Findings of the WMT 2021 Shared Task on Quality Estimation

Lucia Specia,[1,2] Frédéric Blain,[2,3] Marina Fomicheva,[2] Chrysoula Zerva,[4,5]
Zhenhao Li,[1] Vishrav Chaudhary,[6] André F. T. Martins[4,5,7]

[1]Imperial College London, [2]University of Sheffield, [3]University of Wolverhampton,
[4]Instituto de Telecomunicações, [5]Instituto Superior Técnico, [6]Facebook AI, [7]Unbabel
{l.specia,m.fomicheva}@sheffield.ac.uk, f.blain@wlv.ac.uk
chrysoula.zerva@tecnico.ulisboa.pt, zhenhao.li18@imperial.ac.uk,
vishrav@fb.com, andre.t.martins@tecnico.ulisboa.pt

Abstract

We report the results of the WMT 2021 shared task on Quality Estimation, where the challenge is to predict the quality of the output of neural machine translation systems at the word and sentence levels. This edition focused on two main novel additions: (i) prediction for unseen languages, i.e. zero-shot settings, and (ii) prediction of sentences with catastrophic errors. In addition, new data was released for a number of languages, especially post-edited data. Participating teams from 19 institutions submitted altogether 1263 systems to different task variants and language pairs.

1 Introduction

The 10th edition of the shared task on Quality Estimation (QE) builds on its previous editions to further benchmark methods for estimating the quality of neural machine translation (MT) output at runtime, without the use of reference translations. It includes the (sub)tasks of word-level and sentence-level estimation. The document-level task was removed from this edition, since it has not attracted many participants in previous editions. Important elements introduced this year are: a new sentence-level task where sentences are annotated with a binary label reflecting whether or not they contain a critical error that could lead to catastrophic consequences; new test data for languages that are not covered by any training set for zero-shot prediction (both direct assessment and post-editing based labels, at sentence and word levels. scores instead of labels based on post-editing; a new multilingual sentence-level dataset mainly from Wikipedia articles, where the source articles can be retrieved for document-wide context; the availability of NMT models to explore system-internal information for the task.

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

- To extend the MLQE-PE public benchmark datasets;
- To investigate new language independent approaches especially for zero-shot prediction;
- To study the feasibility of unsupervised approaches especially for zero-shot prediction; and
- To create a new task focusing on critical error detection.

We have three subtasks: Task 1 aims at predicting a variant of DA scores at sentence level (Section 2.1); Task 2 aims at predicting post-editing effort scores at both sentence and word levels, i.e. words that need editing, as well as missing words and incorrect source words (Section 2.2); Task 3 aims at predicting a binary label at sentence level to indicate whether the sentence contains one or more critical errors (Section 2.3).

Tasks make use of large datasets annotated by professional translators with either 0-100 DA scoring, post-edition, or critical error flagging. The text domains and languages vary for each subtask. Neural MT systems were built on freely available data using an open-source toolkits to produce translations. We provide new training and test datasets for Tasks 2 and 3, new test sets for Task 1, as well as new *zero-shot* test sets for Tasks 1 and 2. The datasets and models released are publicly available. Participants are also allowed to explore any additional data and resources deemed relevant.

Baseline systems were entered in the platform by the task organisers (Section 3). The shared task uses CodaLab as submission platform, where participants (Section 4) could submit up to 30 systems for each task and language pair, except for the multilingual track of Tasks 1 and 2 (up to 10 systems). Results for all tasks evaluated according to standard metrics are given in Section 5, which this year also included model size. A discussion on the main

Proceedings of the Sixth Conference on Machine Translation (WMT), pages 684–725
November 10–11, 2021. ©2021 Association for Computational Linguistics

goals and findings from this year's task is presented in Section 6.

2 Subtasks

In what follows, we give a brief description for each subtask, including the datasets provided for them.

2.1 Task 1: Predicting sentence-level DA

This task consists in scoring translation sentences according to their perceived quality score – which we refer to as direct assessment (DA). For that, we use the same training sets as last year's Task 1 (Specia et al., 2020), and provided new test sets for all low, medium and high-resource languages:

- English→German (En-De),
- English→Chinese (En-Zh),
- Russian→English (Ru-En),
- Romanian→English (Ro-En),
- Estonian→English (Et-En),
- Sinhala→English (Si-En), and
- Nepali→English (Ne-En).

This data was produced in the same way as the data for last year, with sentences sample from Wikipedia (or Wikipedia and Reddit for Ru-En) and translated by a fairseq Transformer (Ott et al., 2019) bilingual model.

In addition, we provide new test sets for four other languages, for which training data was not provided. The goal was to test the performance of QE models under zero-shot settings. The new test sets contain source Wikipedia sentences sampled in the same way as the previous data, but translated by the ML50 fairseq multilingual Transformer model (Tang et al., 2020),[1] which had been found to perform well especially for low-resource languages. The following languages were used:

- English→Czech (En-Cs),
- English→Japanese (En-Ja),
- Pashto→English (Ps-En), and
- Khmer→English (Km-En),

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. DA scores were standardised using the z-score by rater. Participating systems are required to score sentences according to z-standardised DA scores. Statistics on the dataset are shown in Table 1. This dataset part of the MLQE-PE dataset and more details are given in Fomicheva et al. (2020). The complete data can be downloaded from the public repository.[2]

Participation was encouraged for each language pair and also for the **multilingual variant** of the task, where submissions had to include predictions for all six Wikipedia-based language pairs (all except Ru-En). The latter aimed at fostering work on language-independent models, as well as models that can leverage data from multiple languages.

2.2 Task 2: Predicting post-editing effort

This task concerns scoring translations according to the proportion of the words that need to be edited to obtain a correct translation. The scores are generated using Human-mediated Translation Edit Rate (HTER) (Snover et al., 2006), i.e. calculating the minimum edit distance between the machine translation and its manually post-edited version, as well as detecting where errors are in the translation of source sentences. It comprises two sub-tasks, a sentence level one where the targets are the HTER scores per segment and a word level task where the targets are word level OK/BAD tags to signify the correctness of words and gaps in the source and translation sentences. Both sub-tasks use the same languages pairs and splits described for Task 1 in Table 1. Details on the data, such as label distributions, can be found in Fomicheva et al. (2020).

Sentence-level post-editing effort The label for this task is the percentage of edits that need to be fixed (HTER). The data used for this task is the PE annotations and corresponding HTER scores from the MLQE-PE dataset (Fomicheva et al., 2020). HTER labels are computed using TERCOM,[3] with

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

[2] https://github.com/sheffieldnlp/mlqe-pe

[3] https://github.com/jhclark/tercom

Language Pairs	Sentences Train / Dev / Test21	Tokens Train / Dev / Test21	DA	PE	CE
En-De	7,000 / 1,000 / 1,000	114,980 / 16,519 / 16,545	✓	✓	
En-Zh	7,000 / 1,000 / 1,000	115,585 / 16,307 / 16,637	✓	✓	
Ru-En	7,000 / 1,000 / 1,000	82,229 / 11,992 / 11,650	✓	✓	
Ro-En	7,000 / 1,000 / 1,000	120,198 / 17,268 / 17,359	✓	✓	
Et-En	7,000 / 1,000 / 1,000	98,080 / 14,423 / 14,044	✓	✓	
Ne-En	7,000 / 1,000 / 1,000	104,934 / 15,144 / 15,017	✓	✓	
Si-En	7,000 / 1,000 / 1,000	109,515 / 15,708 / 15,709	✓	✓	
Ps-En	– / – / 1,000	– / – / 27,045	✓	✓	
Km-En	– / – / 1,000	– / – / 21,981	✓	✓	
En-Ja	– / – / 1,000	– / – / 20,626	✓	✓	
En-Cs	– / – / 1,000	– / – / 20,394	✓	✓	
En-Cs	7,476 / 1,000 / 1,000	122,275 / 16,270 / 16,106			✓
En-De	7,878 / 1,000 / 1,000	127,778 / 16,114 / 16,371			✓
En-Ja	7,658 / 1,000 / 1,000	126,307 / 16,400 / 16,412			✓
En-Zh	6,859 / 1,000 / 1,000	110,717 / 16,283 / 15,989			✓

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.

default settings (tokenised, case insensitive, exact matching only) with scores capped to 1.

Word-level errors This sub-task focuses on detecting word-level errors in the MT output. The goal in this case is to annotate each token with binary correctness (OK/BAD) tags. The token-level annotations include the annotation of gaps, which allows us to account for omission errors. All annotations are produced with respect to a post-edited sentence, which is treated as the ground truth reference. Similarly to the sentence-level tasks, the MLQE-PE data is used for all language pairs (see Table 1). The following types of labels are used:

- **Source side:** Each word in the source side is labelled as OK (correctly translated) or BAD (caused a translation error).
- **Target side:** Each word in the target side is labelled as OK (a correct translation) or BAD (should be replaced or deleted). Additionally, we consider gap 'tokens' at the beginning of the sentence, at the end and between each two words. They are labelled OK if no word should be inserted in that position (according to the post-edited version), and BAD otherwise.

2.3 Task 3: Predicting Catastrophic Errors

This is a new task introduced this year. It aims to predict a sentence-level binary score indicating whether a translation contains (at least one) critical error (CE). Translations with such errors are defined as translations that deviate in meaning as compared to the source sentence in such a way that they are misleading and may carry health, safety, legal, reputation, religious or financial implications. Meaning deviations from the source sentence can happen in three ways:

- **Mistranslation**: critical content is translated incorrectly into a different meaning, or not translated (i.e. it remains in the source language) or translated into gibberish.
- **Hallucination**: critical content that is not in the source is introduced in the translation, for example, profanity words are introduced that were not in the source.
- **Deletion**: critical content that is in the source sentence is not present in the translation. For example, the source sentence may contain a negation or hateful word that is removed in the translation.

We focus on the following set of critical error categories:

- **TOX.** Deviation in toxicity (hate, violence or profanity) be against an individual or a group (a religion, race, gender, etc.). This error can happen because toxicity is introduced in the translation when it is not in the source, deleted in the translation when it was in the source, or mistranslated into different (toxic or not) words, or not translated at all (i.e. the toxicity remains in the source language or it is transliterated).

- **SAF.** Deviation in health or safety risks, i.e. the translation contains errors that may bring a risk to the reader. This issue can happen because content is introduced in the translation when it is not in the source, deleted in the translation when it was in the source, or mistranslated into different words, or not translated at all (i.e. it remains in the source language).

- **NAM.** Deviation in named entities. A named entity (people, organisation, location, etc.) is deleted, mistranslated by either another incorrect named entity or a common word or gibberish, or left untranslated when it should be translated, or transliterated where the transliteration makes no sense in the target language (i.e. the reader cannot recover the actual named entity from it), or introduced when it was not in the source text. If the named entity is translated partially correctly but one can still understand that it refers to the same entity, it should not be an error.

- **SEN.** Deviation in sentiment polarity or negation. The MT either introduces or removes a negation (with or without an explicit negation word), or reverses the sentiment of the sentence (e.g. a negative sentence becomes positive or vice-versa). We note that SEN errors do not always involve a full negation, for example, replacing "possibly" with "with certainty" constitutes a SEN error.

- **NUM.** Deviation in units/time/date/numbers. The MT translated a number/date/time or unit incorrectly (or translated it as gibberish), or removed it, which could lead someone to miss an appointment, get lost, etc.

Data for this task was annotated at the word/span level by professional translators not only for the presence of an error, but also with the error category. Each instance was annotated by three professional translators using a modified version of MT-EQuAl.[4] We instructed the translators to ignore other translation errors, be them critical (there may be other types of critical errors outside these five categories) or non-critical, e.g. minor grammatical or typographical errors. We also instructed them to indicate source sentences that were unintelligible, or translation sentences that contained too many errors to be annotated. Figure 1 shows three examples of different error annotations for the translations.

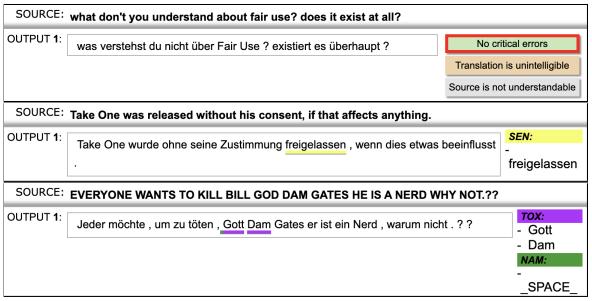


Figure 1: Example of fine-grained sentence annotation. Spans in the same colour belong to the same catastrophic error type. In the first case, the translation contains no critical error; in the second case, the translation contains only one SEN error; in the last case, the translation contains two errors: one TOX and one NAM (the space is annotated to indicate a missing named entity).

For this the first edition of this task, we aggregated these labels in two ways: First, for each of the three annotated versions of a sentence, we extrapolated the word-level labels into a sentence-level label: if the sentence contained at least one critical error, it was annotated as critical. Second: we took the majority sentence-level label from the three annotators to create a single sentence-level label for each sentence, resulting in the following binary labels:[5]

- ERR: the translated sentence contains at least one (any) token or whitespace (for deletion errors) annotated with a critical error in any categories, according to at least 2 out of 3 annotators, or otherwise

- NOT: the sentence does not contain any token with a critical error.

[4] https://mt4cat.fbk.eu/software/mt-equal

[5] We removed from the dataset sentences that had been annotated by the majority as having an unintelligible source or a translation with too many errors.

Thus, the task does not expect the errors to be categorised or to have their spans identified in the sentence, but rather to have a binary prediction on a sentence basis. For example, the first sample in Figure 1 would have resulted in the sentence labelled as NOT by that annotator, while the last two samples would have resulted in the sentence labelled as ERR.

An initial set of 10K English samples for training, development and test data was created from Wikipedia comments, extracted from two sources: the Jigsaw Toxic Comment Classification Challenge[6] and the Wikipedia Comments Corpus.[7] Machine translations were generated by the ML50 fairseq multilingual translation model for the following languages:

- English-Czech,
- English-Japanese,
- English-Chinese, and
- English-German.

After filtering for unintelligible source sentences and translations with too many errors, the statistics for the resulting splits are presented in Table 1. As expected, critical errors are rare. Give the nature of this dataset (user generated content with high chances of toxicity, named entities, etc.), we observed a fairly large proportion of sentences with such errors. Nevertheless, the distribution of labels is skewed towards the NOT class. The proportion of instances with NOT labels in the training set (similar for dev and test sets) is as follows: 83% for En-Cs, 72% for En-De, 91% or En-Ja, and 84% for En-Zh.

3 Baseline systems

Sentence-level baseline systems: For Tasks 1 and 2, both 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., 2020). Both baselines are implemented in OpenKiwi (Kepler et al., 2019) and trained using the concatenated train portions of the data for training (combining all 7 language pairs) and the concatenated development portions of the data for validation/early-stopping. In all cases, the XLM-RoBERTa transformer was used for the encoding (predictor) part of the architecture, using `xlm-roberta-base` for all experiments. The XLM-RoBERTa encoder is initially trained on the concatenated train and development segments using ULM fine-tuning (Howard and Ruder, 2018) and then this fine-tuned encoder is used in the full predictor-estimator model which is fine-tuned separately for each task scores (DA or HTER).

