

Findings of the WMT 2020 shared task on quality estimation

Item Type	Conference contribution
Authors	Specia, Lucia;Blain, Frédéric;Fomicheva, Marina;Fonseca, Erick;Chaudhary, Vishrav;Guzmán, Francisco;Martins, André FT
Citation	Specia, L., Blain, F., Fomicheva, M. et al. (2020) Findings of the WMT 2020 shared task on quality estimation, Proceedings of the Fifth Conference on Machine Translation, November 2020, pp. 743–764.
Publisher	Association for Computational Linguistics
Download date	2026-05-18 06:20:56
License	https://creativecommons.org/licenses/by/4.0/
Link to Item	http://hdl.handle.net/2436/623855

Findings of the WMT 2020 Shared Task on Quality Estimation

Lucia Specia,^{1,2} Frédéric Blain,^{2,3} Marina Fomicheva,² Erick Fonseca,⁴
Vishrav Chaudhary,⁵ Francisco Guzmán,⁵ André F. T. Martins^{4,6}

¹Imperial College London, ²University of Sheffield, ³University of Wolverhampton,
⁴Instituto de Telecomunicações, ⁵Facebook AI, ⁶Unbabel
{l.specia,m.fomicheva}@sheffield.ac.uk, f.blain@wlv.ac.uk
erick.fonseca@lx.it.pt, {vishrav, fguzman}@fb.com,
andre.martins@unbabel.com

Abstract

We report the results of the WMT20 shared task on Quality Estimation, where the challenge is to predict the quality of the output of neural machine translation systems at the word, sentence and document levels. This edition included new data with open domain texts, direct assessment annotations, and multiple language pairs: English–German, English–Chinese, Russian–English, Romanian–English, Estonian–English, Sinhala–English and Nepali–English data for the sentence-level subtasks, English–German and English–Chinese for the word-level subtask, and English–French data for the document-level subtask. In addition, we made neural machine translation models available to participants. 19 participating teams from 27 institutions submitted altogether 1374 systems to different task variants and language pairs.

1 Introduction

This shared task builds on its previous eight editions to further examine automatic methods for estimating the quality of neural machine translation (MT) output at run-time, without the use of reference translations. As in previous editions, it includes the (sub)tasks of word-level, sentence-level and document-level estimation. Important elements introduced this year are: a variant of the sentence-level task where sentences are annotated with *direct assessment* (DA)¹ 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

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

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 create a new set of public benchmarks for tasks in quality estimation.
- To investigate models for predicting DA scores and their relationship with models trained for predicting post-editing effort,
- To study the feasibility of multilingual (or even language independent) approaches to QE.
- To study the influence of source-language document-level context for the task of QE.
- To analyse the applicability of NMT model information for QE.

We have three subtasks: Task 1 aims at predicting 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 score for an entire document as a function of the proportion of incorrect words in such a document, weighted by the severity of the different errors (Section 2.3).

Tasks make use of large datasets produced from either post-editions or DA annotations, or error annotation, all done by professional translators. The text domains vary for each subtask. Neural MT systems were built on freely available data using an open-source toolkit to produce translations, and these models were made available to participants. We provide new training and test datasets for Tasks 1 and 2, and a new test set for Task 3. 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. Results for all tasks evaluated according to standard metrics are given in Section 5, while a discussion on the main 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, a **new dataset**, was created containing seven languages pairs using sentences mostly from Wikipedia². These language pairs are divided into 3 categories: the high-resource English→German (En-De), English→Chinese (En-Zh) and Russian→English (Ru-En) pairs; the medium-resource Romanian→English (Ro-En) and Estonian→English (Et-En) pairs; and the low-resource Sinhala→English (Si-En) and Nepali→English (Ne-En) pairs.

Translations were produced with state-of-the-art transformer-based NMT models trained using publicly available data and the fairseq toolkit (Ott et al., 2019); and were manually annotated for perceived quality. The quality label for this task ranges 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.

