	**0.352**	**0.390**	**0.308**

Table 6: **Matthew Correlation Coefficient** (MCC) for the submissions to WMT22 Quality Estimation **Task 1 (word-level)**. For each language pair, results marked in bold correspond to the winning submissions, as they are not significantly outperformed by any other system based on randomisation tests with Bonferroni correction (Yeh, 2000). Baseline systems are highlighted in grey.

Model	En-Cs	En-Ja	En-Mr	En-Ru	En-De	En-Yo	Km-En	Ps-En	Zh-En
IST-Unbabel	**0.561**	**0.466**	**0.317**	**0.390**	**0.365**	**0.234**	0.665	0.672	**0.379**
HW-TSC	**0.536**	**0.462**	**0.280**	0.313	0.252	–	**0.686**	**0.715**	0.220
BASELINE (OpenKiwi+LIME)	0.417	0.367	0.194	0.135	0.074	0.111	0.580	0.615	0.048
BASELINE (Random)	0.363	0.336	0.167	0.148	0.124	0.144	0.565	0.614	0.093
UT-QE	–	–	–	–	–	–	0.622	0.668	–

Table 7: **Recall at Top-K** for the submissions to the WMT22 Quality Estimation **Task 2 (Explainable QE)**. For each language pair, results marked in bold correspond to the winning submissions, as they are not significantly outperformed by any other system based on randomisation tests with Bonferroni correction (Yeh, 2000). Baseline systems are highlighted in grey.

Model	En-De (Cons)	En-De (UN-cons)	Pt-En (Cons)	Pt-En (UN-cons)
KU X Upstage	–	0.964	–	0.984
IST-Unbabel	0.564	–	0.721	–
BASELINE	0.074	0.855	-0.001	0.934
aiXplain	–	0.219	–	0.179

Table 8: **Matthews Correlation Coefficient** (MCC) for the submissions to WMT21 Quality Estimation **Task 3 (Critical Error Detection)**. For each language pair, results marked in bold correspond to the winning submissions, as they are not significantly outperformed by any other system based on randomisation tests with Bonferroni correction (Yeh, 2000). Baseline systems are highlighted in grey.

tasks are correlated (as one may intuitively expect): a QE system that yields better sentence-level scores also highlights word-level errors more correctly.

5.3 Task 3

In this task, we divide participants into *unconstrained* and *constrained* settings, and address each group in separate. As in the last year, this task attracted few participants, which we attribute to the recentness of the task.

In the *unconstrained* setting, there are two participants: KU X Upstage and HyperMT - aiXplain. The first achieved very high values for the measured metrics, and is the best performer for this setting for both language pairs. The second obtained lower values, falling below the baseline on both language pairs.

In the *constrained* setting, a single submission was received: IST-Unbabel. Their system outperformed the baseline on both language pairs.

6 Discussion

In what follows, we discuss the main findings of this year's shared task based on the goals we had previously identified for it.

General progress Participating systems achieved very promising results for most languages, including the newly introduced language-pairs as well as the new annotation style (MQM). The **best performing submissions showed moderate to strong correlation for sentence-level DA and MQM prediction tasks.** While it is hard to draw direct comparisons with the previous editions, the overall correlation scores obtained are similar or improved for the common language-pairs. In combination with the outcomes of previous editions, it seems that multi-lingual and multi-task systems that are able to take advantage of multiple resources, are showing better and more robust results. However, the **word-level quality prediction is still a challenging task and there is ample room for improvement**. Along the same lines, further **exploring explainability tasks, that support the sentence level predictions with word level scores seems a promising path to motivate finer-grained approaches to word-level** quality annotations.

DA vs MQM annotations To further understand the observed discrepancies between top performances in the DA and MQM sub-tasks for sentence-level quality estimation, we analyse the distributions of predicted scores vs gold scores for each language pair, as presented in Figure 2.

We can see in the scatter plots that there are multiple test-segments which are annotated as perfect translations (maximum possible normalised MQM score), which fail to be classified accordingly as indicated by the top parts of the MQM scatter plots in Figure 2. Overall, even with DA annotations we can see that **language pairs with more balanced distribution between high and low quality segments (Km-En, Ps-En) are those for which QE systems obtain better correlations**, compare to more skewed language pairs (En-Mr, En-Ja).

Additionally, we can see that the **MQM scores are significantly skewed towards higher scores**, with long-tails of few very low quality instances. This provides motivation to revisit the quantification of MQM annotations to generate sentence level scores and further experiments into consolidating MQM annotations from different annotators. Furthermore, perhaps providing access to finer-grained MQM annotations (using the category or severity labels as targets) could aid in obtaining more meaningful outcomes. In future editions we intend to further expand the coverage of languages for MQM annotations that will allow us to draw further conclusions and push the state-of-the-art further in this track.

Zero shot predictions We found that **even without development data or prior knowledge about the language pair, the systems that submitted predictions for En-Yo still achieved meaningful correlations**. For the quality assessment and explainability tasks, the achieved correlations are lower compared to the "seen" language pairs, but still comparable. We can also observe the scatter plot distributions that show the correlation obtained by the top performing system that is comparable with the other DA distributions.

However, we noticed that the **availability of the zero-shot languages in the frequently used pretrained encoders posed an additional challenge** for the participants as the performance on En-Yo seemed dependent on whether the pretrained language model had seen Yoruba text during pretraining. In future editions, we hope that mixing different zero-shot languages will further motivate unsupervised approaches.

Explainable quality estimation The performance of the baselines in Task 2 suggests that applying a model-agnostic explanation method (i.e., LIME) to a relatively good sentence-level QE system (i.e., OpenKiwi) straightforwardly may not result in plausible explanations. In particular, the OpenKiwi+LIME baseline got higher Recall at Top-K than the random baseline for only 5 LPs. Using randomisation tests with Bonferroni correction, we found that the OpenKiwi+LIME baseline can significantly outperform the random baseline for only 2 LPs (En-Cs and En-Ja). Despite its higher Pearson's correlation at the sentence level, OpenKiwi+LIME yielded random-like (or even worse) explanations for MQM language pairs. This also calls for a stronger baseline for the future edition of the QE shared task. Additional signals/heuristics might be added to the future shared task's baselines such as sparsity of the rationales (as used by IST-Unbabel) and alignments between source and target sentences (as used by HW-TSC and UT-QE).

Critical error detection. By comparing the performance of the submitted systems, in particular the baselines, we see that the difficulty of the *constrained* setting is much higher. We attribute this discrepancy to the fact that the artificially generated data follows a specific set of patterns, which can be captured by current methods when given enough examples. The HyperMT - aiXplain submission seems to be an exception. However, although this system is *unconstrained*, it is composed of fine-tuned decision trees where the base features are *constrained*. We consider that these features are unable to provide sufficient information for the decision trees to be able to identify critical errors, even when fine-tuned on the provided training data.