Word-level baseline systems: For Task 2, we used the same architecture and encoder as above, but it was trained to predict jointly word-level OK/BAD tags and sentence-level HTER scores.

Catastrophic error baseline system: The baseline model for Task 3 follows the MonoTransQuest architecture proposed by Ranasinghe et al. (2020) for sentence-level classification. As input, the model takes a sequence of tokens including the `[CLS]` token, and the source and translated sentence tokens, separated by a `[SEP]` token. This string is fed into a transformer encoder and the output of the encoder is given to a classification head where cross-entropy is adopted as the loss function. We use the pre-trained XLM-RoBERTa-base released by HuggingFace's model repository (**?**) for the implementation.

4 Participants

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 two submissions for each task variant and language pair. In the descriptions below, participation in specific tasks is denoted by a task identifier (T1 = Task 1, T2 = Task 2, T3 = Task 3).

Bergamot (T1): Bergamot explores the use of a teacher-student knowledge distillation framework to transfer knowledge from a strong QE teacher model to a much smaller student model with a different, shallower architecture. Namely, the system distill a large and powerful QE model TransQuest [1] based on XLM-Roberta into a small BiRNN-based DeepQuest model [2]. The predictions from a teacher QE model trained on MLQE data [3] is used to train the lightweight student. Additionally, the system employs data augmentation through teacher predictions on monolingual data sampled from Wikipedia following

[6] https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge/data

[7] https://meta.wikimedia.org/wiki/Research:Detox/Data_Release#Wikipedia_Comments_Corpus

ID	Participating team	
Bergamot	University of Sheffield & Imperial College London & University of Wolverhampton, UK	(Gajbhiye et al., 2021)
Bergamot-UTartu	University of Tartu, Estonia	(Yankovskaya and Fishel, 2021)
ENSBRT	University of Illinois at Chicago & IQVIA, USA	(Chowdhury et al., 2021)
HW-TSC	Huawei Translation Services Center & Nanjing University, China	(Chen et al., 2021)
IST-Unbabel	Instituto de Telecomunicações Lisbon & Instituto Superior Técnico Lisbon & Unbabel, Portugal	(Zerva et al., 2021)
JHU-Microsoft	Johns Hopkins University & Microsoft	(Ding et al., 2021)
LAMA-ICL	LAMA - Imperial College London, UK	(Jiang et al., 2021)
NICT Kyoto	National Institute of ICT, Japan	(Rubino et al., 2021)
Papago	Naver, Republic of Korea	(Lim et al., 2021)
POSTECH	Pohang University of Science and Technology, Republic of Korea	(Heo et al., 2021)
QEMind	Alibaba, China	(Wang et al., 2021)
RTM	Boğaziçi University, Turkey	(Biçici, 2021)
SMOB-ECEIIT	Technion - Israel Institute of Technology, Israel	–
TUDa	Technische Universität Darmstadt, Germany	(Geigle et al., 2021)

Table 2: Participants to the WMT21 Quality Estimation shared task.

Task/LP	# submission
Task 1 – Sent-level Direct Assessment	**725**
Multilingual	32
English-German	99
English-Chinese	78
Romanian-English	58
Estonian-English	56
Nepalese-English	52
Sinhala-English	65
Russian-English	54
English-Czech	54
English-Japanese	62
Pashto-English	50
Khmer-English	65
Task 2 – Sent-/Word-level PE Effort	**163/178**
Multilingual	7/–
English-German	37/33
English-Chinese	22/32
Romanian-English	13/14
Estonian-English	13/19
Nepalese-English	6/11
Sinhala-English	7/10
Russian-English	11/11
English-Czech	10/14
English-Japanese	13/14
Pashto-English	6/8
Khmer-English	18/12
Task 3 – Sent-Level Critical Error Det.	**197**
English-German	56
English-Chinese	30
English-Czech	36
English-Japanese	75
Total	**1263**

Table 3: Number of submissions to each sub-task and language-pair at the WMT21 Quality Estimation shared task.

the procedure described in [3]. Further details about the distillation framework used for submission can be found in [4].

Bergamot-UTartu (T1, T2): Bergamot-UTartu proposes CNN-models based on attention weights extracted from NMT systems. For Task 1, they explored three QE models: i) CNN-DA trained on human-labelled data; ii) CNN-BLEURT a "zero-shot" system that requires only synthetic data, for which they used BLEURT scores (Sellam et al., 2020) as training data; iii) CNN-BLEURT+ a fine-tuned version of CNN-BLEURT. For Task 2, CNN-HTER is a model similar to CNN-DA, but trained on the post-editing scores.

ENSBRT (T2): ENSBRT propose a system that is an ensemble of multilingual BERT (mBERT)-based regression models, which are generated by fine-tuning on different input settings. They adapted their system for the zero-shot setting by exploiting target language-relevant language pairs and pseudo-reference translations.

HW-TSC (T1, T2, T3): HW-TSC's submissions in the three sub tasks follow the framework of Predictor-Estimator (Kim et al., 2017), with a pre-trained XLM-Roberta as Predictor and task-specific classifier or regressor as Estimator. They further explore to incorporate additional high-quality translation sentences in the way of multitask learning or encoding it with the Predictor directly. For Task1, they enable the model to jointly learn to score translations with a regression task and to distinguish between translations and additional better translations (i.e. post-edits from Task2 dataset) with a classification task. They also exploit

a data augmentation strategy based on MC dropout to improve zero-shot performance. They ensemble multi results with MC dropout to keep a relatively small number of parameters and model size. For Task 2, they leverage additional translation sentence generated by a NMT system trained for WMT21 News shared task in the way of directly concatenating it with source and original translation. A unified model is trained under multi-task learning framework where losses of source word, translation token and gap, additional translation token and gap, HTER scores are all added up to train the model. For Task 3, they translate source sentences with Google and Baidu translation API. Each new translation is then concatenated to the corresponding source and translation pair, to get a sentence feature. They ensemble the results of three different models and take their majority voting as final result.

IST-Unbabel (T1, T2): For Task 1, IST-Unbabel's system is an ensemble of an XLM-R with stacked adapter layers and an mBART that incorporates different types of uncertainty (annotation uncertainty and MT uncertainty). For Task 2, the submitted system is an ensemble of two XLM-R with adapters (the difference being the XLM-R checkpoint, while one uses the xlm-roberta-large normal checkpoint the other uses an XLM-R checkpoint pertained on data from the metrics shared task). The ensembled checkpoints learn to predict both word level tags and sentence level HTER scores in a multi-task setting.

JHU-Microsoft (T2): The JHU-Microsoft submission focuses on the word-level subtask of Task 2, for which they adopt Levenshtein Transformer (Gu et al. 2019) as their model architecture. The training procedure starts with training a non-autoregressive translation model using a Levenshtein Transformer, with its encoder and decoder initialized with those from the M2M multilingual translation model (Fan et al. 2020). They then fine-tune the model to perform the word-level QE task on the human-annotated training set, or optionally also on automatically generated pseudo-post-editing translation triplets. The final submission is an ensemble of 4-8 best models on the 2020 test set for each language pair, and the ensemble is performed by linear interpolation of scores from each model, with the interpolation weights tuned by the Nelder-Meade method (Nelder and Meade, 1965).

LAMA-ICL (T3): LAMA-ICL's approach builds on cross-lingual pre-trained representations in a sequence classification model. We further improve the base classifier by (i) adding a weighted sampler to deal with unbalanced data and (ii) introducing feature engineering, where features related to toxicity, named-entities and sentiment, which are potentially indicative of critical errors, are extracted using existing tools and integrated to the model in different ways. We train models with one type of feature at a time and ensemble those models that improve over the base classifier on the development (dev) set.

NICT Kyoto (T3): NICT Kyoto submission for the Critical Error Detection task consists in large scale QE pretraining with synthetic data in a multilingual and multimetric setting. A total of six sentence- and word-level quality indicators were involved in continued training of an XLM-R checkpoint using QE oriented training objectives in a multi-task fashion, based on a corpus of 70 million sentence pairs including twelve languages. Fine-tuning on the official dataset was then performed and resulting models from different initializations were ensembled to constitute the final submission.

Papago (T1): Papago's submission is a multilingual Quality Estimation system that explores the combination of pre-trained language models and multi-task Learning architectures. They propose an iterative training pipeline based on pretraining with large amounts of in-domain synthetic data and fine-tuning with gold (labeled) data. They then compress our system via knowledge distillation in order to reduce parameters yet maintain strong performance.

POSTECH (T2): POSTECH's model uses two pre-trained monolingual encoders to first produce monolingual representations of the two input data separately and then exchanges the information of these representations through

two additional cross attention networks. The two pre-trained monolingual encoders are an English ELECTRA and a German ELECTRA, respectively. They fine-tuned their system in two stages: the QE pre-training stage and the QE fine-tuning stage. In the former, they used a large quantity of artificial training data, and the loss value equals to the sum of the losses for the estimated HTER (sentence-level QE), OK-BAD for the source tokens, OK-BAD for the MT tokens, and OK-BAD for the gap tokens in between two MT tokens. In the latter, they only used human labelled training data, and the loss value is one of the four above mentioned loss values, depending on the targeted subtask.

QEMind (T1, T3): QEMind propose novel glass-box QE features to estimate the uncertainty of machine translations and incorporate them into the transfer learning from the large-scale pre-trained model, XLM-Roberta. In addition, three important strategies are particularly utilized for improving the QE system's performance such as multilingual training, data augmentation and model ensemble.

RTM (T1, T2): Referential Translation Machines (RTMs) Superlearner results combine individual machine learning model results via cross-validation on the training set. The combined models guarantee lower error on the validation set than the model that minimises the overall error. A superlearner model improves the results over non-mixture results.

SMOB-ECEIIT (T1): SMOB-ECEIIT's participation is fully unsupervised, as created without using any annotated data or even parallel bilingual data. The system is composed of two novel different methods. The first method is based on soft alignment of multilingual contextual embeddings, generated by pre-trained mBert or XLM-R (depending on the specific language). The soft alignment is calculated by the Sinkhorn distance (Cuturi, 2013), which is an optimal transportation distance with an entropic regularization term. The second method is based on the assumption that word embedding spaces are approximately isometric (Vulić et al., 2020), and on an isometric-invariant method known as Persistent Homology (Edelsbrunner, 2013). Each sentence is represented by the distances between its own word embeddings (either static or contextual). The created distance matrices are compared using the Wasserstein distance between their persistence barcodes (the output of persistent homology computation). Finally, the two methods are linearly combined.

TUDa (T1): TUDa's submissions are produced with pre-trained multilingual language models which they extended to new languages and unseen scripts using recent adapter-based methods.

5 Results

5.1 Task 1

Submissions for Task 1 are **evaluated** against the true z-normalised direct assessment label using Pearson's r correlation score as primary metric. This is what was used for ranking system submissions. Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE) were also computed as secondary metrics. Statistical significance on Pearson r was computed using William's test.[8]

Table 4 summarises the results for all language pairs, as well as the multilingual variant, in terms of Pearson's r correlation with direct assessments, ranking systems by their average performance for all language pairs (using 0 as Pearson score for other languages). In the Appendix, Tables 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19 and 20 provide the detailed results for all language pairs and the multilingual variant, ranking participants by their performance for each of these cases.

Best performers The best system varies slightly across language pairs, with QEMind winning the multilingual task, i.e. the average performance for all language pairs (including zero-shot). Overall, the three top performing systems, QEMind, HW-TSC and IST-Unbabel, perform very closely on average, and also for each given language. The three make use of the XML-R large pre-trained representations in a predictor-estimator fashion, and model ensembling. Another recurring theme is to explore data augmentation (QEMind and HW-TSC) and model uncertainty (QEMind and IST-Unbabel). While the baseline system also uses XLM-R as predictor, it uses its 'base' version, and only the provided 'train' part of the data to train the estimator. In addition, it does not resort to ensembling.

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

Model	Multi	En-De	En-Zh	Ro-En	Et-En	Ne-En	Si-En	Ru-En	En-Cs	En-Ja	Ps-En	Km-En
QEMind	0.675	0.567	**0.603**	**0.908**	**0.812**	**0.867**	**0.596**	**0.806**	**0.582**	**0.359**	**0.647**	**0.679**
HW-TSC	0.665	**0.584**	0.583	0.901	**0.808**	0.858	0.581	0.878	**0.573**	**0.364**	0.622	0.659
IST-Unbabel	0.665	**0.579**	0.586	0.899	0.796	0.856	**0.605**	0.792	**0.577**	**0.355**	0.628	0.650
papago (IKT)	0.658	**0.568**	0.567	0.901	0.759	0.853	**0.595**	0.793	**0.572**	0.332	**0.637**	0.662
TUDa	0.631	0.473	0.558	0.886	0.792	0.834	0.571	0.764	0.545	0.330	0.609	0.639
Inmon‡	0.623	–	–	–	–	–	–	–	0.547	0.297	0.592	0.630
papago (KD)	0.613	0.551	0.553	0.879	0.794	0.823	0.582	0.744	0.497	0.276	0.582	0.625
BASELINE	0.541	0.403	0.525	0.818	0.660	0.738	0.513	0.677	0.352	0.230	0.476	0.562
SMOB-ECEIIT	0.348	0.226	0.131	0.650	0.329	0.544	0.347	0.420	0.195	0.153	0.424	0.409
Bergamot	–	–	0.687	0.544	0.626	0.425	–	–	–	–	–	–
Bergamot-UTartu	–	0.369	–	–	0.547	–	–	–	0.300	–	–	–
RTM	–	0.143	0.248	0.287	0.099	0.127	0.061	0.356	0.104	0.082	–	–

Table 4: Pearson correlation with direct assessments for the submissions to WMT21 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.

To gain a better understanding in the performance of different QE approaches for different language pairs, Figure 2 shows the scatter plots for the baseline and the best performing system for each language pair.