²This dataset is a superset of MLQE (Fomicheva et al., 2020c) which included 6 language pairs and is sourced entirely from Wikipedia. The newly-added English-Russian DAs follow the same guidelines, but come from diverse sources.

Statistics on the dataset are shown in Table 1. More details are given in Fomicheva et al. (2020a). The complete data can be downloaded from the public repository³.

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 follows from previous editions of the WMT shared task and consists in scoring translations according to the proportion of their words that need to be fixed using HTER as label, i.e. 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 uses a subset of the languages from Task 1, namely the two high-resource language pairs (En-De and En-Zh, Table 1).

Sentence-level post-editing effort The label for this task is the percentage of edits that need to be fixed (HTER). Starting with the En-De and En-Zh source-machine translation segment pairs, the machine translation sentences were post-edited by two human translators, one per language, who are paid editors from the Unbabel community. The two translators had no access to the direct assessments above. In other words, the DA and HTER annotations were collected independently.

The average human translation error rate between the machine translated text and the post-edited text was 0.32 for En-De, and 0.62 for En-Zh. HTER labels were computed using TERCOM⁴ with default settings (tokenised, case insensitive, exact matching only), with scores capped to 1.

Word-level errors This variant evaluates the extent to which we can detect word-level errors in MT output. Based on the post-edited translations, as described above, we annotate each token of the target and the source sentence, as well as word omission errors. The code to produce this set of tags from any prior WMT corpora is available for

³<https://github.com/sheffieldnlp/mlqe-pe>

⁴<https://github.com/jhclark/tercom>

Languages	Sentences			Tokens			DA	PE
	Train	Dev	Test	Train	Dev	Test		
En-De	7,000	1,000	1,000	114,980	16,519	16,371	✓	✓
En-Zh	7,000	1,000	1,000	115,585	16,307	16,765	✓	✓
Ru-En	7,000	1,000	1,000	82,229	11,992	11,760	✓	
Ro-En	7,000	1,000	1,000	120,198	17,268	17,001	✓	
Et-En	7,000	1,000	1,000	98,080	14,423	14,358	✓	
Ne-En	7,000	1,000	1,000	104,934	15,144	14,770	✓	
Si-En	7,000	1,000	1,000	109,515	15,708	15,821	✓	

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

download.⁵ More specifically, the following types of labels were produced:

- **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.

In order to obtain the labels, we first align source and MT using the IBM Model 2 alignments from FastAlign (Dyer et al., 2013), and compute edit distances between the generated and post-edited translations with TERCOM, using default settings and disabled shifts.

2.3 Task 3: Predicting document-level MQM

This task consists in finding document-level translation errors and estimating a quality score according to the amount of minor, major, and critical errors present in the translation. The predictions are compared to a ground-truth obtained from annotations produced by crowd-sourced human translators from Unbabel community.

Each document contains zero or more errors, annotated according to the MQM taxonomy⁶, and

⁵<https://github.com/deep-spin/ge-corpus-builder>

⁶Multidimensional Quality Metrics; see <http://www.qt21.eu/mqm-definition/definition-2015-12-30.html> for details.



Figure 1: Example of fine-grained document annotation. Spans in the same color belong to the same annotation. Error severity and type are not shown for brevity.

may span one or more tokens, not necessarily contiguous. Errors have a label specifying their type, such as wrong word order, missing words, agreement, etc. They provide additional information, but do not need to be predicted by the systems. Additionally, there are three severity levels for errors: *minor* (if it is not misleading nor changes meaning), *major* (if it changes meaning), and *critical* (if it changes meaning and carries any kind of implication, possibly offensive).

Figure 1 shows an example of fine-grained error annotations for a sentence. Note that there is an annotation composed by two discontinuous spans: a whitespace and the token *Grip* — in this case, the annotation indicates wrong word order, and *Grip* should have been at the whitespace position.