Due to the scarcity of annotated data containing critical errors, we argue that the *constrained* setting presents a much more realist challenge, where systems are trained for correlating with human judgements but are tested for robustness to critical errors.

For a future edition of this task, we envision a design that simultaneously considers both correlations with human judgements and robustness to critical errors when evaluating a QE system. This can be combined with Task 1, where besides the current evaluation method, participants would also receive a robustness score for their systems, mea-

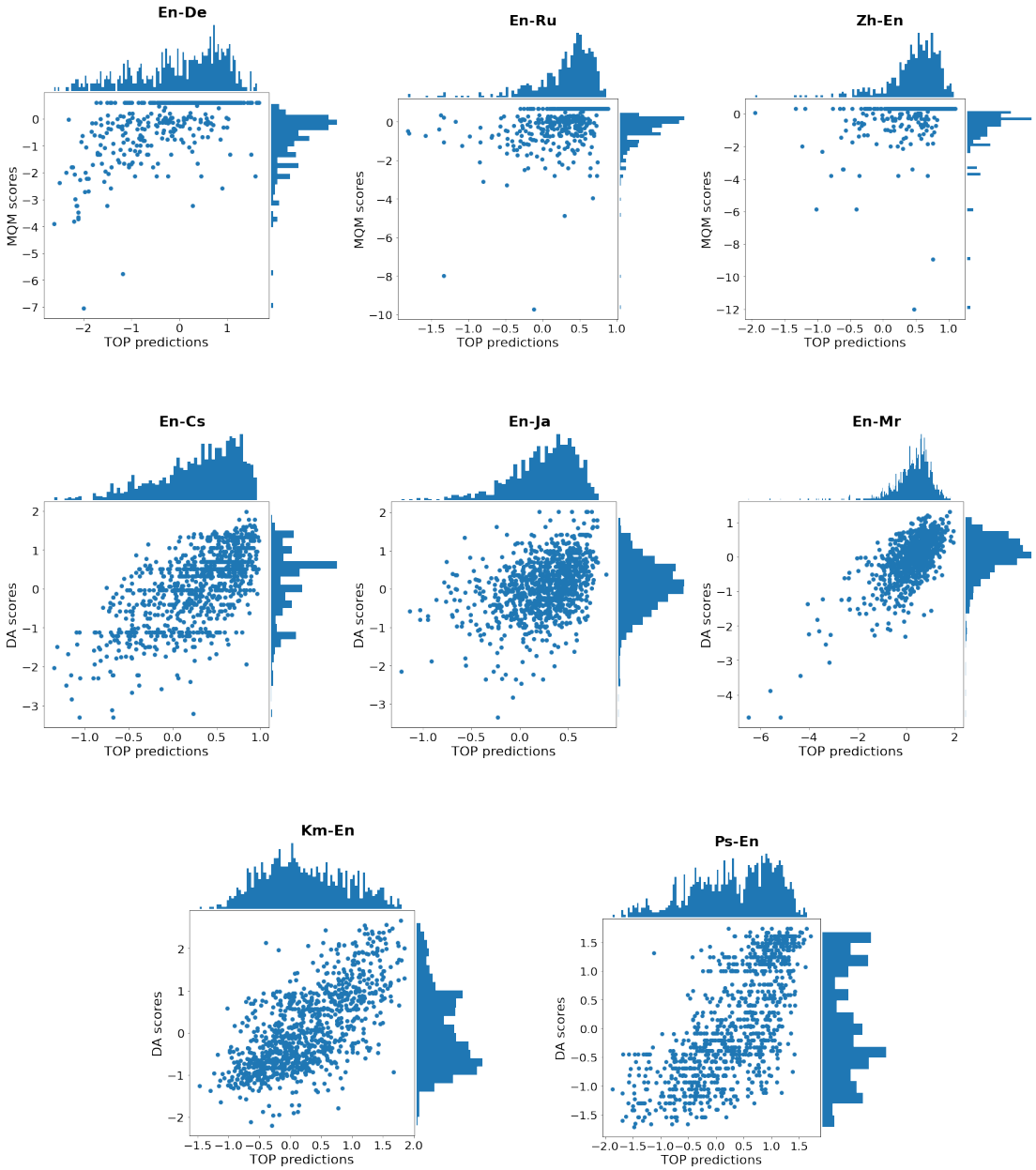

Figure 2: Scatter plots for the predictions against true DA/MQM scores for the top-performing system for each language pair. The histograms show the corresponding marginal distributions of predicted and true scores.

sured on a test set with critical errors. We hope that this configuration would both attract more participants to this task (as it would not required training a specific system for critical error detection) and further motivate the treatment of critical errors in the development of QE systems.

7 Conclusions

This year's edition of the QE Shared Task introduced a number of new elements: new low-resource language pairs (Marathi and Yoruba), new annotation conventions for sentence and word level quality (MQM), new test sets, and new versions of explainability and critical error detection subtasks. The tasks attracted a steady number of participating teams and we believe the overall results are a great reflection of the state-of-the-art in QE.

We have made the gold labels and all submissions to all tasks available for those interested in further analysing the results, while newly interested participants can still access the competition instances on codalab and directly compare their performance to other models. We aspire for the future editions to continue the efforts set in this and previous years and expand the resources and coverage of QE, while further exploring recent and more challenging subtasks such as fine-grained QE, explainable QE and critical error detection.

Acknowledgments

Ricardo Rei and José G. C. de Souza are supported by the P2020 program (MAIA: contract 045909) and by European Union's Horizon Europe Research and Innovation Actions (UTTER: contract 101070631)

André Martins and Chrysoula Zerva are supported by the P2020 program (MAIA: contract 045909), by the European Research Council (ERC StG DeepSPIN 758969), and by the Fundação para a Ciência e Tecnologia through contract UIDB/50008/2020.

Marina Fomicheva and Lucia Specia were supported by funding from the Bergamot project (EU H2020 Grant No. 825303).

References

Hervé Abdi. 2007. The bonferroni and šidák corrections for multiple comparisons. *Encyclopedia of measurement and statistics*, 3:103–107.