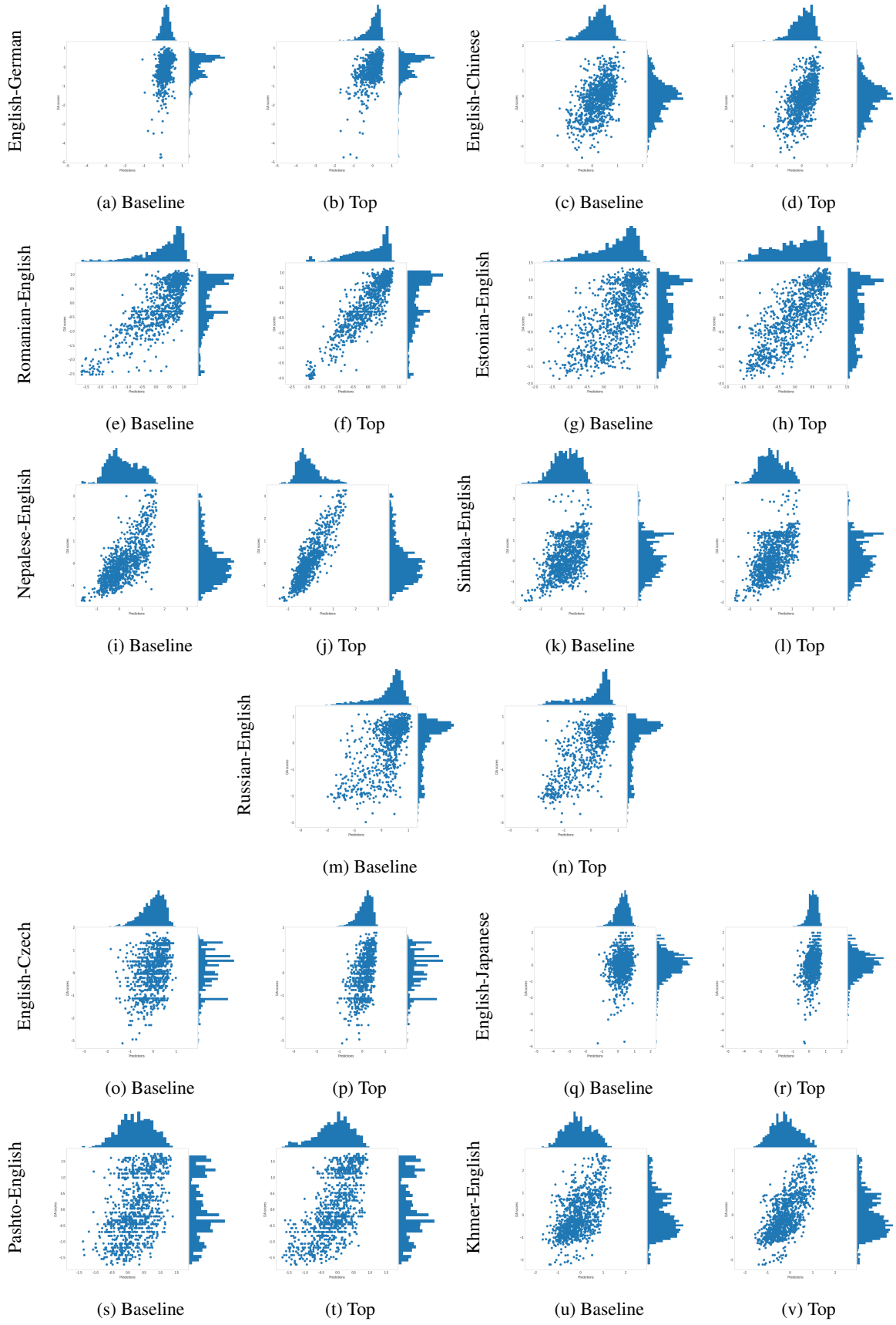
The performance of all except four systems is substantially better than that of the baseline system for all languages. The systems below the baseline correspond to unsupervised systems (Bergamot, Bergamot-UTartu - using NMT glass-box features; and SMOB-ECEIIT, using alignment over XLM-R representations), as well as RTM, which does not rely on pre-trained representations altogether.

Zero-shot languages On the zero-shot languages, the performance was comparable to those of the average non-zero-shot language pairs, except for the En-Ja language pair, where it was substantially lower. Most systems achieved such good performance by relying on multilingual prediction models trained on cross-lingual representations from XLM-R. With En-Ja, we believe there may have been an issue with the segmentation of the Japanese data after translation, which led to annotation issues and/or issues of mapping of characters against the vocabulary of pre-trained language models. We will investigate this further.

High vs low-resource performance Similar to last year, MT quality for the high-resource language pairs, in particular En-De but also En-Zh, proved to be more challenging to predict. This could be an indication of less variability in the MT outputs for these language pairs, given that the NMT models are likely to perform overall well for these languages. This would lead to little variability in perceived MT quality by humans, and thus a harder data to learn from. Interestingly, this also seemed to be the case in the zero-shot QE setting for En-Cs, which is relatively higher resource than Ps-En and Km-En. We observe that these differences in correlation also happen with the HTER predictions for these language pairs (see the analysis of the sentence level task in §5.2 and Table 5).

All medium and low and medium-resource language pairs achieve high correlations, in particular in the supervised settings with Ro-En and Ne-En. This again is an indication of the potential of multilingual or cross-lingual pre-trained representations. It could also indicate that the models (and human annotators to some extent) rely heavily on the target language (English), which is well represented in the pre-trained representations.

High correlations Just like in WMT2020, the very high correlation for some language pairs, particularly for Ro-En (r = 0.91) but also for Ne-En (r = 0.87) could be explained by the fact that there is a number of very low-quality sentences that the QE systems are able to successfully detect. Esp. for Ro-En, we find that they correspond to 'hallucinated' outputs that do not have anything to do with the original sentences. Detecting such cases should be trivial for QE systems, which explains the particularly high correlation values for this language pair.



(a) Baseline (b) Top (c) Baseline (d) Top

(e) Baseline (f) Top (g) Baseline (h) Top

(i) Baseline (j) Top (k) Baseline (l) Top

(m) Baseline (n) Top

(o) Baseline (p) Top (q) Baseline (r) Top

(s) Baseline (t) Top (u) Baseline (v) Top

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

5.2 Task 2

Sentence-level post-editing effort For this task variant, **evaluation** was performed against the true HTER label using the same metrics as in Task 1, with Pearson's r correlation score as the primary metric. Table 5 summarises the results for all language pairs, including the multilingual performance. Systems are ranked by their averaged performance over all language pairs based on the Pearson r coefficient. Statistical significance on Pearson r was computed using the William's test. In the Appendix, Tables 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31 and 32 provide the detailed results for all language pairs and the multilingual variant, ranking participants by their performance for each of these cases. Note that for the *multilingual* track (Table 21 and 1st column of table 5) we present only the performance of systems that submitted multilingual models (HW-TSC and IST-Unbabel) for that specific track.

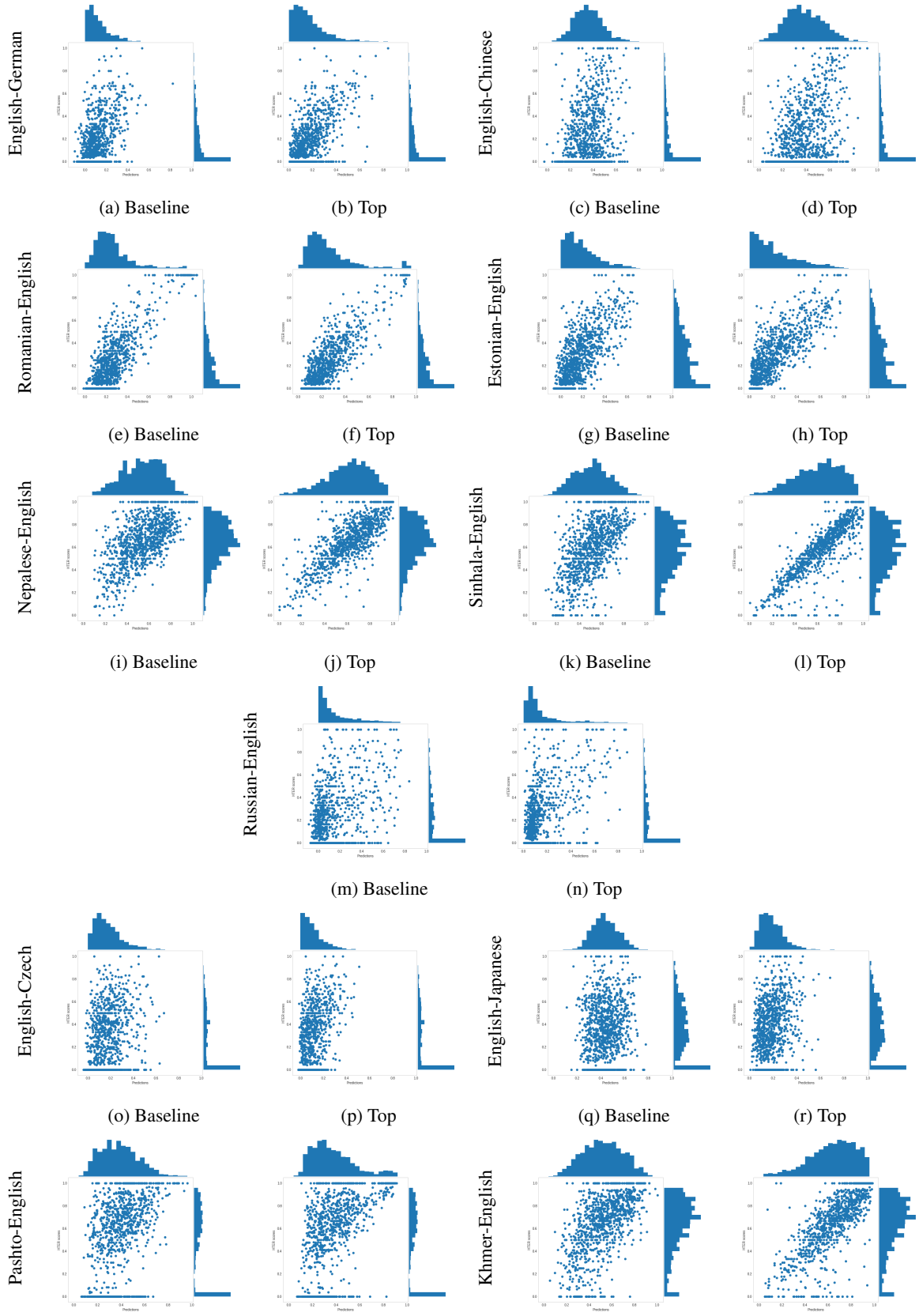
Best performers Both multilingual system submissions outperform the monolingual approaches in the individual language pairs, with HW-TSC ranking first for the majority of the 'supervised' language pairs (En-De, En-Zh, Ne-En, Si-En and Ru-En) as well as the multilingual track and IST-Unbabel leading the majority of the zero-shot language pairs (En-Cs, En-Ja, Ps-En). Apart from the multilingual aspect, the two top systems have more in common: they both use the Predictor-Estimator framework with XLM-Roberta encoders for the predictor and task-specific classifiers for the estimator. Additionally, they both address the sentence- and word-level task using a multi-task approach. HW-TSC enhances their approach using additional pseudo-references as input (generated by another NMT system), while IST-Unbabel system uses additional external data from the WMT Metrics shared task and incorporate adapters in their architecture.

Overall, submitted systems used a variety of approaches to improve performance and address the zero-shot tasks, which revolved around augmenting the training data either by including synthetic data (Bergamot-UTartu, POSTECH) and/or external data (IST-Unbabel) or by using pseudo-references generated by other MT systems (ENSBRT, HW-TSC, JHU-Microsoft). Additionally, ensembling approaches were used to boost performance (HW-TSC, ENSBRT, IST-Unbabel).

Figure 3 shows the scatter plots for the baseline and the best performing system for each language pair. We can see that in most language pairs (perhaps with the exception of En-Zh) the scatter plots for the 'Top' system are much narrower and closer to the identity line, compared to these for the corresponding baseline. More importantly, language pairs with a high proportion of HTER score values close to 0 (many segments without post edits) prove to be more challenging for the submitted models. For example, comparing En-Zh, Ru-En against En-De to Si-En, Ne-En and Et-En, we can see that the latter have narrower, better correlated scatter plots in Figure 3, which is reflected in higher performance in Table 5. This observation seemingly extends to the zero-shot languages, where we observe that performance for the Km-En language pair is consistently higher for all systems compared to the other zero-shot pairs.

Word-level errors For this task, the primary **evaluation** metric is Matthews correlation coefficient (MCC, Matthews, 1975). We also report the F_1-scores for the OK and BAD classes. Similarly to the previous editions, we evaluate separately the source and target side, and this year we also provide a separate evaluation for the target gap tag predictions. We also calculate the performance for combined gaps and words in MT, although it was not considered in the overall ranking process. Systems are primarily ranked by their MCC performance for the word tags on the target side (denoted as 'Words in MT' in the tables). The word-level results for Task 2 are summarised in Tables 6, ordered by the MCC metric, while in the Appendix, Tables 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43 and 44 provide the detailed results for all language pairs and the multilingual variant, ranking participants by their performance for each of these cases. Statistical significance was calculated based on the MCC metric for each language pair using randomization tests with Bonferroni correction (Yeh, 2000a).

Best performers For the multilingual track the picture is similar to the sentence level sub-task, with the HW-TSC system ranking first across all performance indicators, and also leading most of the individual language pairs. Apart from the two multilingual approaches, most of the systems participating in the sentence level sub-task did not submit predictions for the word-level task with the exception of POSTECH which submitted predictions for En-De. However, the JHU-Microsoft which



(a) Baseline (b) Top (c) Baseline (d) Top

(e) Baseline (f) Top (g) Baseline (h) Top

(i) Baseline (j) Top (k) Baseline (l) Top

(m) Baseline (n) Top

(o) Baseline (p) Top (q) Baseline (r) Top

(s) Baseline (t) Top (u) Baseline (v) Top

Figure 3: Scatter plots for the predictions against true HTER scores for the baseline and top-performing system for each language pair. The histograms show the corresponding marginal distributions of predicted and true scores.

Model	Multi	En-De	En-Zh	Ro-En	Et-En	Ne-En	Si-En	Ru-En	En-Cs	En-Ja	Ps-En	Km-En
HW-TSC	0.631	**0.653**	**0.368**	0.862	**0.809**	**0.798**	**0.869**	**0.562**	0.475	**0.262**	**0.534**	**0.753**
IST-Unbabel	0.597	0.617	0.290	**0.879**	**0.811**	0.718	0.710	**0.539**	**0.529**	**0.275**	**0.555**	0.655
BASELINE	0.502	0.529	0.282	0.831	0.714	0.626	0.607	0.448	0.306	0.098	0.503	0.576
ENSBRT	–	0.520	–	0.795	0.666	0.572	0.522	0.376	–	–	–	0.530
Abulice‡	–	0.577	0.312	–	–	–	–	–	–	–	–	–
POSTECH	–	0.546	–	–	–	–	–	–	–	–	–	–
Bergamot-UTartu	–	0.531	–	–	0.562	–	–	–	–	–	–	–
RTM	–	–	0.087	–	–	–	–	–	–	–	–	–

Table 5: Pearson correlation with post-editing effort for the submissions to WMT21 Quality Estimation Task 2 (sentence-level). 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.