Document-level scores were then generated from the word-level errors and their severity using the method described in Sanchez-Torron and Koehn (2016, footnote 6). Namely, denoting by n the number of words in the document, and by n_{\min} , n_{maj} , and n_{cri} the number of annotated minor, major, and critical errors, the final quality scores were computed as:

$$\text{MQM} = 1 - \frac{n_{\text{minor}} + 5n_{\text{major}} + 10n_{\text{crit}}}{n} \quad (1)$$

Note that MQM values can be negative if the total severity exceeds the number of words.

As this year’s dataset, we reused the training data from previous years, adding the test sets from 2018 and 2019 to the training set, keeping the same development set from 2019, and released a new test set. The documents are short product title and descriptions in English, extracted from the Amazon Product Reviews dataset (McAuley et al., 2015; He and McAuley, 2016) (Sports and Outdoors category). The documents were machine translated into French using a state of the art online neural MT system. The dataset statistics are presented in Table 2.

3 Baseline systems

Sentence-level baseline systems: For Tasks 1 and 2, both word and sentence-level, we used the LSTM-based Predictor-Estimator approach (Kim et al., 2017), implemented in OpenKiwi (Kepler et al., 2019b). The Predictor model was trained on the same parallel data as the NMT systems for each language pair (made available at the task website),⁷ while the the Estimator was trained on the 7, 000 QE labelled data for each task.

Word-level baseline systems: For Task 2, we also used the Predictor-Estimator as above, but it was trained to predict jointly word-level tags and sentence-level scores.

Document-level baseline system: For Task 3, similarly as last year, we used a baseline which treats sentences independently and casts the problem as word-level QE, such that all words and gaps within an error span are given the tag BAD. We then trained a Predictor-Estimator model, regrouping any contiguous sequence of tokens tagged as BAD in a single error annotation. In order to get MQM scores, instead of computing the value according to its definition, we compute it simply as 1 minus the the ratio of BAD tags.

4 Participants

Table 3 lists all participating teams submitting systems to any of the tasks, and Table 4 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,

⁷<http://statmt.org/wmt20/quality-estimation-task.html>

participation in specific tasks is denoted by a task identifier (T1 = Task 1, T2 = Task 2, T3 = Task 3).

Bergamot-LATTE (T1): Bergamot-LATTE submitted two systems to the two variants of sentence-level predictions: (i) a black-box approach based on pre-trained representations; (ii) an unsupervised glass-box approach that leverages information extracted from the neural MT system. The black-box model consists of stacking a 2-layer multilayer perceptron on the vector representation of the CLS token from the contextualised representation from XLM-R (Conneau et al., 2020), using both the source and the target sentences as input. The glass-box approach explores the best-performing unsupervised quality indicators presented in Fomicheva et al. (2020c) that rely on uncertainty quantification based on the Monte Carlo dropout method: D-TP and D-Lex-Sim.

Bergamot (T1, T2): Bergamot explores recent work on glass-box QE that exploits NMT output distribution and attention to capture uncertainty as a proxy to MT quality. Specifically, they use three groups of unsupervised quality indicators described in Fomicheva et al. (2020c) as features for a regression model.

Bering Lab (T2): Bering Lab proposes a fine-tuned version of a pre-trained XLM-R model. The model is first trained on a huge artificial QE data that is created by (i) translating a parallel corpus with an OpenNMT system; and (ii) using the TER tool to produce artificial labels for both word- and sentence-levels. The model is then fine-tuned using the shared task’s data. For predictions at word-level, the final hidden vector of each token, including the $\langle S \rangle$, is fed into a linear layer with sigmoid activation in order to predict the probability of each of these token to be BAD. Quality labels for tokens and gaps are predicted separately with two distinct binary classification layers. For predictions at sentence-level, the final hidden vector of the first $\langle S \rangle$ token, considered as a pooled representation, is fed into two linear layers with \tanh activation. Submitted predictions are results of an ensemble of 5 models trained with different seeds: averaged predictions for sentence-level, and majority voting for word-level.

	Documents			Sentences			Tokens		
	Train	Dev	Test	Train	Dev	Test	Train	Dev	Test
En-Fr	1,448	200	180	8,592	1,301	895	189,735	28,092	18,545

Table 2: Statistics of the data used for Task 3. The number of tokens is computed based on the source sentences.