Farhad Akhbardeh, Arkady Arkhangorodsky, Magdalena Biesialska, Ondřej Bojar, Rajen Chatterjee, Vishrav Chaudhary, Marta R. Costa-jussa, Cristina España-Bonet, Angela Fan, Christian Federmann, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Leonie Harter, Kenneth Heafield, Christopher Homan, Matthias Huck, Kwabena Amponsah-Kaakyire, Jungo Kasai, Daniel Khashabi, Kevin Knight, Tom Kocmi, Philipp Koehn, Nicholas Lourie, Christof Monz, Makoto Morishita, Masaaki Nagata, Ajay Nagesh, Toshiaki Nakazawa, Matteo Negri, Santanu Pal, Allahsera Auguste Tapo, Marco Turchi, Valentin Vydrin, and Marcos Zampieri. 2021. Findings of the 2021 conference on machine translation (WMT21). In *Proceedings of the Sixth Conference on Machine Translation*, pages 1–88, Online. Association for Computational Linguistics.

Duarte M. Alves, Ricardo Rei, Ana C. Farinha, José G. C. de Souza, and André F. T. Martins. 2022. Robust MT evaluation with Sentence-level Multilingual data Augmentation. In *Proceedings of the Seventh Conference on Machine Translation*, Abu Dhabi. Association for Computational Linguistics.

Chantal Amrhein and Rico Sennrich. 2022. Identifying Weaknesses in Machine Translation Metrics Through Minimum Bayes Risk Decoding: A Case Study for COMET. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing*, Online. Association for Computational Linguistics.

Fatemeh Azadi, Heshaam Faili, and Mohammad Javad Dousti. 2022. Mismatching-Aware Unsupervised Translation Quality Estimation for Low-Resource Languages. *arXiv preprint arXiv:2208.00463*.

Keqin Bao, Yu Wan, Dayiheng Liu, Baosong Yang, Wenqiang Lei, Xiangnan He, Derek F. Wong, and Jun Xie. 2022. Alibaba-translate china's submission for wmt 2022 quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation*, Abu Dhabi. Association for Computational Linguistics.

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

Valerio Basile, Federico Cabitza, Andrea Campagner, and Michael Fell. 2021. Toward a perspectivist turn in ground truthing for predictive computing. *arXiv preprint arXiv:2109.04270*.

Zewen Chi, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-Ling Mao, Heyan Huang, and Ming Zhou. 2021. InfoXLM: An information-theoretic framework for cross-lingual language model pre-training. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3576–3588, Online. Association for Computational Linguistics.

George Chrysostomou and Nikolaos Aletras. 2022. An empirical study on explanations in out-of-domain settings. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6920–6938, Dublin, Ireland. Association for Computational Linguistics.

Hyung Won Chung, Thibault Fevry, Henry Tsai, Melvin Johnson, and Sebastian Ruder. 2021. Rethinking Embedding Coupling in Pre-trained Language Models. In *International Conference on Learning Representations*.

Qu Cui, Shujian Huang, Jiahuan Li, Xiang Geng, Zaixiang Zheng, Guoping Huang, and Jiajun Chen. 2021. Directqe: Direct pretraining for machine translation quality estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 12719–12727.

David M Eberhard, Gary F Simons, and Charles D Fennig. 2020. Ethnologue: Languages of the world (2020). *URL: https://www. ethnologue. com/(visited on Apr. 11, 2020)(cit. on p. 14)*.

Sugyeong Eo, Chanjun Park, Hyeonseok Moon, Jaehyung Seo, and Heuiseok Lim. 2022. KU X Upstage's submission for the WMT22 Quality Estimation: Critical Error Detection Shared Task. In *Proceedings of the Seventh Conference on Machine Translation*, Abu Dhabi. Association for Computational Linguistics.

Marina Fomicheva, Piyawat Lertvittayakumjorn, Wei Zhao, Steffen Eger, and Yang Gao. 2021. The Eval4NLP shared task on explainable quality estimation: Overview and results. In *Proceedings of the 2nd Workshop on Evaluation and Comparison of NLP Systems*, pages 165–178, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Marina Fomicheva, Shuo Sun, Erick Fonseca, Chrysoula Zerva, Frédéric Blain, Vishrav Chaudhary, Francisco Guzmán, Nina Lopatina, Lucia Specia, and André F. T. Martins. 2022. MLQE-PE: A multilingual quality estimation and post-editing dataset. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4963–4974, Marseille, France. European Language Resources Association.

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

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo George Foster, Craig Stewart, Tom Kocmi, Eleftherios Avramidis, Alon Lavie, and André F. T. Martins. 2022. Results of the WMT22 Metrics Shared Task. In *Proceedings of the Seventh Conference on Machine Translation*, Abu Dhabi. 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.

Xiang Geng, Yu Zhang, Shujian Huang, Shimin Tao, Hao Yang, and Jiajun Chen. 2022. NJUNLP's Participation for the WMT2022 Quality Estimation Shared Task. In *Proceedings of the Seventh Conference on Machine Translation*, Abu Dhabi. Association for Computational Linguistics.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2016. Can machine translation systems be evaluated by the crowd alone. *Natural Language Engineering*, FirstView:1–28.

Francisco Guzmán, Peng-Jen Chen, Myle Ott, Juan Pino, Guillaume Lample, Philipp Koehn, Vishrav Chaudhary, and Marc'Aurelio Ranzato. 2019. The FLORES evaluation datasets for low-resource machine translation: Nepali–English and Sinhala–English. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6098–6111, Hong Kong, China. Association for Computational Linguistics.

Hui Huang, Hui Di, Chunyou Li, Hanming Wu, Kazushige Oushi, Yufeng Chen, Jian Liu, and Jin'an Xu. 2022. BJTU-Toshiba's Submission to WMT22 Quality Estimation Shared Task. In *Proceedings of the Seventh Conference on Machine Translation*, Abu Dhabi. Association for Computational Linguistics.

Diptesh Kanojia, Marina Fomicheva, Tharindu Ranasinghe, Frédéric Blain, Constantin Orăsan, and Lucia Specia. 2021. Pushing the right buttons: Adversarial evaluation of quality estimation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 625–638, Online. Association for Computational Linguistics.

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.

Anoop Kunchukuttan, Pratik Mehta, and Pushpak Bhattacharyya. 2018. The IIT Bombay English-Hindi parallel corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Seunghyun S. Lim and Jeonghyeok Park. 2022. Papago's submission to the wmt22 quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation*, Abu Dhabi. Association for Computational Linguistics.

Arle Lommel, Aljoscha Burchardt, Maja Popović, Kim Harris, Eleftherios Avramidis, and Hans Uszkoreit. 2014. Using a new analytic measure for the annotation and analysis of mt errors on real data. In *Proceedings of the 17th Annual conference of the European Association for Machine Translation*, pages 165–172.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press, USA.