Model	Multi	En-De	En-Zh	Ro-En	Et-En	Ne-En	Si-En	Ru-En	En-Cs	En-Ja	Ps-En	Km-En
Words in MT												
HW-TSC	0.530	**0.510**	**0.354**	**0.666**	**0.606**	**0.674**	**0.847**	**0.451**	**0.380**	**0.258**	**0.450**	**0.636**
IST-Unbabel	0.430	0.466	0.310	**0.649**	0.570	0.508	0.528	0.332	**0.376**	0.169	0.370	0.448
BASELINE	0.346	0.370	0.247	0.536	0.461	0.440	0.425	0.256	0.273	0.131	0.313	0.351
JHU-Microsoft	–	**0.523**	0.149	0.634	0.572	0.329	–	0.303	–	–	0.191	–
Abulice‡	–	0.437	0.033	–	–	–	–	–	–	–	–	–
POSTECH	–	0.413	–	–	–	–	–	–	–	–	–	–
GAPs in MT												
HW-TSC	0.337	**0.300**	**0.172**	**0.446**	**0.312**	**0.403**	**0.639**	**0.388**	**0.213**	**0.152**	**0.260**	**0.419**
IST-Unbabel	0.196	0.183	0.068	0.357	0.254	0.268	0.258	0.165	0.125	0.025	0.177	0.259
BASELINE	0.126	0.116	0.065	0.205	0.136	0.215	0.208	0.073	0.039	0.036	0.134	0.175
JHU-Microsoft	–	**0.256**	0.035	0.208	0.218	0.207	–	0.167	–	–	0.118	–
Abulice‡	–	–	–	–	–	–	–	–	–	–	–	–
POSTECH	–	0.110	–	–	–	–	–	–	–	–	–	–
Words in SRC												
HW-TSC	0.432	**0.450**	**0.310**	**0.614**	**0.549**	**0.545**	**0.616**	**0.426**	**0.313**	**0.217**	**0.304**	**0.410**
IST-Unbabel	0.378	0.404	0.286	**0.603**	0.522	**0.445**	0.406	0.351	**0.294**	**0.210**	**0.294**	**0.345**
BASELINE	0.307	0.322	0.241	0.511	0.405	0.390	0.335	0.215	0.224	0.175	0.249	0.279
JHU-Microsoft	–	–	–	–	–	–	–	–	–	–	–	–
Abulice‡	–	0.392	0.011	–	–	–	–	–	–	–	–	–
POSTECH	–	0.320	–	–	–	–	–	–	–	–	–	–
Combined Words and Gaps in MT												
HW-TSC	n/a	**0.496**	**0.359**	**0.656**	**0.584**	**0.749**	**0.868**	**0.456**	0.336	0.180	**0.533**	**0.677**
IST-Unbabel	n/a	**0.468**	**0.369**	**0.640**	**0.582**	0.705	0.690	0.339	**0.400**	0.217	**0.538**	0.631
BASELINE	n/a	0.378	0.320	0.543	0.482	0.672	0.642	0.319	0.339	0.250	0.517	0.587
JHU-Microsoft	n/a	**0.500**	0.240	0.642	**0.572**	0.657	–	0.329	–	–	**0.523**	–
Abulice‡	n/a	0.442	0.118	–	–	–	–	–	–	–	–	–
POSTECH	n/a	0.403	–	–	–	–	–	–	–	–	–	–

Table 6: Matthews correlation coefficient with the OK and BAD classes labels for the submissions to WMT21 Quality Estimation Task 2 (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, 2000a). Baseline systems are highlighted in grey; ‡ indicates Codalab username of participants from whom we have not received further information.

Model	En-De	En-Zh	En-Cs	En-Ja
NICT Kyoto	**0.546**	**0.311**	**0.511**	0.252
HW-TSC	0.490	**0.353**	**0.448**	0.318
LAMA-ICL	0.498	0.305	0.473	0.314
QEMind	0.480	0.278	0.454	0.260
BASELINE	0.397	0.187	0.388	0.214
silence1024‡	0.449	**0.343**	–	0.277
Jason_pogba‡	–	–	–	0.278
serkan‡	–	0.141	–	–

Table 7: Matthews correlation coefficient with the binary critical error labels for the submissions to WMT21 Quality Estimation Task 3. 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 William's test. Baseline systems are highlighted in grey; ‡ indicates Codalab username of participants from whom we have not received further information.

was trained only for the target predictions of the word-level task, obtained the best performance in En-De. JHU-Microsoft also seemed to obtain competitive performance for the Et-En and Ro-En tasks, indicating a strength in languages closer to English.

Overall, the performance of the top two systems is closer for the high – and some of the medium – resource languages, both in the supervised and zero-shot tracks (En-De, En-Zh, En-Cs). Much like the WMT20 shared task, the performance for the target word tags is considerably higher compared to the source tags. This phenomenon is observed across language pairs with the exception of Ru-En where predictions for source and target words are close for all systems. This year we can also observe the performance on target gaps separately, which is consistently lower, even when compared to the source tags, across all language pairs and submitted systems.

It is important to note that when focusing on the combined target performance, i.e., the combination of word and gap quality predictions for the MT, the order and performance differences between the top scoring teams can vary compared to the MT word prediction ones. Overall, there are fewer language pairs where we have a clear winner (Ne-En, Si-En and Km-En for HW-TSC and Cs-En for IST-Unbabel) while for the rest there is no statistically significant difference between the top pairs. Still, HW-TSC is consistently among the top systems, with the exception of Cs-En.

5.3 Task 3

Table 7 summarises the results for all language pairs, ranked by their performance in terms of Matthews correlation coefficient (MCC, Matthews, 1975). In the Appendix, Tables 45, 46, 47 and 48 provide the detailed results for all language pairs, ranking participants by their performance for each of these cases. Statistical significance is calculated using the William's test.

This task attracted fewer participants than the others, most likely because it is new. All described systems perform better than the baseline for all language pairs. Across languages, the order of MCC scores roughly corresponds to the skewness of data distribution obtained for languages: For En-De, which achieved the highest MCC score, the NOT (no error class) accounted for 72% of the training instances, while for En-Ja, with the lowest MCC score, the NOT class accounted for 91% of the training instances.

Best performers NICT Kyoto ranked among the top systems for all language pairs. However, only for En-De it did significantly outperform all other systems. For the rest of the language pairs there could not be a clear winner based on statistical significance testing; HW-TSC was in the top-ranked systems for En-Zh and En-Cs, while for En-Ja no system managed to significantly outperform the others, but they all performed significantly better than the baseline.

In terms of the approaches applied by the best performing systems, they all use the baseline architecture as starting point, but HW-TSC also uses machine translations for the source sentences by top online systems. These are concatenated to the provided source and translation pair. NICT Kyoto also added synthetic data for multiple language pairs with multitask learning and model ensembling. LAMA-ICL used additional features to detect the presence/deviation of toxicity, sentiment and named entities, also followed by ensenbling of models with different individual features.

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, with the best performing submissions showing moderate to strong correlation for sentence-level DA and HTER prediction tasks. One reason for high correlation levels is likely

to be that top performing systems are based on pre-trained representations. Even for zero-shot languages (see below), relatively high (above 0.5) correlation was achieved for most languages for sentence-level tasks. The same applies for the word-level tasks, where the performance was behind that of supervised prediction, but still high for Km-En and Ps-En.

A comparison to previous years submissions is possible for Task 1 on the non-zero shot languages. The training data is the same and the test sets (test20 and test21) were created at the same time, with data sampled and annotated in the same way. Comparing Pearson correlation scores from the 2020 official results to this year's official results, as we can see from Table 8, for the languages which had already achieved very strong correlations, it remained the same (Ro-En, Et-En, Ru-En) or improved (Ne-En), whereas for the languages with average correlation, it mostly improved substantially (En-De, En-Zh). The exception was Si-En, where the correlation was lower in 2021, which needs further investigation. Overall, we believe the numbers show steady progress over previous models, even though the core of most of the winning systems is the same for the 2020 and 2021 editions.

Model size. When interpreting the results for all tasks, it should be noted that most of the participants use extremely resource-heavy systems, ensembles of multiple models with more than 500M parameters, which could make them difficult to use in practice. In this year's edition of the Shared Task on QE we asked the participants to provide information on the size of their models. Figures 4 and 5 illustrate the performance-efficiency trade-off for the submitted systems. On the x-axis we plot the Pearson correlation with sentence-level DA judgements (Task 1), while the y-axis shows the number of model parameters, as reported by the participants. Pareto-optimal submissions are marked in blue. These plots give us a different view of the performance of the submitted systems. Thus, for the higher quality models, the best results are achieved by QEMind and HW-TSC, whereas Bergamot, Bergamot-UTartu and BASELINE are optimal in terms of model size.

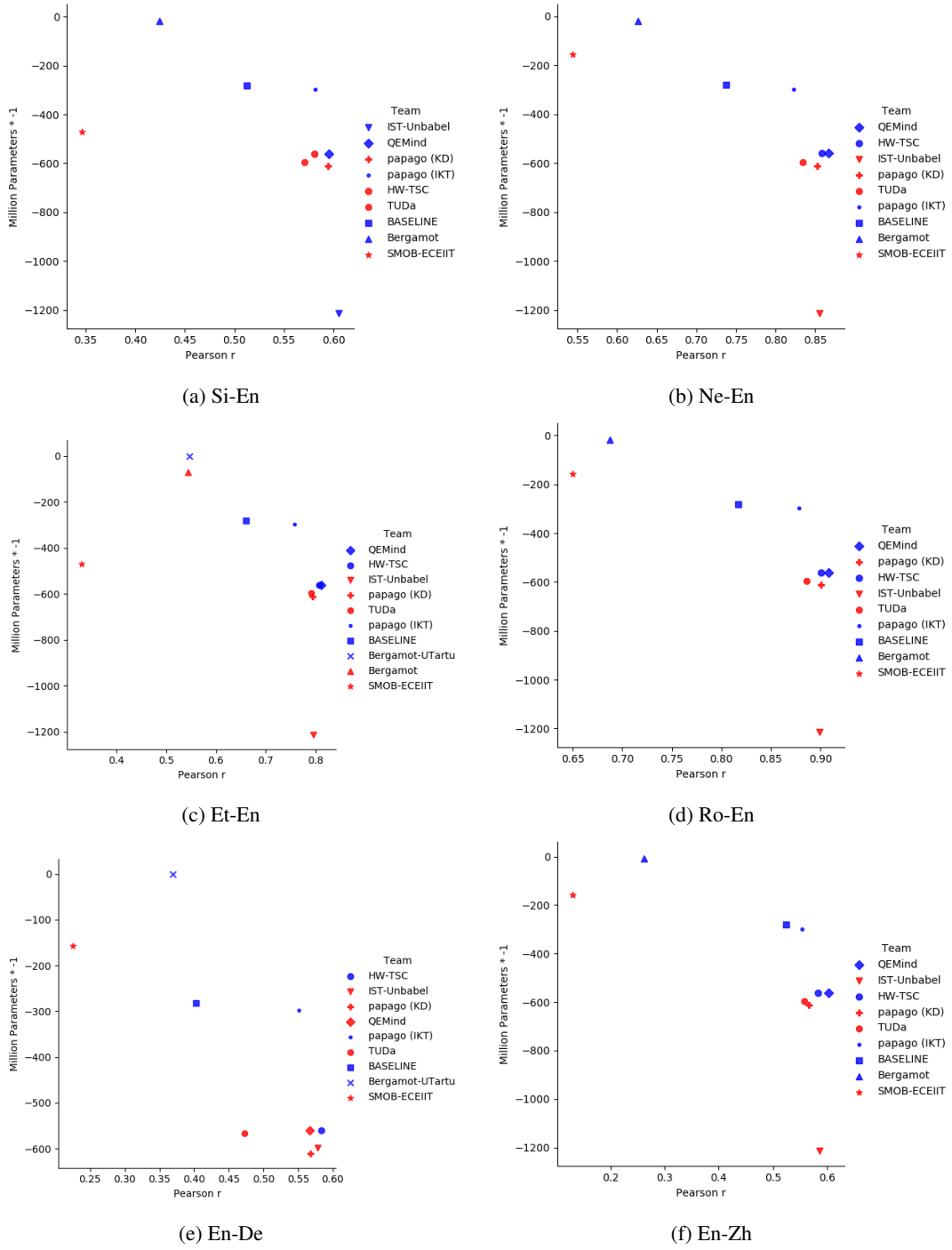
Extending publicly available benchmarks. This year counted with substantial new data. One the one hand, we extended the MLQE-PE dataset with more DA test sets (for all seven previous language pairs and four new zero-shot language pairs), as well as post-editing training and test sets for five additional language pairs (which only had DA scores before), as well as the four zero-shot language pairs. On the other hand, we created sizeable data for the new Task 3, a unique set focusing on critical errors, based on three annotations by professional translators. We hope that others will also contribute by adding new languages to this dataset in the future.

Zero-shot prediction. For the first time, we introduced language pairs for which no training data was available. This challenge was addressed mainly in two ways: synthetic data creation with using parallel data for the relevant languages, and use of indicators coming from the NMT system for unsupervised prediction. Overall, the performance for these languages was surprisingly good (except for En-Ja, potentially for data segmentation issues), comparable to non-zero-shot languages in the dataset. We attribute this high performance mostly to the use of fine-tuning on synthetic data for the relevant languages. In future editions, we may consider blinder zero-shot settings where participants will not be informed of the actual languages the models will require to predict the quality for, to encourage the development of truly multilingual or language-agnostic models.

Critical error detection. We posit that the detection of critical errors is a very important problem for two main reasons: (i) high-quality NMT models may produce fluent translations that may appear very good, but contain localised errors which are not always obvious and may go unnoticed, even by human translators post-editing the translation; and (ii) certain types of content are particularly challenging for MT models, such as social media data posts containing named entities, and could lead to critical errors especially if translations are to be used without human editing. While in the past we have provided word-level QE tasks where errors were annotated not only with error categories, but also error severity (e.g. MQM data in last year's WMT QE Task 3), this was the first attempt to predict specifically (and only) critical errors. This seems a much harder problem, as we expect the QE model to be able not only to find errors, but to distinguish minor (and even major) errors from critical errors. That was the reasoning for our "simplification" of the task this year, i.e. for making

Shared task	En-De	En-Zh	Ro-En	Et-En	Ne-En	Si-En	Ru-En
WMT 2021	**0.58**	**0.60**	0.91	0.81	**0.87**	0.61	0.81
WMT 2020	0.55	0.54	0.91	0.82	0.82	**0.68**	0.81

Table 8: Pearson correlation with direct assessments - comparison between top submission in 2020 and 2021. While the test set is different, it was taken from the same distribution. The training set is the same.