ID	Participating team	
Bergamot-LATTE	University of Sheffield & Imperial College London, UK & Johns Hopkins University & Facebook AI, US & University of Tartu, Estonia	(Fomicheva et al., 2020b)
Bergamot	University of Tartu, Estonia	(Fomicheva et al., 2020b)
Bering Lab	Bering Lab, Republic of Korea	(Lee, 2020)
Elturco.AI	Elturco AI, Turkey	–
FVCRC	Nagoya University, Japan & University of Sydney, Australia	(Zhou et al., 2020)
HW-TSC	Huawei Translation Services & East China Normal University, China	(Wang et al., 2020a)
IST & Unbabel	Instituto Superior Técnico Lisbon & Unbabel, Portugal	(Moura et al., 2020)
JXNU-CCLQ	Jiangxi Normal University, China	–
Mak	University of Wolverhampton, UK	–
NICT Kyoto	National Institute of ICT, Japan	(Rubino, 2020)
NiuTrans	Northeastern University & NiuTrans Research, China	(Hu et al., 2020)
NJUNLP	Nanjing University, China	(Cui et al., 2020)
Papago	KAIST & Naver, Republic of Korea	(Baek et al., 2020)
RTM	Boğaziçi University, Turkey	(Biçici, 2020)
TMUOU	Osaka University & Tokyo Metropolitan University, Japan	(Nakamachi et al., 2020)
Tencent Inc.	Tencent Inc, China	(Wang et al., 2020b)
TransQuest	University of Wolverhampton, UK	(Ranasinghe et al., 2020)
WL Research	WL Research, US, Canada and Turkey	(Kane et al., 2020)
XC	Imperial College London, UK	–

Table 3: Participants in the WMT20 Quality Estimation shared task.

Task/LP	# submission
Task 1 – Sent-level Direct Assessment	747
Multilingual	43
English-German	132
English-Chinese	146
Romanian-English	150
Nepali-English	56
Estonian-English	68
Sinhala-English	74
Russian-English	78
Task 2 – Post-Editing Effort	435
English-German (sent-level)	131
English-Chinese (sent-level)	235
English-German (word-level)	38
English-Chinese (word-level)	31
Task 3 – Document-Level QE	192
English-French (annot.)	97
English-French (score)	95
Total	1374

Table 4: Number of submissions to each sub-task and language-pair at the WMT20 Quality Estimation shared task. In the results (Section 5) we only report the top two submissions per team for each task and language pair.

Elturco.AI (T2): Elturco.AI uses a generative model and a discriminative model, inspired by Electra (Clark et al., 2020). The two mod-

els are jointly trained on a parallel corpus, in order to create increasingly difficult artificial samples for quality estimation. The generative model consists of a transformer encoder and two transformer decoders, for forward and backward direction. In addition to predicting tokens, it is also trained to predict gap locations on the target side given the source sentence and left and right contexts on the target side. Distorted translations are generated by sampling on generator outputs on token and gap locations, which can be shorter or longer than the original translation. The distorted translations are compared to original translations for generating token and gap tags. The discriminator, a transformer encoder-decoder with full attention mask on the decoder side, is trained to predict the generated tags given the source and distorted translation. Once trained, the discriminator is fine-tuned on the actual quality estimation dataset.

FVCRC (T1): FVCRC’s system builds on BERTScore, a text generation evaluation system based on pretrained BERT contextual embeddings, originally for Metrics tasks. By

using pre-trained multilingual BERT-based model, they experiment with BERTScore on QE tasks. Without reference translations, it makes more errors in terms of word (or sub-word) alignments when perform greedy matching on pairwise cosine similarity, which is believed to be main cause of its drop of performance in QE tasks. They introduce GIZA++ word (subword) alignments and n-grams similarity matching to tackle misalignments and sentence perplexity of candidate translation as additional information to the evaluation score. Otherwise, the default setting of BERTScore (Zhang* et al., 2020) is used: pre-trained bert-base-multilingual-cased and xlm-mlm-100-1280 for embedding extraction, with a single model. This system is not trained on human labels (DA) and is not optimised on additional data.