Nikita Mehandru, Marine Carpuat, and Niloufar Selehi. 2022. Quality Estimation by Backtranslation at the WMT 2022 Quality Estimation Task. In *Proceedings of the Seventh Conference on Machine Translation*, Abu Dhabi. Association for Computational Linguistics.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A Fast, Extensible Toolkit for Sequence Modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

Gowtham Ramesh, Sumanth Doddapaneni, Aravinth Bheemaraj, Mayank Jobanputra, Raghavan AK, Ajitesh Sharma, Sujit Sahoo, Harshita Diddee, Mahalakshmi J, Divyanshu Kakwani, Navneet Kumar, Aswin Pradeep, Srihari Nagaraj, Kumar Deepak, Vivek Raghavan, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh Shantadevi Khapra. 2022. Samanantar: The Largest Publicly Available Parallel Corpora Collection for 11 Indic Languages. *Transactions of the Association for Computational Linguistics*, 10:145–162.

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

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C. Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. CometKiwi: IST-Unbabel 2022 Submission for the Quality Estimation Shared Task. In *Proceedings of the Seventh Conference on Machine Translation*, Abu Dhabi. Association for Computational Linguistics.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231.

Lucia Specia, Frédéric Blain, Marina Fomicheva, Chrysoula Zerva, Zhenhao Li, Vishrav Chaudhary, and André F. T. Martins. 2021. Findings of the WMT 2021 shared task on quality estimation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 684–725, Online. Association for Computational Linguistics.

Chang Su, Miaomiao Ma, Shimin Tao, Hao Yang, Min Zhang, Xiang Geng, Shujian Huang, Jiaxin Guo, Minghan Wang, and Yinglu Li. 2022. CrossQE: HW-TSC 2022 Submission for the Quality Estimation Shared. In *Proceedings of the Seventh Conference on Machine Translation*, Abu Dhabi. Association for Computational Linguistics.

Y. Tang, C. Tran, Xian Li, P. Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning. *ArXiv*, abs/2008.00401.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

Yu Wan, Dayiheng Liu, Baosong Yang, Haibo Zhang, Boxing Chen, Derek Wong, and Lidia Chao. 2022. UniTE: Unified translation evaluation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8117–8127, Dublin, Ireland. Association for Computational Linguistics.

Jiayi Wang, Ke Wang, Boxing Chen, Yu Zhao, Weihua Luo, and Yuqi Zhang. 2021. QEMind: Alibaba's submission to the WMT21 quality estimation shared task. In *Proceedings of the Sixth Conference on Machine Translation*, pages 948–954, Online. Association for Computational Linguistics.

Evan J. Williams. 1959. *Regression Analysis*, volume 14. Wiley, New York, USA.

Alexander Yeh. 2000. More Accurate Tests for the Statistical Significance of Result Differences. In *Coling-2000: the 18th Conference on Computational Linguistics*, pages 947–953, Saarbrücken, Germany.

Eirini Zafeiridou and Sokratis Sofianopoulos. 2022. Welocalize-ARC/NKUA's Submission to the WMT 2022 Quality Estimation Shared Task. In *Proceedings of the Seventh Conference on Machine Translation*, Abu Dhabi. Association for Computational Linguistics.

A Official Results of the WMT22 Quality Estimation Task 1 (Direct Assessment)

Tables 9, 10, 11, 12, 13, 14 and 15 show the results for all language pairs and the multilingual variants, ranking participating systems best to worst using Spearman correlation as primary key for each of these cases.

Model	**Spearman**	RMSE	MAE	Disk footprint (B)	# Model params
• IST-Unbabel	0.572	0.689	0.539	2,260,735,089	583,891,109
Papago	0.502	2.404	2.077	2,243,044,839	560,713,447
Welocalize-ARC/NKUA	0.448	0.794	0.632	2,307,101,417	576,733,248
BASELINE	0.415	0.979	0.820	2,280,011,066	564,527,011
lp_sunny ‡	0.414	1.054	0.898	2,356,736,392	580,792,183

Table 9: Official results of the WMT22 Quality Estimation Task 1 **Direct Assessment** for the **Multilingual** variant. Teams marked with "•" are the winners, as they are not significantly outperformed by any other system according to the Williams Significance Test (Williams, 1959). Baseline systems are highlighted in grey. ‡ indicates Codalab usernames of participants from whom we have not received further information.

Model	**Spearman**	RMSE	MAE	Disk footprint (B)	# Model params
• IST-Unbabel	0.605	0.671	0.521	2,260,735,089	583,891,109
Alibaba Translate	0.587	0.675	0.533	2,191,440	560,981,507
Papago	0.571	1.793	1.451	2,243,044,839	560,713,447
Welocalize-ARC/NKUA	0.506	0.733	0.571	2,307,068,585	576,725,041
BASELINE	0.497	0.748	0.585	2,280,011,066	564,527,011
lp_sunny ‡	0.485	0.757	0.596	2,356,736,392	580,792,183

Table 10: Official results of the WMT22 Quality Estimation Task 1 **Direct Assessment** for the **Multilingual (w/o English-Yoruba)** variant. Teams marked with "•" are the winners, as they are not significantly outperformed by any other system according to the Williams Significance Test (Williams, 1959). Baseline systems are highlighted in grey. ‡ indicates Codalab usernames of participants from whom we have not received further information.

Model	**Spearman**	RMSE	MAE	Disk footprint (B)	# Model params
• IST-Unbabel	0.655	0.720	0.545	2,260,735,089	583,891,109
• Papago	0.636	1.371	1.081	2,243,044,839	560,713,447
Alibaba Translate	0.635	0.746	0.607	2,191,440	560,981,507
HW-TSC	0.626	0.712	0.545	540,868,112	222,353,517
Welocalize-ARC/NKUA	0.563	0.785	0.610	2,307,068,585	576,725,041
BASELINE	0.560	0.804	0.608	2,280,011,066	564,527,011
lp_sunny ‡	0.511	0.786	0.614	2,356,736,392	580,792,183
aiXplain	0.477	0.825	0.679	745,679,835	12,345
UCBerkeley-UMD*	0.285	1.252	0.961	–	177,853,440

Table 11: Official results of the WMT22 Quality Estimation Task 1 **Direct Assessment** for the **English-Czech** dataset. Teams marked with "•" are the winners, as they are not significantly outperformed by any other system according to the Williams Significance Test (Williams, 1959). Baseline systems are highlighted in grey. ‡ indicates Codalab usernames of participants from whom we have not received further information and * indicates late submissions that were not considered for the official ranking of participating systems