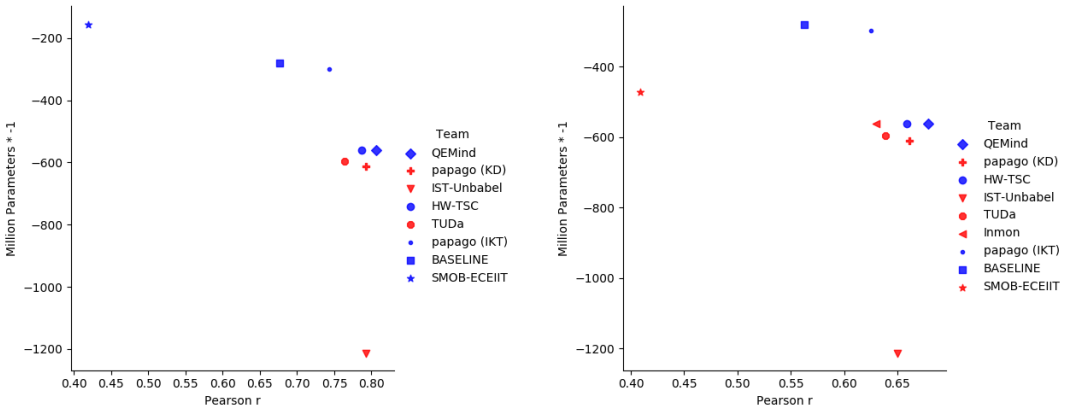


(a) Si-En (b) Ne-En

(c) Et-En (d) Ro-En

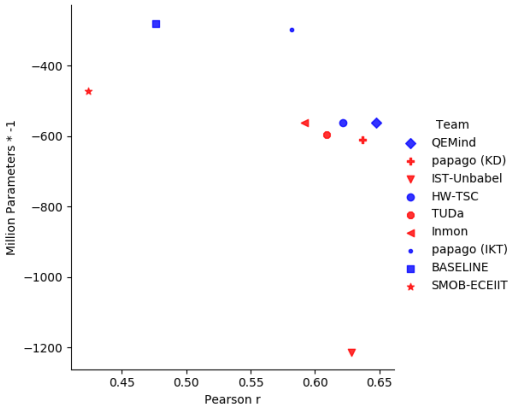
(e) En-De (f) En-Zh

Figure 4: Performance of the submitted systems on Task 1 for Si-En, Ne-En, Et-En, Ro-En, En-De and En-Zh. The x-axis shows Pearson correlation with human judgements and the y-axis corresponds to the number of model parameters multiplied by -1. Pareto optimal submissions are marked in blue, while the rest are shown in red.

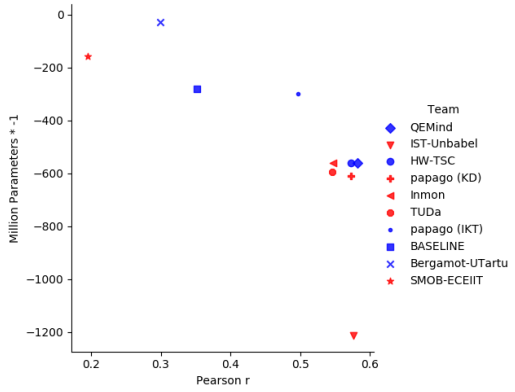


(a) Ru-En

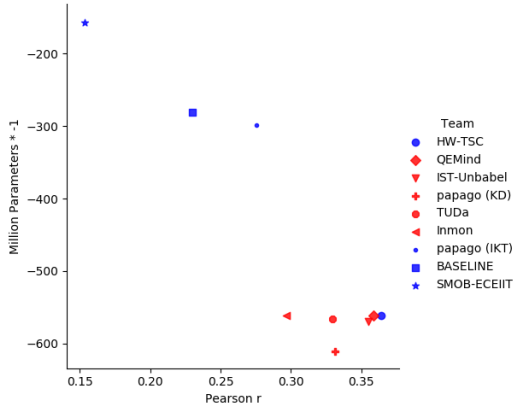
(b) Km-En



(c) Ps-En



(d) En-Cs



(e) En-Ja

Figure 5: Performance of the submitted systems on Task 1 for Ru-En, Km-En, Ps-En, En-Cs and En-Ja. The x-axis shows Pearson correlation with human judgements and the y-axis corresponds to the number of model parameters multiplied by -1. Pareto optimal submissions are marked in blue, while the rest are shown in red.

it a sentence-level binary classification problem. This might be enough for filtering purposes, i.e. to avoid offering/using automatic translations that may contain critical errors. If the goal is to support human translators in the task of post-editing, more fine-grained prediction may be needed.

The overall results for this task in terms of MCC are promising, especially for En-Cs and En-De. Considering the detailed results for this task in Tables 45, 46, 47 and 48, we see that despite the skewed distribution between the two classes, the models achieve a high F1 score at detecting errors, around 0.9 or higher for all language pairs.

7 Conclusions

This year's edition of the QE Shared Task introduced a number of new elements: new data covering five more language pairs with post-edits for sentence and word-level prediction, new test sets for all tasks, including four new zero-shot language pairs, and a new task focusing on critical error detection. Our analysis also paid close attention to model size, an important aspect for deploying QE systems in realistic applications, such as real-time inference and devices with limited resources. The tasks attracted a steady number of participating teams and systems and we believe the overall results are a great reflection of the SotA in QE. Continuing from the effort we set forward last year, this edition the tasks in this edition, with its zero-shot variant, cover a broad range of challenges in QE, such as improving performance for languages with skewed distributions, addressing low (or zero) resource languages, predicting source words that lead to errors, multilingual models, etc.

We are making the gold labels and all submissions to all tasks available for those interested in further analysing the results, investigating approaches for prediction ensembling, among others.

Acknowledgments

Marina Fomicheva and Lucia Specia were supported by funding from the Bergamot project (EU H2020 Grant No. 825303). André Martins and Chrysoula Zerva were funded by the P2020 programs Unbabel4EU (contract 042671) and 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. We would like to thank Erick Fonseca for answering questions about data preprocessing, and Marina Sánchez-Torrón and Camila Pohlman for monitoring the post-editing process for English-German, English-Chinese, and Russian-English for Task 2. We thank IQT Labs for providing the Russian-English dataset for Task 1. We thank Genze Jiang for helping with the baseline model for Task 3.

References

Ergun Biçici. 2021. Rtm superlearner results at quality estimation task. In *Proceedings of the Sixth Conference on Machine Translation Shared Tasks Papers*, Online. Association for Computational Linguistics.

Yimeng Chen, Chang Su, Yingtao Zhang, Yuxia Wang, Xiang Geng, Hao Yang, Shimin Tao, Guo Jiaxin, Wang Minghan, Min Zhang, Yujia Liu, and Shujian Huang. 2021. Hw-tsc's participation at wmt 2021 quality estimation shared tasks. In *Proceedings of the Sixth Conference on Machine Translation Shared Tasks Papers*, Online. Association for Computational Linguistics.

Shaika Chowdhury, Naouel Baili, and Brian Vannah. 2021. Ensemble fine-tuned mbert for wmt21 translation quality estimation shared task. In *Proceedings of the Sixth Conference on Machine Translation Shared Tasks Papers*, Online. Association for Computational Linguistics.

Marco Cuturi. 2013. Sinkhorn distances: Lightspeed computation of optimal transportation distances. *arXiv preprint arXiv:1306.0895*.

Shuoyang Ding, Marcin Junczys-Dowmunt, Matt Post, Christian Federmann, and Philipp Koehn. 2021. The jhu-microsoft submission for wmt21 quality estimation shared task. In *Proceedings of the Sixth Conference on Machine Translation Shared Tasks Papers*, Online. Association for Computational Linguistics.

Herbert Edelsbrunner. 2013. Persistent homology: theory and practice.

Marina Fomicheva, Shuo Sun, Erick Fonseca, Frédéric Blain, Vishrav Chaudhary, Francisco Guzmán, Nina Lopatina, Lucia Specia, and André F. T. Martins. 2020. MLQE-PE: A multilingual quality estimation and post-editing dataset. *arXiv preprint arXiv:2010.04480*.

Amit Gajbhiye, Marina Fomicheva, Fernando Alva-Manchego, Frédéric Blain, Abiola Obamuyide, Nikolaos Aletras, and Lucia Specia. 2021. Knowledge distillation for quality estimation.

Gregor Geigle, Jonas Elias Stadtmüller, Wei Zhao, Jonas Pfeiffer, and Steffen Eger. 2021. Tuda at wmt21: Sentence-level direct assessment with adapters. In *Proceedings of the Sixth Conference on Machine Translation Shared Tasks Papers*, Online. Association for Computational Linguistics.

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.

Dam Heo, WonKee Lee, Baikjin Jung, and Jong-Hyeok Lee. 2021. Quality estimation using dual encoders with transfer learning. In *Proceedings of the Sixth Conference on Machine Translation Shared Tasks Papers*, Online. Association for Computational Linguistics.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339.

Genze Jiang, Zhenhao Li, and Lucia Specia. 2021. Icl's submission to the wmt21 critical error detection shared task. In *Proceedings of the Sixth Conference on Machine Translation Shared Tasks Papers*, Online. Association for Computational Linguistics.

Fábio Kepler, Jonay Trénous, Marcos Treviso, Miguel Vera, and André F. T. Martins. 2019. OpenKiwi: An open source framework for quality estimation. In *Proceedings of ACL 2019 System Demonstrations*.

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.

Seunghyun Shaun Lim, Hantae Kim, and Hyunjoong Kim. 2021. Papago submission for the wmt21 quality estimation shared task. In *Proceedings of the Sixth Conference on Machine Translation Shared Tasks Papers*, Online. Association for Computational Linguistics.

Brian W Matthews. 1975. Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochimica et Biophysica Acta (BBA)-Protein Structure*, 405(2):442–451.

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.

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

Raphaël Rubino, Atsushi Fujita, and Benjamin Marie. 2021. Nict kyoto submission for the wmt'21 quality estimation task: Multimetric multilingual pretraining for critical error detection. In *Proceedings of the Sixth Conference on Machine Translation Shared Tasks Papers*, Online. Association for Computational Linguistics.

Thibault Sellam, Dipanjan Das, and Ankur P Parikh. 2020. Bleurt: Learning robust metrics for text generation. In *Proceedings of ACL*.

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

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.

Ivan Vulić, Sebastian Ruder, and Anders Søgaard. 2020. Are all good word vector spaces isomorphic? *arXiv preprint arXiv:2004.04070*.

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 Shared Tasks Papers*, Online. Association for Computational Linguistics.

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

Lisa Yankovskaya and Mark Fishel. 2021. Bergamot-utartu participation in the wmt'2021 quality estimation shared task. In *Proceedings of the Sixth Conference on Machine Translation Shared Tasks Papers*, Online. Association for Computational Linguistics.

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

Alexander Yeh. 2000b. More accurate tests for the statistical significance of result differences. *arXiv preprint cs/0008005*.

Chrysoula Zerva, Daan van Stigt, Ricardo Rei, Ana C Farinha, Pedro G. Ramos, José G. C. de Souza, Taisiya Glushkova, Miguel Vera, Fabio Kepler, and

André F. T. Martins. 2021. Ist-unbabel 2021 submission for the quality estimation shared task. In *Proceedings of the Sixth Conference on Machine Translation Shared Tasks Papers*, Online. Association for Computational Linguistics.

A Official Results of the WMT21 Quality Estimation Task 1

Tables 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19 and 20 show the results for all language pairs and the multilingual variant, ranking participating systems best to worst using Pearson's r correlation as primary key for each of these cases.

Model	Rank	**Pearson** r	MAE	RMSE	Disk footprint (B)	# Model params
QEMind	3	0.675	0.627	0.486	2,244,030,744	560,981,507
HW-TSC	2.4	0.665	0.627	0.482	2,243,941,083	560,941,057
IST-Unbabel	5	0.665	0.642	0.495	4,872,322,439	1,214,683,792
papago (IKT)	5.2	0.658	0.645	0.496	2,503,797,760	611,278,859
TUDa	6.2	0.631	0.688	0.526	2,382,759,964	595,689,991
Inmon‡	5.2	0.623	0.687	0.526	2,243,941,083	560,941,057
papago (KD)	4.2	0.613	0.687	0.524	1,249,902,592	297,974,795
BASELINE	5.2	0.541	0.729	0.562	1,142,413,043	281,291,535
SMOB-ECEIIT	6.6	0.348	1.057	0.821	1,886,937,088	471,716,864

Table 9: Official results of the WMT21 Quality Estimation Task 1 for the **Multilingual** variant. 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). ‡ indicates Codalab usernames of participants from whom we have not received further information.

Model	Rank	**Pearson** r	MAE	RMSE	Disk footprint (B)	# Model params
• HW-TSC	2.6	0.584	0.544	0.390	2,243,941,083	560,941,057
• IST-Unbabel	4.4	0.579	0.567	0.393	2,409,244,995	598,943,476
• papago (IKT)	5.8	0.568	0.580	0.430	2,445,115,000	611,278,859
QEMind	4.8	0.567	0.579	0.432	2,244,030,744	560,981,507
papago (KD)	4.2	0.551	0.587	0.426	1,249,902,592	297,974,795
TUDa	6.6	0.473	0.626	0.440	2,264,844,300	566,211,075
BASELINE	5.2	0.403	0.629	0.433	1,142,413,043	281,291,535
Bergamot-UTartu	5.2	0.369	0.854	0.605	6,985,478	421,537
SMOB-ECEIIT	6.2	0.226	1.070	0.834	626,401,280	156,589,824
RTM	n/a	0.143	1.150	0.538	61,203,283,968	380

Table 10: Official results of the WMT21 Quality Estimation Task 1 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. "Rank" indicates the averaged ranking of participants with regards to all metrics (including memory print and number of parameters).

Model	Rank	**Pearson** r	MAE	RMSE	Disk footprint (B)	# Model params
• QEMind	3	0.603	0.580	0.450	2,244,030,744	560,981,507
IST-Unbabel	5.6	0.586	0.631	0.499	4,872,322,439	1,214,683,792
HW-TSC	3.6	0.583	0.627	0.487	2,243,941,083	560,941,057
papago (IKT)	5	0.567	0.623	0.490	2,503,797,760	611,278,859
TUDa	6.6	0.558	0.687	0.541	2,382,759,964	595,689,991
papago (KD)	4.8	0.553	0.643	0.500	1,249,902,592	297,974,795
BASELINE	5	0.525	0.683	0.534	1,142,413,043	281,291,535
Bergamot	5.4	0.262	1.088	0.914	28,949,742	6,941,751
RTM	n/a	0.248	1.924	1.772	61,203,283,968	380
SMOB-ECEIIT	6	0.131	1.149	0.838	626,401,280	156,589,824

Table 11: Official results of the WMT21 Quality Estimation Task 1 for the **English-Chinese** 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. "Rank" indicates the averaged ranking of participants with regards to all metrics (including memory print and number of parameters).