Model	**Spearman**	RMSE	MAE	Disk footprint (B)	# Model params
• IST-Unbabel	0.385	0.689	0.528	2,260,735,089	583,891,109
Alibaba Translate	0.348	0.673	0.522	2,191,440	560,981,507
HW-TSC	0.341	0.726	0.555	540,868,112	222,353,517
Papago	0.327	2.253	1.957	2,243,044,839	560,713,447
lp_sunny ‡	0.290	0.718	0.556	2,356,736,392	580,792,183
Welocalize-ARC/NKUA	0.276	0.755	0.579	2,307,068,585	576,725,041
aiXplain	0.274	0.704	0.547	745,679,835	12,345
BASELINE	0.272	0.747	0.576	2,280,011,066	564,527,011

Table 12: Official results of the WMT22 Quality Estimation Task 1 **Direct Assessment** for the **English-Japanese** dataset. Teams marked with "•" are the winners, as they are not significantly outperformed by any other system according to the Williams Significance Test (Williams, 1959). Baseline systems are highlighted in grey. ‡ indicates Codalab usernames of participants from whom we have not received further information.

Model	**Spearman**	RMSE	MAE	Disk footprint (B)	# Model params
• Papago	0.604	0.658	0.514	2,243,044,839	560,713,447
• Alibaba Translate	0.597	0.456	0.349	2,191,440	560,981,507
• IST-Unbabel	0.592	0.498	0.365	6,932,353,559	583,891,109
• NJUNLP	0.585	0.617	0.414	3,264,730,349	560,145,557
HW-TSC	0.567	0.506	0.372	222,353,517	540,868,112
aiXplain	0.493	0.540	0.396	745,679,835	12,345
Welocalize-ARC/NKUA	0.444	0.534	0.401	2,307,068,585	576,725,041
BASELINE	0.436	0.628	0.461	2,280,011,066	564,527,011
lp_sunny ‡	0.395	0.570	0.443	2,356,736,392	580,792,183

Table 13: Official results of the WMT22 Quality Estimation Task 1 **Direct Assessment** for the **English-Marathi** dataset. Teams marked with "•" are the winners, as they are not significantly outperformed by any other system according to the Williams Significance Test (Williams, 1959). Baseline systems are highlighted in grey. ‡ indicates Codalab usernames of participants from whom we have not received further information.

Model	**Spearman**	RMSE	MAE	Disk footprint (B)	# Model params
• IST-Unbabel	0.669	0.714	0.569	2,260,735,089	583,891,109
Alibaba Translate	0.657	0.778	0.596	2,191,440	560,981,507
Papago	0.653	2.786	2.291	2,243,044,839	560,713,447
Welocalize-ARC/NKUA	0.623	0.794	0.619	2,307,068,585	576,725,041
lp_sunny ‡	0.611	0.784	0.621	2,356,736,392	580,792,183
BASELINE	0.579	0.774	0.616	2,280,011,066	564,527,011
HW-TSC	0.509	1.043	0.804	222,353,517	540,868,112

Table 14: Official results of the WMT22 Quality Estimation Task 1 **Direct Assessment** for the **Khmer-English** dataset. Teams marked with "•" are the winners, as they are not significantly outperformed by any other system according to the Williams Significance Test (Williams, 1959). Baseline systems are highlighted in grey. ‡ indicates Codalab usernames of participants from whom we have not received further information.

Model	**Spearman**	RMSE	MAE	Disk footprint (B)	# Model params
• IST-Unbabel	0.722	0.719	0.575	2,260,735,089	583,891,109
Alibaba Translate	0.697	0.720	0.594	2,191,440	560,981,507
Papago	0.671	0.763	0.646	2,243,044,839	560,713,447
HW-TSC	0.661	0.729	0.592	540,868,112	222,353,517
BASELINE	0.641	0.788	0.663	2,280,011,066	564,527,011
lp_sunny ‡	0.637	0.954	0.775	2,356,736,392	580,792,183

Table 15: Official results of the WMT22 Quality Estimation Task 1 **Direct Assessment** for the **Pashto-English** dataset. Teams marked with "•" are the winners, as they are not significantly outperformed by any other system according to the Williams Significance Test (Williams, 1959). Baseline systems are highlighted in grey. ‡ indicates Codalab usernames of participants from whom we have not received further information.

B Official Results of the WMT22 Quality Estimation Task 1 (MQM)

Tables 16, 17, 18 and 19 show the results for all language pairs and the multilingual variant, ranking participating systems best to worst using Spearman correlation as primary key for each of these cases.

Model	**Spearman**	RMSE	MAE	Disk footprint (B)	# Model params
• IST-Unbabel	0.474	0.973	0.559	2,260,735,089	583,891,109
NJUNLP	0.468	0.945	0.579	3,264,730,349	560,145,557
Alibaba Translate	0.456	0.855	0.493	2,260,733,079	565,137,999
Papago	0.449	1.332	0.990	2,243,044,839	560,713,447
lp_sunny ‡	0.415	0.952	0.536	2,356,736,392	580,792,183
BASELINE	0.317	1.041	0.575	2,280,011,066	564,527,011

Table 16: Official results of the WMT22 Quality Estimation Task 1 **MQM** for the **Multilingual** variant. Baseline systems are highlighted in grey. ‡ indicates Codalab usernames of participants from whom we have not received further information.

Model	**Spearman**	RMSE	MAE	Disk footprint (B)	# Model params
• NJUNLP	0.635	0.838	0.594	3,264,730,349	560,145,557
• BJTU-Toshiba	0.621	0.818	0.545	2,239,711,849	559,893,507
pu_nlp ‡	0.611	0.997	0.716	1,326,455,799	237,846,178
Papago	0.582	0.906	0.556	2,243,044,839	560,713,447
IST-Unbabel	0.561	0.854	0.521	2,260,743,851	565,139,485
Alibaba Translate	0.550	0.769	0.466	2,260,733,079	565,137,999
lp_sunny ‡	0.495	0.875	0.534	2,356,736,392	580,792,183
HW-TSC	0.494	0.953	0.612	470,693,617	117,653,760
BASELINE	0.455	0.970	0.576	2,280,011,066	564,527,011
aiXplain	0.376	0.995	0.747	368,857,948	12,345

Table 17: Official results of the WMT22 Quality Estimation Task 1 **MQM** for the **English-German** dataset. Teams marked with "•" are the winners, as they are not significantly outperformed by any other system according to the Williams Significance Test (Williams, 1959). Baseline systems are highlighted in grey. ‡ indicates Codalab usernames of participants from whom we have not received further information.

Model	**Spearman**	RMSE	MAE	Disk footprint (B)	# Model params
• IST-Unbabel	0.519	0.963	0.531	2,260,743,915	565,139,485
• Alibaba Translate	0.505	0.961	0.590	2,260,733,079	565,137,999
• Papago	0.496	1.428	1.126	2,243,044,839	560,713,447
• NJUNLP	0.474	0.997	0.666	3,264,730,349	560,145,557
lp_sunny ‡	0.453	0.915	0.548	2,356,736,392	580,792,183
BJTU-Toshiba	0.434	1.011	0.659	2,239,711,849	559,893,507
HW-TSC	0.433	1.257	0.809	2,260,780,823	565,137,436
aiXplain	0.338	1.116	0.785	368,857,948	12,345
BASELINE	0.333	1.051	0.606	2,280,011,066	564,527,011

Table 18: Official results of the WMT22 Quality Estimation Task 1 **MQM** for the **English-Russian** dataset. Teams marked with "•" are the winners, as they are not significantly outperformed by any other system according to the Williams Significance Test (Williams, 1959). Baseline systems are highlighted in grey. ‡ indicates Codalab usernames of participants from whom we have not received further information.

Model	**Spearman**	RMSE	MAE	Disk footprint (B)	# Model params
• HW-TSC	0.369	1.163	0.770	2,260,780,823	565,137,436
• IST-Unbabel	0.348	1.073	0.559	2,260,735,089	583,891,109
• Alibaba Translate	0.347	0.989	0.490	2,260,733,079	565,137,999
• Papago	0.325	0.980	0.397	2,243,044,839	560,095,633
• BJTU-Toshiba	0.299	1.128	0.612	1,736,199,083	434,015,235
lp_sunny ‡	0.298	1.064	0.525	2,356,736,392	580,792,183
NJUNLP	0.296	0.999	0.476	3,264,730,349	560,145,557
aiXplain	0.194	1.481	1.079	368,857,948	12,345
BASELINE	0.164	1.102	0.543	2,280,011,066	564,527,011

Table 19: Official results of the WMT22 Quality Estimation Task 1 **MQM** for the **Chinese-English** dataset. Teams marked with "•" are the winners, as they are not significantly outperformed by any other system according to the Williams Significance Test (Williams, 1959). Baseline systems are highlighted in grey. ‡ indicates Codalab usernames of participants from whom we have not received further information.

C Official Results of the WMT22 Quality Estimation Task 1 (Word-level)

Tables 20, 21, 22, 23, 24, 25, 26, 27, 28 and 29 show the results for all language pairs and the multilingual variants, ranking participating systems best to worst using Matthews correlation coefficient (MCC) as primary key for each of these cases.

Model	**MCC**	Recall	Precision	Disk footprint (B)	# Model params
• IST-Unbabel	0.341	0.466	0.810	2,260,744,555	565,139,485
Papago	0.317	0.422	0.787	2,241,394,304	560,301,035
BASELINE	0.235	0.356	0.765	2,280,011,066	564,527,011

Table 20: Official results of the WMT22 Quality Estimation Task 1 (word-level) for the **Multilingual** task. Teams marked with "•" are the winners, as they are not significantly outperformed by any other system according to the Williams Significance Test (Williams, 1959). Baseline systems are highlighted in grey.

Model	**MCC**	Recall	Precision	Disk footprint (B)	# Model params
• IST-Unbabel	0.361	0.494	0.830	2,260,744,555	565,139,485
• Papago	0.343	0.451	0.858	2,241,394,304	560,301,035
BASELINE	0.257	0.378	0.838	2,280,011,066	564,527,011
HW-TSC	0.218	0.404	0.628	2,336,352,552	612,368,384

Table 21: Official results of the WMT22 Quality Estimation Task 1 (word-level) for the **Multilingual w/o English-Yoruba** task. Teams marked with "•" are the winners, as they are not significantly outperformed by any other system according to the Williams Significance Test (Williams, 1959). Baseline systems are highlighted in grey.

Model	**MCC**	Recall	Precision	Disk footprint (B)	# Model params
• IST-Unbabel	0.436	0.578	0.852	2,260,744,555	565,139,485
• HW-TSC	0.424	0.570	0.848	2,260,780,823	565,137,436
• Papago	0.396	0.549	0.739	2,240,570,795	560,095,834
BASELINE	0.325	0.426	0.870	2,280,011,066	564,527,011

Table 22: Official results of the WMT22 Quality Estimation Task 1 (word-level) for the **English-Czech** dataset. Teams marked with "•" are the winners, as they are not significantly outperformed by any other system based on randomisation tests with Bonferroni correction (Yeh, 2000). Baseline systems are highlighted in grey.

Model	**MCC**	Recall	Precision	Disk footprint (B)	# Model params
• HW-TSC	0.258	0.497	0.728	2,260,780,823	565,137,436
• Papago	0.257	0.502	0.699	2,241,394,304	560,301,035
• IST-Unbabel	0.238	0.491	0.687	2,260,743,979	565,139,485
BASELINE	0.175	0.375	0.795	2,280,011,066	564,527,011

Table 23: Official results of the WMT22 Quality Estimation Task 1 (word-level) for the **English-Japanese** dataset. Teams marked with "•" are the winners, as they are not significantly outperformed by any other system based on randomisation tests with Bonferroni correction (Yeh, 2000). Baseline systems are highlighted in grey. "Rank" indicates the averaged ranking of participants with regards to all metrics (including memory print and number of parameters) and not the final shared task ranking which is decided according to MCC.

Model	**MCC**	Recall	Precision	Disk footprint (B)	# Model params
• Papago	0.418	0.420	0.951	2,241,394,304	560,301,035
• NJUNLP	0.412	0.472	0.939	3,264,730,349	560,145,557
• IST-Unbabel	0.392	0.414	0.947	2,260,744,107	565,139,485
HW-TSC	0.351	0.428	0.917	2,260,780,823	565,137,436
BASELINE	0.306	0.282	0.946	2,280,011,066	564,527,011

Table 24: Official results of the WMT22 Quality Estimation Task 1 (word-level) for the **English-Marathi** dataset. Teams marked with "•" are the winners, as they are not significantly outperformed by any other system based on randomisation tests with Bonferroni correction (Yeh, 2000). Baseline systems are highlighted in grey.