Model	Rank	**Pearson** r	MAE	RMSE	Disk footprint (B)	# Model params
• QEMind	4	0.908	0.393	0.316	2,244,030,744	560,981,507
papago (IKT)	4.6	0.901	0.393	0.288	2,503,797,760	611,278,859
HW-TSC	3	0.901	0.384	0.286	2,243,941,083	560,941,057
IST-Unbabel	5.8	0.899	0.393	0.289	4,872,322,439	1,214,683,792
TUDa	6.2	0.886	0.453	0.335	2,382,759,964	595,689,991
papago (KD)	4.6	0.879	0.427	0.316	1,249,902,592	297,974,795
BASELINE	5.4	0.818	0.556	0.408	1,142,413,043	281,291,535
Bergamot	5.6	0.687	1.024	0.748	70,044,344	16,772,151
SMOB-ECEIIT	5.8	0.650	0.794	0.628	626,401,280	156,589,824
RTM	n/a	0.287	3.749	3.607	61,203,283,968	380

Table 12: Official results of the WMT21 Quality Estimation Task 1 for the **Romanian-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. "Rank" indicates the averaged ranking of participants with regards to all metrics (including memory print and number of parameters).

Model	Rank	**Pearson** r	MAE	RMSE	Disk footprint (B)	# Model params
• QEMind	3.4	0.812	0.488	0.393	2,244,030,744	560,981,507
• HW-TSC	4.4	0.808	0.520	0.409	2,243,941,083	560,941,057
IST-Unbabel	5.8	0.796	0.519	0.404	4,872,322,439	1,214,683,792
papago (KD)	5.2	0.794	0.510	0.397	2,503,797,760	611,278,859
TUDa	6.4	0.792	0.563	0.424	2,382,759,964	595,689,991
papago (IKT)	5	0.759	0.550	0.434	1,249,902,592	297,974,795
BASELINE	5.4	0.660	0.700	0.543	1,142,413,043	281,291,535
Bergamot-UTartu	6	0.547	1.840	1.701	1,705,478	421,537
Bergamot	5.8	0.544	0.966	0.761	284,339,184	70,969,501
SMOB-ECEIIT	7.6	0.329	1.072	0.862	1,886,937,088	471,716,864
RTM	n/a	0.099	2.520	2.346	61,203,283,968	380

Table 13: Official results of the WMT21 Quality Estimation Task 1 for the **Estonian-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. "Rank" indicates the averaged ranking of participants with regards to all metrics (including memory print and number of parameters).

Model	Rank	**Pearson** r	MAE	RMSE	Disk footprint (B)	# Model params
• QEMind	4.8	0.867	0.570	0.426	2,244,030,744	560,981,507
HW-TSC	2.8	0.858	0.504	0.384	2,243,941,083	560,941,057
IST-Unbabel	5.2	0.856	0.515	0.401	4,872,322,439	1,214,683,792
papago (IKT)	5	0.853	0.522	0.399	2,503,797,760	611,278,859
TUDa	5.4	0.834	0.540	0.426	2,382,759,964	595,689,991
papago (KD)	5	0.823	0.562	0.441	1,249,902,592	297,974,795
BASELINE	5.4	0.738	0.657	0.524	1,142,413,043	281,291,535
Bergamot	5.6	0.626	0.977	0.818	83,907,600	19,220,401
SMOB-ECEIIT	5.8	0.544	0.931	0.717	626,401,280	156,589,824
RTM	n/a	0.127	2.286	2.017	61,203,283,968	380

Table 14: Official results of the WMT21 Quality Estimation Task 1 for the **Nepalese-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. "Rank" indicates the averaged ranking of participants with regards to all metrics (including memory print and number of parameters).

Model	Rank	**Pearson** r	MAE	RMSE	Disk footprint (B)	# Model params
• IST-Unbabel	4.2	0.605	0.742	0.583	4,872,322,439	1,214,683,792
• QEMind	5	0.596	0.783	0.609	2,244,030,744	560,981,507
• papago (IKT)	4.6	0.595	0.745	0.585	2,503,797,760	611,278,859
papago (KD)	3.2	0.582	0.768	0.597	1,249,902,592	297,974,795
HW-TSC	4.8	0.581	0.776	0.602	2,243,941,083	560,941,057
TUDa	6	0.571	0.774	0.609	2,382,759,964	595,689,991
BASELINE	5	0.513	0.797	0.626	1,142,413,043	281,291,535
Bergamot	5.2	0.425	0.920	0.773	74,490,910	17,079,701
SMOB-ECEIIT	7	0.347	1.115	0.864	1,886,937,088	471,716,864
RTM	n/a	0.061	2.822	2.485	61,203,283,968	380

Table 15: Official results of the WMT21 Quality Estimation Task 1 for the **Sinhala-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. "Rank" column indicates the averaged ranking of participants with regards to all metrics (including memory print and number of parameters).

Model	Rank	**Pearson** r	MAE	RMSE	Disk footprint (B)	# Model params
• QEMind	2.6	0.806	0.534	0.388	2,244,030,744	560,981,507
papago (IKT)	4.2	0.793	0.572	0.392	2,503,797,760	611,278,859
IST-Unbabel	5.4	0.792	0.583	0.412	4,872,322,439	1,214,683,792
HW-TSC	3.4	0.787	0.554	0.397	2,243,941,083	560,941,057
TUDa	5.8	0.764	0.629	0.437	2,382,759,964	595,689,991
papago (KD)	4.4	0.744	0.615	0.421	1,249,902,592	297,974,795
BASELINE	5	0.677	0.702	0.492	1,142,413,043	281,291,535
SMOB-ECEIIT	5.2	0.420	1.026	0.795	626,401,280	156,589,824
RTM	n/a	0.356	1.126	0.841	61,203,283,968	380

Table 16: Official results of the WMT21 Quality Estimation Task 1 for the **Russian-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. "Rank" indicates the averaged ranking of participants with regards to all metrics (including memory print and number of parameters).

Model	Rank	**Pearson** r	MAE	RMSE	Disk footprint (B)	# Model params
• QEMind	3.4	0.582	0.746	0.599	2,244,030,744	560,981,507
• IST-Unbabel	5	0.577	0.751	0.583	4,872,322,439	1,214,683,792
• HW-TSC	3.8	0.573	0.747	0.602	2,243,941,083	560,941,057
• papago (IKT)	5	0.572	0.748	0.585	2,503,797,760	611,278,859
Inmon ‡	5.8	0.547	0.809	0.624	2,243,941,083	560,941,057
TUDa	6.2	0.545	0.808	0.619	2,382,759,964	595,689,991
papago (KD)	5.2	0.497	0.765	0.621	1,249,902,592	297,974,795
BASELINE	6	0.352	0.845	0.686	1,142,413,043	281,291,535
Bergamot-UTartu	6.2	0.300	1.420	1.166	111,300,550	27,815,809
SMOB-ECEIIT	6.4	0.195	1.199	0.967	626,401,280	156,589,824
RTM	n/a	-0.104	2.159	1.902	61,203,283,968	380

Table 17: Official results of the WMT21 Quality Estimation Task 1 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. "Rank" indicates the averaged ranking of participants with regards to all metrics (including memory print and number of parameters). ‡ indicates Codalab usernames of participants from whom we have not received further information.

Model	Rank	**Pearson** r	MAE	RMSE	Disk footprint (B)	# Model params
• HW-TSC	2.2	0.364	0.755	0.556	2,243,941,083	560,941,057
• QEMind	3.2	0.359	0.757	0.560	2,244,030,744	560,981,507
• IST-Unbabel	4.6	0.355	0.764	0.566	2,277,509,716	569,330,715
papago (IKT)	6	0.332	0.853	0.648	2,503,797,760	611,278,859
TUDa	6.6	0.330	0.917	0.705	2,264,844,300	566,211,075
Inmon ‡	5.6	0.297	0.882	0.665	2,243,941,083	560,941,057
papago (KD)	5	0.276	0.865	0.649	1,249,902,592	297,974,795
BASELINE	4	0.230	0.816	0.617	1,142,413,043	281,291,535
SMOB-ECEIIT	5.8	0.153	1.174	0.870	626,401,280	156,589,824
RTM	n/a	-0.082	2.694	2.576	61,203,283,968	380

Table 18: Official results of the WMT21 Quality Estimation Task 1 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. "Rank" indicates the averaged ranking of participants with regards to all metrics (including memory print and number of parameters). ‡ indicates Codalab usernames of participants from whom we have not received further information.

Model	Rank	**Pearson** r	MAE	RMSE	Disk footprint (B)	# Model params
• QEMind	2.8	0.647	0.736	0.605	2,244,030,744	560,981,507
• papago (IKT)	4	0.637	0.738	0.605	2,503,797,760	611,278,859
IST-Unbabel	5.8	0.628	0.780	0.658	4,872,322,439	1,214,683,792
HW-TSC	3.4	0.622	0.737	0.616	2,243,941,083	560,941,057
TUDa	6.2	0.609	0.824	0.674	2,382,759,964	595,689,991
Inmon ‡	5.2	0.592	0.795	0.665	2,243,941,083	560,941,057
papago (KD)	3.8	0.582	0.771	0.632	1,249,902,592	297,974,795
BASELINE	5.2	0.476	0.852	0.711	1,142,413,043	281,291,535
SMOB-ECEIIT	6.6	0.424	1.044	0.832	1,886,937,088	471,716,864

Table 19: Official results of the WMT21 Quality Estimation Task 1 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. "Rank" indicates the averaged ranking of participants with regards to all metrics (including memory print and number of parameters). ‡ indicates Codalab usernames of participants from whom we have not received further information.

Model	Rank	**Pearson** r	MAE	RMSE	Disk footprint (B)	# Model params
• QEMind	2.8	0.679	0.729	0.564	2,244,030,744	560,981,507
papago (IKT)	6	0.662	0.815	0.641	2,503,797,760	611,278,859
HW-TSC	3.6	0.659	0.744	0.578	2,243,941,083	560,941,057
IST-Unbabel	4.6	0.650	0.721	0.568	4,872,322,439	1,214,683,792
TUDa	4.8	0.639	0.740	0.585	2,382,759,964	595,689,991
Inmon ‡	4.8	0.630	0.765	0.599	2,243,941,083	560,941,057
papago (KD)	5.4	0.625	0.879	0.693	1,249,902,592	297,974,795
BASELINE	4.4	0.562	0.788	0.614	1,142,413,043	281,291,535
SMOB-ECEIIT	6.6	0.409	1.057	0.830	1,886,937,088	471,716,864

Table 20: Official results of the WMT21 Quality Estimation Task 1 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. "Rank" indicates the averaged ranking of participants with regards to all metrics (including memory print and number of parameters). ‡ indicates Codalab usernames of participants from whom we have not received further information.

B Official Results of the WMT21 Quality Estimation Task 2 (Sentence-level)

Tables 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31 and 32 show the results for all language pairs and the multilingual variant, ranking participating systems best to worst using Pearson's r correlation as primary key for each of these cases.

Model	Rank	**Pearson** r	MAE	RMSE	Disk footprint (B)	# Model params
HW-TSC	1.4	0.631	0.202	0.153	2,243,954,093	560,944,640
IST-Unbabel	2.4	0.597	0.219	0.171	2,294,887,576	569,368,609
BASELINE	2.2	0.502	0.235	0.188	1,142,441,796	281,297,685

Table 21: Official results of the WMT21 Quality Estimation Task 2 for the **Multilingual** variant. 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).

Model	Rank	**Pearson** r	MAE	RMSE	Disk footprint (B)	# Model params
• HW-TSC	3	0.653	0.151	0.108	2,243,954,093	560,944,640
IST-Unbabel	4.6	0.617	0.172	0.116	2,294,887,576	569,368,609
Abulice ‡	4.2	0.577	0.174	0.115	2,243,439,613	560,814,661
POSTECH	4.6	0.546	0.172	0.139	1,561,188,430	390,210,052
Bergamot-UTartu	3.2	0.531	0.171	0.135	55,632,317	48
BASELINE	4.4	0.529	0.183	0.129	1,142,441,796	281,297,685
ENSBRT	4	0.520	0.171	0.129	1,363,652,116	502,000,000

Table 22: Official results of the WMT21 Quality Estimation Task 2 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. "Rank" indicates the averaged ranking of participants with regards to all metrics (including memory print and number of parameters). ‡ indicates Codalab usernames of participants from whom we have not received further information.

Model	Rank	**Pearson** r	MAE	RMSE	Disk footprint (B)	# Model params
• HW-TSC	2.6	0.368	0.297	0.248	2,243,954,093	560,944,640
Abulice ‡	2.4	0.312	0.340	0.280	100,000	9,501,148
IST-Unbabel	2.6	0.290	0.266	0.220	2,294,887,576	569,368,609
BASELINE	2.4	0.282	0.287	0.246	1,142,441,796	281,297,685
RTM	n/a	0.087	0.668	0.621	61,203,283,968	380

Table 23: Official results of the WMT21 Quality Estimation Task 2 for the **English-Chinese** 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. "Rank" indicates the averaged ranking of participants with regards to all metrics (including memory print and number of parameters). ‡ indicates Codalab usernames of participants from whom we have not received further information.

Model	Rank	**Pearson** r	MAE	RMSE	Disk footprint (B)	# Model params
• IST-Unbabel	2.2	0.879	0.122	0.098	2,294,887,576	569,368,609
HW-TSC	2.6	0.862	0.144	0.111	2,243,954,093	560,944,640
BASELINE	2	0.831	0.142	0.115	1,142,441,796	281,297,685
ENSBRT	3.2	0.795	0.171	0.141	1,363,652,116	502,000,000

Table 24: Official results of the WMT21 Quality Estimation Task 2 for the **Romanian-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. "Rank" indicates the averaged ranking of participants with regards to all metrics (including memory print and number of parameters).

Model	Rank	**Pearson** r	MAE	RMSE	Disk footprint (B)	# Model params
• IST-Unbabel	2.8	0.811	0.153	0.112	2,294,887,576	569,368,609
• HW-TSC	2.6	0.809	0.154	0.110	2,243,954,093	560,944,640
BASELINE	3.4	0.714	0.195	0.149	1,142,441,796	281,297,685
ENSBRT	3.2	0.666	0.171	0.132	1,363,652,116	502,000,000
Bergamot-UTartu	3	0.562	0.191	0.149	65,310,657	48

Table 25: Official results of the WMT21 Quality Estimation Task 2 for the **Estonian-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. "Rank" indicates the averaged ranking of participants with regards to all metrics (including memory print and number of parameters).

Model	Rank	**Pearson** r	MAE	RMSE	Disk footprint (B)	# Model params
• HW-TSC	1.8	0.798	0.136	0.099	2,243,954,093	560,944,640
IST-Unbabel	2.8	0.718	0.161	0.126	2,294,887,576	569,368,609
BASELINE	2.6	0.626	0.205	0.160	1,142,441,796	281,297,685
ENSBRT	2.8	0.572	0.176	0.139	1,363,652,116	502,000,000

Table 26: Official results of the WMT21 Quality Estimation Task 2 for the **Nepalese-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. "Rank" indicates the averaged ranking of participants with regards to all metrics (including memory print and number of parameters).