Model	**MCC**	Recall	Precision	Disk footprint (B)	# Model params
• Papago	0.429	0.762	0.660	2,241,394,304	560,301,035
• IST-Unbabel	0.425	0.779	0.555	2,260,744,107	565,139,485
• NJUNLP	0.421	0.744	0.677	3,264,730,349	560,145,557
BASELINE	0.402	0.769	0.567	2,280,011,066	564,527,011
HW-TSC	0.353	0.759	0.395	2,260,780,823	565,137,436

Table 25: Official results of the WMT22 Quality Estimation Task 1 (word-level) for the **Khmer-English** dataset. Teams marked with "•" are the winners, as they are not significantly outperformed by any other system based on randomisation tests with Bonferroni correction (Yeh, 2000). Baseline systems are highlighted in grey.

Model	**MCC**	Recall	Precision	Disk footprint (B)	# Model params
• IST-Unbabel	0.424	0.691	0.733	2,260,744,107	565,139,485
Papago	0.374	0.646	0.723	2,241,394,304	560,301,035
BASELINE	0.359	0.695	0.628	2,280,011,066	564,527,011
HW-TSC	0.358	0.699	0.597	2,260,780,823	565,137,436

Table 26: Official results of the WMT22 Quality Estimation Task 1 (word-level) for the **Pashto-English** dataset. Teams marked with "•" are the winners, as they are not significantly outperformed by any other system based on randomisation tests with Bonferroni correction (Yeh, 2000). Baseline systems are highlighted in grey.

Model	**MCC**	Recall	Precision	Disk footprint (B)	# Model params
• NJUNLP	0.352	0.351	0.980	3,264,730,349	560,145,557
• Papago	0.319	0.336	0.960	2,241,394,304	560,301,035
• IST-Unbabel	0.303	0.317	0.956	2,260,744,107	565,139,485
HW-TSC	0.274	0.292	0.954	2,260,780,823	565,137,436
BASELINE	0.182	0.213	0.970	2,280,011,066	564,527,011

Table 27: Official results of the WMT22 Quality Estimation Task 1 (word-level) for the **English-German** dataset. Teams marked with "•" are the winners, as they are not significantly outperformed by any other system based on randomisation tests with Bonferroni correction (Yeh, 2000). Baseline systems are highlighted in grey.

Model	**MCC**	Recall	Precision	Disk footprint (B)	# Model params
• IST-Unbabel	0.427	0.468	0.958	2,260,743,915	565,139,485
• Papago	0.421	0.381	0.966	2,241,394,304	560,713,447
• NJUNLP	0.390	0.440	0.949	3,264,730,349	560,145,557
HW-TSC	0.343	0.396	0.945	2,260,780,823	565,137,436
BASELINE	0.203	0.144	0.960	2,280,011,066	564,527,011

Table 28: Official results of the WMT22 Quality Estimation Task 1 (word-level) for the **English-Russian** dataset. Teams marked with "•" are the winners, as they are not significantly outperformed by any other system based on randomisation tests with Bonferroni correction (Yeh, 2000). Baseline systems are highlighted in grey.

Model	**MCC**	Recall	Precision	Disk footprint (B)	# Model params
• IST-Unbabel	0.360	0.327	0.966	2,260,743,915	565,139,485
• Papago	0.351	0.338	0.973	2,241,394,304	560,713,447
• NJUNLP	0.308	0.303	0.988	3,264,730,349	560,145,557
HW-TSC	0.246	0.181	0.910	2,260,780,823	565,137,436
BASELINE	0.104	0.123	0.965	2,280,011,066	564,527,011

Table 29: Official results of the WMT22 Quality Estimation Task 1 (word-level) for the **Chinese-English** dataset. Teams marked with "•" are the winners, as they are not significantly outperformed by any other system based on randomisation tests with Bonferroni correction (Yeh, 2000). Baseline systems are highlighted in grey.

D Official Results of the WMT22 Quality Estimation Task 2 (Explainable QE)

Tables 30, 31, 32, 33, 34, 35, 36, 37 and 38 show the results for all language pairs, ranking participating systems best to worst using "Recall at Top-K" on target sentences as primary key for each of these cases.

Model	Word-level (Target sentence)			Sentence-level	
	Recall at Top-K	AUC	AP	Pearson's	Spearman's
• IST-Unbabel	0.561	0.725	0.659	0.548	0.511
• HW-TSC	0.536	0.709	0.632	0.314	0.323
BASELINE (OpenKiwi+LIME)	0.417	0.537	0.500	0.342	0.352
BASELINE (Random)	0.363	0.493	0.453	0.011	0.016

Table 30: Official results of the WMT22 Quality Estimation Task 2 for the **English-Czech** dataset. Teams marked with "•" correspond to the winning submissions, as they are not significantly outperformed by any other system based on randomisation tests with Bonferroni correction (Yeh, 2000). Baseline systems are highlighted in grey.

Model	Word-level (Target sentence)			Sentence-level	
	Recall at Top-K	AUC	AP	Pearson's	Spearman's
• IST-Unbabel	0.466	0.641	0.557	0.252	0.243
• HW-TSC	0.462	0.651	0.547	0.132	0.148
BASELINE (OpenKiwi+LIME)	0.367	0.509	0.451	0.202	0.217
BASELINE (Random)	0.336	0.503	0.418	0.028	0.019

Table 31: Official results of the WMT22 Quality Estimation Task 2 for the **English-Japanese** dataset. Teams marked with "•" correspond to the winning submissions, as they are not significantly outperformed by any other system based on randomisation tests with Bonferroni correction (Yeh, 2000). Baseline systems are highlighted in grey.

Model	Word-level (Target sentence)			Sentence-level	
	Recall at Top-K	AUC	AP	Pearson's	Spearman's
• IST-Unbabel	0.317	0.667	0.448	0.585	0.467
• HW-TSC	0.280	0.625	0.412	0.317	0.426
BASELINE (OpenKiwi+LIME)	0.194	0.479	0.310	0.336	0.372
BASELINE (Random)	0.167	0.489	0.296	0.043	0.017

Table 32: Official results of the WMT22 Quality Estimation Task 2 for the **English-Marathi** dataset. Teams marked with "•" correspond to the winning submissions, as they are not significantly outperformed by any other system based on randomisation tests with Bonferroni correction (Yeh, 2000). Baseline systems are highlighted in grey.

Model	Word-level (Target sentence)			Sentence-level	
	Recall at Top-K	AUC	AP	Pearson's	Spearman's
• IST-Unbabel	0.390	0.747	0.511	0.416	0.459
HW-TSC	0.313	0.686	0.422	0.369	0.426
BASELINE (Random)	0.148	0.527	0.256	0.022	0.015
BASELINE (OpenKiwi+LIME)	0.135	0.428	0.230	0.252	0.330

Table 33: Official results of the WMT22 Quality Estimation Task 2 for the **English-Russian** dataset. Teams marked with "•" correspond to the winning submissions, as they are not significantly outperformed by any other system based on randomisation tests with Bonferroni correction (Yeh, 2000). Baseline systems are highlighted in grey.