Model	Rank	**Pearson** r	MAE	RMSE	Disk footprint (B)	# Model params
• HW-TSC	1.8	0.869	0.126	0.075	2,243,954,093	560,944,640
IST-Unbabel	2.8	0.710	0.178	0.136	2,294,887,576	569,368,609
BASELINE	2.2	0.607	0.204	0.159	1,142,441,796	281,297,685
ENSBRT	3.2	0.522	0.206	0.162	1,363,652,116	502,000,000

Table 27: Official results of the WMT21 Quality Estimation Task 2 for the **Sinhala-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. "Rank" indicates the averaged ranking of participants with regards to all metrics (including memory print and number of parameters).

Model	Rank	**Pearson** r	MAE	RMSE	Disk footprint (B)	# Model params
• HW-TSC	2	0.562	0.225	0.160	2,243,954,093	560,944,640
• IST-Unbabel	2.6	0.539	0.224	0.165	2,294,845,131	569,360,411
BASELINE	2.4	0.448	0.255	0.188	1,142,441,796	281,297,685
ENSBRT	3	0.376	0.251	0.189	1,363,652,116	502,000,000

Table 28: Official results of the WMT21 Quality Estimation Task 2 for the **Russian-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. "Rank" indicates the averaged ranking of participants with regards to all metrics (including memory print and number of parameters).

Model	Rank	**Pearson** r	MAE	RMSE	Disk footprint (B)	# Model params
• IST-Unbabel	2.4	0.529	0.271	0.200	2,294,887,576	569,368,609
HW-TSC	1.6	0.475	0.249	0.196	2,243,954,093	560,944,640
BASELINE	2	0.306	0.262	0.206	1,142,441,796	281,297,685

Table 29: Official results of the WMT21 Quality Estimation Task 2 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. "Rank" indicates the averaged ranking of participants with regards to all metrics (including memory print and number of parameters).

Model	Rank	**Pearson** r	MAE	RMSE	Disk footprint (B)	# Model params
• IST-Unbabel	2	0.275	0.279	0.224	2,294,887,576	569,368,609
• HW-TSC	1.8	0.262	0.278	0.228	2,243,954,093	560,944,640
BASELINE	2.2	0.098	0.279	0.232	1,142,441,796	281,297,685

Table 30: Official results of the WMT21 Quality Estimation Task 2 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. "Rank" indicates the averaged ranking of participants with regards to all metrics (including memory print and number of parameters).

Model	Rank	**Pearson** r	MAE	RMSE	Disk footprint (B)	# Model params
• IST-Unbabel	2.2	0.555	0.328	0.284	2,294,887,576	569,368,609
• HW-TSC	1.6	0.534	0.298	0.232	2,243,954,093	560,944,640
BASELINE	2.2	0.503	0.333	0.290	1,142,441,796	281,297,685

Table 31: Official results of the WMT21 Quality Estimation Task 2 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. "Rank" indicates the averaged ranking of participants with regards to all metrics (including memory print and number of parameters).

Model	Rank	**Pearson** r	MAE	RMSE	Disk footprint (B)	# Model params
• HW-TSC	1.8	0.753	0.165	0.111	2,243,954,093	560,944,640
IST-Unbabel	3.4	0.655	0.243	0.199	2,294,887,576	569,368,609
BASELINE	2	0.576	0.241	0.196	1,142,441,796	281,297,685
ENSBRT	2.8	0.530	0.262	0.197	1,363,652,116	167,357,185

Table 32: Official results of the WMT21 Quality Estimation Task 2 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. "Rank" indicates the averaged ranking of participants with regards to all metrics (including memory print and number of parameters).

C Official Results of the WMT21 Quality Estimation Task 2 (Word-level)

Tables 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43 and 44 show the results for all language pairs and the multilingual variant, ranking participating systems best to worst using Matthews correlation coefficient (MCC) as primary key for each of these cases.

Model	Rank	**MCC**	F1-BAD	F1-OK	F1-Multi	Disk footprint (B)	# Model params
				Words in MT			
HW-TSC	n/a	0.530	0.679	0.828	0.565	n/a	n/a
IST-Unbabel	n/a	0.430	0.628	0.787	0.486	n/a	n/a
BASELINE	n/a	0.346	0.579	0.717	0.402	n/a	n/a
				GAPs in MT			
HW-TSC	n/a	0.337	0.343	0.939	0.326	n/a	n/a
IST-Unbabel	n/a	0.196	0.209	0.975	0.203	n/a	n/a
BASELINE	n/a	0.126	0.137	0.973	0.133	n/a	n/a
				Words in SRC			
HW-TSC	n/a	0.432	0.592	0.799	0.473	n/a	n/a
IST-Unbabel	n/a	0.378	0.561	0.795	0.437	n/a	n/a
BASELINE	n/a	0.307	0.511	0.751	0.370	n/a	n/a

Table 33: Official results of the WMT21 Quality Estimation Task 2 (word-level) for the **Multilingual** task. 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).

Model	Rank	**MCC**	F1-BAD	F1-OK	F1-Multi	Disk footprint (B)	# Model params
				Words in MT			
• JHU-Microsoft	3	0.523	0.599	0.907	0.543	6,863,178,235	484,431,872
• HW-TSC	3.6	0.510	0.587	0.900	0.528	2,243,954,093	560,944,640
IST-Unbabel	3.8	0.466	0.551	0.914	0.504	2,294,887,576	569,368,609
Abulice‡	4.2	0.437	0.530	0.884	0.468	2,243,439,613	560,814,661
POSTECH	3	0.413	0.497	0.915	0.454	1,561,188,430	390,210,052
BASELINE	3.4	0.370	0.455	0.911	0.415	1,142,441,796	281,297,685
				GAPs in MT			
• HW-TSC	3.2	0.300	0.294	0.969	0.285	2,243,954,093	560,944,640
• JHU-Microsoft	3.4	0.256	0.266	0.985	0.262	6,863,178,235	484,431,872
IST-Unbabel	3.8	0.183	0.178	0.986	0.176	2,294,887,576	569,368,609
BASELINE	2.8	0.116	0.098	0.986	0.097	1,142,441,796	281,297,685
POSTECH	3.8	0.110	0.124	0.982	0.122	1,561,188,430	390,210,052
Abulice‡	–	–	–	–	–	–	–
				Words in SRC			
• HW-TSC	3.2	0.450	0.516	0.894	0.461	2,243,954,093	560,944,640
IST-Unbabel	3.8	0.404	0.483	0.921	0.445	2,294,887,576	569,368,609
Abulice‡	3.8	0.392	0.468	0.875	0.409	2,243,439,613	560,814,661
BASELINE	2.8	0.322	0.393	0.924	0.363	1,142,441,796	281,297,685
POSTECH	3.4	0.320	0.395	0.922	0.364	1,561,188,430	390,210,052
JHU-Microsoft	–	–	–	–	–	–	–
				Combined MT Words & Gaps			
• JHU-Microsoft	n/a	0.500	0.546	0.947	0.517	6,863,178,235	484,431,872
• HW-TSC	n/a	0.496	0.533	0.939	0.5	2,243,954,093	560,944,640
• IST-Unbabel	n/a	0.468	0.514	0.954	0.49	2,294,887,576	569,368,609
Abulice‡	n/a	0.442	0.488	0.934	0.456	2,243,439,613	560,814,661
BASELINE	n/a	0.378	0.42	0.952	0.4	1,142,441,796	281,297,685
POSTECH	n/a	0.403	0.45	0.952	0.428	1,561,188,430	390,210,052

Table 34: Official results of the WMT21 Quality Estimation Task 2 (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, 2000b). 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). ‡ indicates Codalab usernames of participants from whom we have not received further information.

Model	Rank	**MCC**	F1-BAD	F1-OK	F1-Multi	Disk footprint (B)	# Model params
				Words in MT			
• HW-TSC	2	0.354	0.497	0.806	0.401	2,243,954,093	560,944,640
IST-Unbabel	3	0.310	0.467	0.792	0.370	2,294,887,576	569,368,609
BASELINE	3	0.247	0.426	0.723	0.308	1,142,441,796	281,297,685
JHU-Microsoft	4	0.149	0.357	0.751	0.268	6,863,178,235	484,431,872
Abulice‡	3	0.033	0.254	0.770	0.196	100,000	9,501,148
				GAPs in MT			
• HW-TSC	2.6	0.172	0.160	0.934	0.149	2,243,954,093	560,944,640
IST-Unbabel	3	0.068	0.083	0.982	0.082	2,294,887,576	569,368,609
BASELINE	2.4	0.065	0.092	0.969	0.089	1,142,441,796	281,297,685
JHU-Microsoft	3.6	0.035	0.051	0.981	0.050	6,863,178,235	484,431,872
Abulice‡	–	–	–	–	–	–	–
				Words in SRC			
• HW-TSC	2.2	0.310	0.443	0.813	0.360	2,243,954,093	560,944,640
IST-Unbabel	3.2	0.286	0.427	0.803	0.343	2,294,887,576	569,368,609
BASELINE	3.2	0.241	0.394	0.751	0.295	1,142,441,796	281,297,685
Abulice‡	3	0.011	0.222	0.769	0.171	100,000	9,501,148
JHU-Microsoft	–	–	–	–	–	–	–
				Combined MT Words & Gaps			
• IST-Unbabel	n/a	0.369	0.441	0.904	0.398	2,294,887,576	569,368,609
• HW-TSC	n/a	0.359	0.424	0.88	0.373	2,243,954,093	560,944,640
BASELINE	n/a	0.32	0.393	0.871	0.342	1,142,441,796	281,297,685
JHU-Microsoft	n/a	0.24	0.337	0.884	0.298	6,863,178,235	484,431,872
Abulice‡	n/a	0.118	0.228	0.884	0.201	100,000	9,501,148

Table 35: Official results of the WMT21 Quality Estimation Task 2 (word-level) for the **English-Chinese** 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, 2000b). 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). ‡ indicates Codalab usernames of participants from whom we have not received further information.

Model	Rank	**MCC**	F1-BAD	F1-OK	F1-Multi	Disk footprint (B)	# Model params
				Words in MT			
• HW-TSC	2	0.666	0.740	0.910	0.673	2,243,954,093	560,944,640
• IST-Unbabel	2.6	0.649	0.729	0.915	0.667	2,294,881,977	569,368,609
• JHU-Microsoft	2.6	0.634	0.713	0.922	0.657	6,863,178,235	484,431,872
BASELINE	2.8	0.536	0.642	0.862	0.553	1,142,441,796	281,297,685
				GAPs in MT			
• HW-TSC	2.2	0.446	0.449	0.974	0.437	2,243,954,093	560,944,640
IST-Unbabel	2.6	0.357	0.377	0.980	0.370	2,294,881,977	569,368,609
JHU-Microsoft	2.8	0.208	0.162	0.983	0.159	6,863,178,235	484,431,872
BASELINE	2.4	0.205	0.229	0.976	0.223	1,142,441,796	281,297,685
				Words in SRC			
• HW-TSC	2	0.614	0.694	0.898	0.623	2,243,954,093	560,944,640
• IST-Unbabel	2.6	0.603	0.689	0.910	0.627	2,294,881,977	569,368,609
BASELINE	2.6	0.511	0.618	0.871	0.539	1,142,441,796	281,297,685
JHU-Microsoft	–	–	–	–	–	–	–
				Combined MT Words & Gaps			
• HW-TSC	n/a	0.656	0.694	0.947	0.657	2,243,954,093	560,944,640
• IST-Unbabel	n/a	0.64	0.686	0.952	0.653	2,294,881,977	569,368,609
JHU-Microsoft	n/a	0.612	0.656	0.954	0.626	6,863,178,235	484,431,872
BASELINE	n/a	0.543	0.598	0.929	0.556	1,142,441,796	281,297,685

Table 36: Official results of the WMT21 Quality Estimation Task 2 (word-level) for the **Romanian-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, 2000b). 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).

Model	Rank	**MCC**	F1-BAD	F1-OK	F1-Multi	Disk footprint (B)	# Model params
				Words in MT			
• HW-TSC	1.6	0.606	0.703	0.902	0.634	2,243,954,093	560,944,640
JHU-Microsoft	2.4	0.572	0.688	0.882	0.607	6,863,178,235	484,431,872
IST-Unbabel	3.2	0.570	0.687	0.880	0.605	2,294,887,576	569,368,609
BASELINE	2.8	0.461	0.589	0.869	0.512	1,142,441,796	281,297,685
				GAPs in MT			
• HW-TSC	2.2	0.312	0.334	0.969	0.324	2,243,954,093	560,944,640
IST-Unbabel	2.8	0.254	0.271	0.977	0.265	2,294,887,576	569,368,609
JHU-Microsoft	2.6	0.218	0.213	0.980	0.209	6,863,178,235	484,431,872
BASELINE	2.4	0.136	0.135	0.979	0.132	1,142,441,796	281,297,685
				Words in SRC			
• HW-TSC	1.8	0.549	0.650	0.899	0.584	2,243,954,093	560,944,640
IST-Unbabel	2.8	0.522	0.633	0.885	0.561	2,294,887,576	569,368,609
BASELINE	2.6	0.405	0.522	0.879	0.459	1,142,441,796	281,297,685
JHU-Microsoft	–	–	–	–	–	–	–
				Combined MT Words & Gaps			
• HW-TSC	n/a	0.584	0.644	0.94	0.605	2,243,954,093	560,944,640
• IST-Unbabel	n/a	0.582	0.644	0.937	0.604	2,294,881,977	569,368,609
• JHU-Microsoft	n/a	0.572	0.636	0.936	0.595	6,863,178,235	484,431,872
BASELINE	n/a	0.482	0.545	0.932	0.508	1,142,441,796	281,297,685

Table 37: Official results of the WMT21 Quality Estimation Task 2 (word-level) for the **Estonian-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, 2000b). 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).

Model	Rank	**MCC**	F1-BAD	F1-OK	F1-Multi	Disk footprint (B)	# Model params
				Words in MT			
• HW-TSC	1.6	0.674	0.876	0.795	0.696	2,243,954,093	560,944,640
IST-Unbabel	2.6	0.508	0.842	0.652	0.549	2,294,881,977	569,368,609
BASELINE	2.2	0.440	0.828	0.583	0.483	1,142,441,796	281,297,685
JHU-Microsoft	3.6	0.329	0.813	0.299	0.243	6,863,178,235	484,431,872
				GAPs in MT			
• HW-TSC	2	0.403	0.435	0.961	0.418	2,243,954,093	560,944,640
IST-Unbabel	2.4	0.268	0.284	0.969	0.276	2,294,881,977	569,368,609
BASELINE	2.2	0.215	0.249	0.963	0.240	1,142,441,796	281,297,685
JHU-Microsoft	3.4	0.207	0.253	0.953	0.241	6,863,178,235	484,431,872
				Words in SRC			
• HW-TSC	1.8	0.545	0.787	0.754	0.594	2,243,954,093	560,944,640
• IST-Unbabel	2.8	0.445	0.782	0.631	0.493	2,294,881,977	569,368,609
BASELINE	2.6	0.390	0.768	0.570	0.438	1,142,441,796	281,297,685
JHU-Microsoft	–	–	–	–	–	–	–
				Combined MT Words & Gaps			
• HW-TSC	n/a	0.749	0.833	0.915	0.763	2,243,954,093	560,944,640
IST-Unbabel	n/a	0.705	0.809	0.894	0.723	2,294,881,977	569,368,609
BASELINE	n/a	0.672	0.79	0.877	0.693	1,142,441,796	281,297,685
JHU-Microsoft	n/a	0.637	0.77	0.83	0.639	6,863,178,235	484,431,872

Table 38: Official results of the WMT21 Quality Estimation Task 2 (word-level) for the **Nepalese-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, 2000b). 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).

Model	Rank	**MCC**	F1-BAD	F1-OK	F1-Multi	Disk footprint (B)	# Model params
				Words in MT			
• HW-TSC	1.4	0.847	0.937	0.910	0.853	2,243,954,093	560,944,640
IST-Unbabel	2.4	0.528	0.822	0.683	0.561	2,294,887,576	569,368,609
BASELINE	2.2	0.425	0.793	0.574	0.456	1,142,441,796	281,297,685
				GAPs in MT			
• HW-TSC	1.4	0.639	0.651	0.979	0.638	2,243,954,093	560,944,640
IST-Unbabel	2.4	0.258	0.271	0.972	0.263	2,294,887,576	569,368,609
BASELINE	2.2	0.208	0.239	0.966	0.231	1,142,441,796	281,297,685
				Words in SRC			
• HW-TSC	1.4	0.616	0.804	0.810	0.651	2,243,954,093	560,944,640
IST-Unbabel	2.4	0.406	0.722	0.627	0.452	2,294,887,576	569,368,609
BASELINE	2.2	0.335	0.698	0.544	0.379	1,142,441,796	281,297,685
				Combined MT Words & Gaps			
• HW-TSC	n/a	0.868	0.909	0.958	0.872	2,243,954,093	560,944,640
IST-Unbabel	n/a	0.69	0.79	0.896	0.708	2,294,881,977	569,368,609
BASELINE	n/a	0.642	0.758	0.87	0.660	1,142,441,796	281,297,685

Table 39: Official results of the WMT21 Quality Estimation Task 2 (word-level) for the **Sinhala-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, 2000b). 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).

Model	Rank	**MCC**	F1-BAD	F1-OK	F1-Multi	Disk footprint (B)	# Model params
				Words in MT			
• HW-TSC	1.8	0.451	0.553	0.892	0.493	2,243,954,093	560,944,640
IST-Unbabel	2.6	0.332	0.430	0.896	0.386	2,294,887,576	569,368,609
JHU-Microsoft	3	0.303	0.439	0.847	0.372	6,863,178,235	484,431,872
BASELINE	2.6	0.256	0.360	0.889	0.319	1,142,441,796	281,297,685
				GAPs in MT			
• HW-TSC	2.2	0.388	0.393	0.962	0.378	2,243,954,093	560,944,640
JHU-Microsoft	2.6	0.167	0.159	0.978	0.156	6,863,178,235	484,431,872
IST-Unbabel	3	0.165	0.160	0.978	0.156	2,294,887,576	569,368,609
BASELINE	2.2	0.073	0.051	0.979	0.050	1,142,441,796	281,297,685
				Words in SRC			
• HW-TSC	2.2	0.426	0.540	0.876	0.473	2,243,954,093	560,944,640
IST-Unbabel	2.6	0.351	0.438	0.899	0.394	2,294,887,576	569,368,609
BASELINE	2.4	0.251	0.326	0.893	0.292	1,142,441,796	281,297,685
JHU-Microsoft	–	–	–	–	–	–	–
				Combined MT Words & Gaps			
• HW-TSC	n/a	0.456	0.514	0.931	0.479	2,243,954,093	560,944,640
IST-Unbabel	n/a	0.339	0.39	0.941	0.367	2,294,881,977	569,368,609
JHU-Microsoft	n/a	0.329	0.406	0.919	0.373	6,863,178,235	484,431,872
BASELINE	n/a	0.319	0.939	0.299	0.139	1,142,441,796	281,297,685

Table 40: Official results of the WMT21 Quality Estimation Task 2 (word-level) for the **Russian-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, 2000b). 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).

Model	Rank	**MCC**	F1-BAD	F1-OK	F1-Multi	Disk footprint (B)	# Model params
				Words in MT			
• HW-TSC	1.6	0.380	0.502	0.864	0.433	2,243,954,093	560,944,640
• IST-Unbabel	2.2	0.376	0.493	0.865	0.426	2,294,887,576	569,368,609
BASELINE	2.2	0.273	0.454	0.819	0.372	1,142,441,796	281,297,685
				GAPs in MT			
• HW-TSC	1.8	0.213	0.188	0.945	0.178	2,243,954,093	560,944,640
IST-Unbabel	2.4	0.125	0.143	0.981	0.141	2,294,887,576	569,368,609
BASELINE	1.8	0.039	0.054	0.983	0.053	1,142,441,796	281,297,685
				Words in SRC			
• HW-TSC	1.4	0.313	0.426	0.886	0.377	2,243,954,093	560,944,640
• IST-Unbabel	2.4	0.294	0.410	0.883	0.362	2,294,887,576	569,368,609
BASELINE	2.2	0.224	0.362	0.862	0.312	1,142,441,796	281,297,685
				Combined MT Words & Gaps			
• IST-Unbabel	n/a	0.4	0.459	0.931	0.427	2,294,881,977	569,368,609
BASELINE	n/a	0.339	0.425	0.914	0.389	1,142,441,796	281,297,685
HW-TSC	n/a	0.336	0.427	0.909	0.388	2,243,954,093	560,944,640

Table 41: Official results of the WMT21 Quality Estimation Task 2 (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, 2000b). 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).

Model	Rank	**MCC**	F1-BAD	F1-OK	F1-Multi	Disk footprint (B)	# Model params
				Words in MT			
• HW-TSC	1.6	0.258	0.495	0.625	0.309	2,243,954,093	560,944,640
IST-Unbabel	2.4	0.169	0.416	0.742	0.309	2,294,887,576	569,368,609
BASELINE	2	0.131	0.437	0.497	0.217	1,142,441,796	281,297,685
				GAPs in MT			
• HW-TSC	1.8	0.152	0.180	0.763	0.137	2,243,954,093	560,944,640
BASELINE	1.6	0.036	0.060	0.962	0.057	1,142,441,796	281,297,685
IST-Unbabel	2.6	0.025	0.016	0.969	0.015	2,294,887,576	569,368,609
				Words in SRC			
• HW-TSC	1.8	0.217	0.416	0.602	0.250	2,243,954,093	560,944,640
• IST-Unbabel	2.2	0.210	0.394	0.808	0.318	2,294,887,576	569,368,609
BASELINE	2	0.175	0.393	0.693	0.272	1,142,441,796	281,297,685
				Combined MT Words & Gaps			
BASELINE	n/a	0.25	0.403	0.79	0.319	1,142,441,796	281,297,685
IST-Unbabel	n/a	0.217	0.352	0.865	0.304	2,294,881,977	569,368,609
HW-TSC	n/a	0.186	0.361	0.677	0.244	2,243,954,093	560,944,640

Table 42: Official results of the WMT21 Quality Estimation Task 2 (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, 2000b). 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).

Model	Rank	**MCC**	F1-BAD	F1-OK	F1-Multi	Disk footprint (B)	# Model params
				Words in MT			
• HW-TSC	1.6	0.450	0.723	0.727	0.525	2,243,954,093	560,944,640
IST-Unbabel	2.6	0.370	0.685	0.684	0.469	2,294,887,576	569,368,609
BASELINE	2.4	0.313	0.674	0.631	0.425	1,142,441,796	281,297,685
JHU-Microsoft	3.4	0.191	0.677	0.170	0.115	6,863,178,235	484,431,872
				GAPs in MT			
• HW-TSC	2.2	0.260	0.262	0.942	0.246	2,243,954,093	560,944,640
IST-Unbabel	2.6	0.177	0.193	0.976	0.188	2,294,887,576	569,368,609
BASELINE	2	0.134	0.145	0.977	0.142	1,142,441,796	281,297,685
JHU-Microsoft	3.2	0.118	0.153	0.951	0.146	6,863,178,235	484,431,872
				Words in SRC			
• HW-TSC	2	0.304	0.538	0.723	0.389	2,243,954,093	560,944,640
• IST-Unbabel	2.6	0.294	0.522	0.758	0.396	2,294,887,576	569,368,609
BASELINE	2.6	0.249	0.501	0.720	0.361	1,142,441,796	281,297,685
JHU-Microsoft	–	–	–	–	–	–	–
				Combined MT Words & Gaps			
• IST-Unbabel	n/a	0.538	0.658	0.88	0.579	2,294,881,977	569,368,609
• HW-TSC	n/a	0.533	0.661	0.868	0.574	2,243,954,093	560,944,640
• JHU-Microsoft	n/a	0.523	0.648	0.782	0.507	6,863,178,235	484,431,872
BASELINE	n/a	0.517	0.648	0.867	0.562	1,142,441,796	281,297,685

Table 43: Official results of the WMT21 Quality Estimation Task 2 (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, 2000b). 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).

Model	Rank	**MCC**	F1-BAD	F1-OK	F1-Multi	Disk footprint (B)	# Model params
				Words in MT			
• HW-TSC	1.4	0.636	0.853	0.779	0.664	2,243,954,093	560,944,640
IST-Unbabel	2.4	0.448	0.790	0.638	0.503	2,294,887,576	569,368,609
BASELINE	.2	0.351	0.766	0.534	0.409	1,142,441,796	281,297,685
				GAPs in MT			
• HW-TSC	1.8	0.419	0.426	0.928	0.395	2,243,954,093	560,944,640
IST-Unbabel	2.2	0.259	0.274	0.964	0.264	2,294,887,576	569,368,609
BASELINE	2	0.175	0.204	0.959	0.195	1,142,441,796	281,297,685
				Words in SRC			
• HW-TSC	1.4	0.410	0.698	0.634	0.443	2,243,954,093	560,944,640
• IST-Unbabel	2.4	0.345	0.668	0.618	0.413	2,294,887,576	569,368,609
BASELINE	2.2	0.279	0.644	0.552	0.355	1,142,441,796	281,297,685
				Combined MT Words & Gaps			
• HW-TSC	n/a	0.677	0.783	0.883	0.692	2,243,954,093	560,944,640
IST-Unbabel	n/a	0.631	0.751	0.877	0.659	2,294,881,977	569,368,609
BASELINE	n/a	0.587	0.725	0.853	0.618	1,142,441,796	281,297,685

Table 44: Official results of the WMT21 Quality Estimation Task 2 (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, 2000b). 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).

D Official Results of the WMT21 Quality Estimation Task 3 (Sentence-level)

Tables 45, 46, 47 and 48 show the results for all language pairs, ranking participating systems best to worst using Matthews correlation coefficient (MCC) as primary key for each of these cases.

Model	Rank	**MCC**	F1-ERR	F1-NOT	F1-Multi	Disk footprint (B)	# Model params
• NICT Kyoto	1.5	0.546	0.877	0.667	0.585	2,239,774,281	559,892,482
LAMA-ICL	2.67	0.498	0.868	0.623	0.541	2,239,830,893	559,908,866
HW-TSC	4.17	0.490	0.867	0.613	0.532	2,241,232,523	561,947,562
QEMind	4	0.480	0.854	0.625	0.534	2,244,034,844	560,982,532
silence1024‡	4.33	0.449	0.850	0.597	0.507	2,239,747,529	560,365,209
BASELINE	4.33	0.397	0.848	0.532	0.451	1,114,634,523	278,635,778

Table 45: Official results of the WMT21 Quality Estimation Task 3 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 William's test. 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). ‡ indicates Codalab usernames of participants from whom we have not received further information.

Model	Rank	**MCC**	F1-ERR	F1-NOT	F1-Multi	Disk footprint (B)	# Model params
• HW-TSC	2.83	0.353	0.889	0.462	0.411	2,241,232,523	531,947,562
• silence1024‡	3.5	0.343	0.888	0.453	0.402	2,239,747,529	560,365,209
• NICT Kyoto	4	0.311	0.883	0.426	0.376	2,239,774,281	559,892,482
LAMA-ICL	4.33	0.305	0.892	0.413	0.368	2,239,830,893	559,908,866
QEMind	5.33	0.278	0.893	0.384	0.343	2,244,034,844	560,982,532
BASELINE	4	0.187	0.898	0.269	0.242	1,114,634,523	278,635,778
serkan‡	4	0.141	0.913	0.131	0.120	1,112,236,548	1,024

Table 46: Official results of the WMT21 Quality Estimation Task 3 for the **English-Chinese** dataset. Teams marked with "•" correspond to the winning submissions, as they are not significantly outperformed by any other system based on William's test. 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). ‡ indicates Codalab usernames of participants from whom we have not received further information.

Model	Rank	**MCC**	F1-ERR	F1-NOT	F1-Multi	Disk footprint (B)	# Model params
• NICT Kyoto	1.83	0.511	0.913	0.595	0.543	2,239,774,281	559,892,482
• LAMA-ICL	2.17	0.473	0.911	0.555	0.506	2,239,765,357	559,892,482
QEMind	3.8	0.454	0.909	0.534	0.485	2,244,034,844	560,982,532
HW-TSC	3.33	0.448	0.906	0.537	0.486	2,234,153,425	560,365,209
BASELINE	3.67	0.388	0.899	0.477	0.429	1,114,634,523	278,635,778

Table 47: Official results of the WMT21 Quality Estimation Task 3 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 William's test. 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).

Model	Rank	**MCC**	F1-ERR	F1-NOT	F1-Multi	Disk footprint (B)	# Model params
HW-TSC	2.5	0.318	0.937	0.378	0.354	2,239,747,529	560,365,209
LAMA-ICL	2.83	0.314	0.956	0.336	0.321	2,239,769,453	559,893,506
Jason_pogba‡	3.83	0.278	0.936	0.341	0.319	2,213,468,431	564,554,219
silence1024‡	4	0.277	0.940	0.337	0.317	2,239,747,529	560,365,209
QEMind	5.33	0.260	0.953	0.288	0.274	2,244,034,844	560,982,532
NICT Kyoto	5.17	0.252	0.929	0.319	0.297	2,239,774,281	559,892,482
BASELINE	4.33	0.214	0.951	0.244	0.232	1,114,634,523	278,635,778

Table 48: Official results of the WMT21 Quality Estimation Task 3 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 William's test. 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). ‡ indicates Codalab usernames of participants from whom we have not received further information.