Model	Word-level (Target sentence)			Sentence-level	
	Recall at Top-K	AUC	AP	Pearson's	Spearman's
• IST-Unbabel	0.365	0.776	0.490	0.559	0.553
HW-TSC	0.252	0.689	0.361	0.375	0.435
BASELINE (Random)	0.124	0.504	0.212	-0.049	-0.043
BASELINE (OpenKiwi+LIME)	0.074	0.442	0.172	0.370	0.414

Table 34: Official results of the WMT22 Quality Estimation Task 2 for the **English-German** dataset. Teams marked with "•" correspond to the winning submissions, as they are not significantly outperformed by any other system based on randomisation tests with Bonferroni correction (Yeh, 2000). Baseline systems are highlighted in grey.

Model	Word-level (Target sentence)			Sentence-level	
	Recall at Top-K	AUC	AP	Pearson's	Spearman's
• IST-Unbabel	0.234	0.671	0.359	0.309	0.321
BASELINE (Random)	0.144	0.514	0.246	-0.086	-0.101
BASELINE (OpenKiwi+LIME)	0.111	0.442	0.218	0.085	0.160

Table 35: Official results of the WMT22 Quality Estimation Task 2 for the **English-Yoruba** dataset. Teams marked with "•" correspond to the winning submissions, as they are not significantly outperformed by any other system based on randomisation tests with Bonferroni correction (Yeh, 2000). Baseline systems are highlighted in grey.

Model	Word-level (Target sentence)			Sentence-level	
	Recall at Top-K	AUC	AP	Pearson's	Spearman's
• HW-TSC	0.686	0.720	0.751	0.601	0.610
IST-Unbabel	0.665	0.660	0.751	0.617	0.598
UT-QE	0.622	0.628	0.694	0.222	0.190
BASELINE (OpenKiwi+LIME)	0.580	0.520	0.653	0.417	0.430
BASELINE (Random)	0.565	0.498	0.633	-0.048	-0.045

Table 36: Official results of the WMT22 Quality Estimation Task 2 for the **Khmer-English** dataset. Teams marked with "•" correspond to the winning submissions, as they are not significantly outperformed by any other system based on randomisation tests with Bonferroni correction (Yeh, 2000). Baseline systems are highlighted in grey.

Model	Word-level (Target sentence)			Sentence-level	
	Recall at Top-K	AUC	AP	Pearson's	Spearman's
• HW-TSC	0.715	0.716	0.777	0.393	0.418
IST-Unbabel	0.672	0.612	0.740	0.593	0.601
UT-QE	0.668	0.643	0.727	0.409	0.402
BASELINE (OpenKiwi+LIME)	0.615	0.503	0.676	0.378	0.403
BASELINE (Random)	0.614	0.497	0.662	-0.002	0.002

Table 37: Official results of the WMT22 Quality Estimation Task 2 for the **Pashto-English** dataset. Teams marked with "•" correspond to the winning submissions, as they are not significantly outperformed by any other system based on randomisation tests with Bonferroni correction (Yeh, 2000). Baseline systems are highlighted in grey.

Model	Word-level (Target sentence)			Sentence-level	
	Recall at Top-K	AUC	AP	Pearson's	Spearman's
• IST-Unbabel	0.379	0.785	0.475	0.103	0.190
HW-TSC	0.220	0.652	0.315	0.097	0.159
BASELINE (Random)	0.093	0.463	0.162	0.041	-0.010
BASELINE (OpenKiwi+LIME)	0.048	0.388	0.126	-0.007	0.159

Table 38: Official results of the WMT22 Quality Estimation Task 2 for the **Chinese-English** dataset. Teams marked with "•" correspond to the winning submissions, as they are not significantly outperformed by any other system based on randomisation tests with Bonferroni correction (Yeh, 2000). Baseline systems are highlighted in grey.

E Official Results of the WMT22 Quality Estimation Task 3 (Critical Error Detection)

Tables 39, 40, 41 and 42 show the results for all language pairs and the multilingual variants, ranking participating systems best to worst using Matthews correlation coefficient (MCC) as primary key for each of these cases.

Model	**MCC**	Recall	Precision	Disk footprint (B)	# Model params
• IST-Unbabel	0.564	0.619	0.619	2,260,735,025	565,137,435
BASELINE	0.074	0.191	0.191	2,277,430,785	569,330,715

Table 39: Official results of the WMT22 Quality Estimation Task 3 (Critical Error Detection) for the **English-German (Constrained)** dataset. Teams marked with "•" are the winners, as they are not significantly outperformed by any other system based on randomisation tests with Bonferroni correction (Yeh, 2000). Baseline systems are highlighted in grey.

Model	**MCC**	Recall	Precision	Disk footprint (B)	# Model params
• KU X Upstage	0.964	0.968	0.968	2,244,861,551	559,890,432
BASELINE	0.855	0.873	0.873	2,260,734,129	565,137,435
aiXplain	0.219	0.318	0.318	2,052,963,739	12,345

Table 40: Official results of the WMT22 Quality Estimation Task 3 (Critical Error Detection) for the **English-German (UNconstrained)** dataset. Teams marked with "•" are the winners, as they are not significantly outperformed by any other system based on randomisation tests with Bonferroni correction (Yeh, 2000). Baseline systems are highlighted in grey.

Model	**MCC**	Recall	Precision	Disk footprint (B)	# Model params
• IST-Unbabel	0.721	0.761	0.761	2,260,735,025	565,137,435
BASELINE	-0.001	0.141	0.141	2,277,430,785	569,330,715

Table 41: Official results of the WMT22 Quality Estimation Task 3 (Critical Error Detection) for the **Portuguese-English (Constrained)** dataset. Teams marked with "•" are the winners, as they are not significantly outperformed by any other system based on randomisation tests with Bonferroni correction (Yeh, 2000). Baseline systems are highlighted in grey.

Model	**MCC**	Recall	Precision	Disk footprint (B)	# Model params
• KU X Upstage	0.984	0.986	0.986	2,244,861,551	559,890,432
BASELINE	0.934	0.944	0.944	2,260,734,129	565,137,435
aiXplain	0.179	0.296	0.296	9,395,107	12,345

Table 42: Official results of the WMT22 Quality Estimation Task 3 (Critical Error Detection) for the **Portuguese-English (UNconstrained)** dataset. Teams marked with "•" are the winners, as they are not significantly outperformed by any other system based on randomisation tests with Bonferroni correction (Yeh, 2000). Baseline systems are highlighted in grey.