HW-TSC (T2): HW-TSC submissions follows the Predictor-Estimator architecture (Kim et al., 2017), with a pre-trained Transformer as Predictor, and task-specific classifiers and regressors as Estimators. HW-TSC uses a unified model to solve both word- and sentence-level tasks, trained under multi-task learning. To improve the transfer-learning efficiency across tasks while preventing over-fitting, a Bottleneck Adapter Layer (Houlsby et al., 2019) is added to the Transformer after the self-attention and the feedforward layers, while keeping the original parameters of the Transformer model fixed.

IST & Unbabel (T1, T2, T3): IST & Unbabel submitted two systems per task variant: OPENKIWI-BASE and KIWI-GLASS-BOX-ENSEMBLE for predicitions at both word- and sentence-levels; KIWI-DOC and KIWI-DOC-IOB for document-level predictions. OPENKIWI-BASE is based on the re-implementation of the Predictor-Estimator architecture (Kim et al., 2017) available in OpenKiwi (Kepler et al., 2019b): the Predictor model is replaced with pre-trained contextualised representations (such as BERT or XLM-R) and the bi-LSTM Estimator is replaced by linear layers. KIWI-GLASS-BOX-ENSEMBLE is similar to OPENKIWI-BASE with a bottleneck layer introduced in the Estimator in order to concatenate the features

extracted from the Predictor, with sentence-level uncertainty features extracted from the NMT system provided by the shared task. Those glass-box features are based on work by Fomicheva et al. (2020c) and exploit entropy measures at prediction time. Unlike OPENKIWI-BASE, the KIWI-GLASS-BOX-ENSEMBLE model is trained for source, target and sentence level predictions simultaneously, using a multi-task learning approach. KIWI-DOC is the same as in Kepler et al. (2019a) while KIWI-DOC-IOB frames the task of annotating as Name Entity Recognition task: the severity annotations are projected to tags in IOB format ('O', 'B-major', 'I-major', 'B-critical', etc.) and the model is trained with a CRF output layer to enforce correctness of the tag-sequence at prediction time. The predicted tags are converted into annotations without the resort to a grouping and labelling heuristic.

JXNU-CCLQ (T1): JXNU-CCLQ proposes a model composed of a Transformer bottleneck layer and a bidirectional LSTM. The parameters of the Transformer bottleneck layer are first optimised with a bilingual parallel corpus, and the entire model is then fine-tuned on the training quality labelled dataset of the shared task. At test time, the translation outputs, which are estimated with teacher forcing and special masking, are put together with the source sentences and put through a unified neural network model to predict the quality of the translations.

Mak (T1): Mak represents the source and its translation sentence pairs as a set of 70 black-box sentence-level features extracted with Quest++(Specia et al., 2015), using the resources used to train the English-Russian NMT system (Ng et al., 2019). Those features are then fitted into a support vector regressor with default settings.

NICT Kyoto (T2): The English-German and English-Chinese sentence-level QE systems for Task 2 are ensembles of pre-trained cross-lingual language models (XLM) (Conneau and Lample, 2019), fine-tuned in a multi-task fashion with two linear output layers for sentence and word-level quality estimation. A total of 8 XLM models with various masking hyper-parameters were domain-adapted

using a subset of the additional resources provided by the QE shared task organisers, as well as a subset of the WikiMatrix corpus [2]. The translation language model training approach (TLM) was used before fine-tuning the XLM models for the QE task, complemented with a novel self-supervised learning task which aims to model errors inherent to machine translation outputs.

NiuTrans (T1, T2, T3): For Task 1, NiuTrans explored the combination of pre-trained models and multi-task learning. They used three different model settings, including multilingual-bert (~200M parameters), xlm-roberta-base (~300M parameters) and xlm-roberta-large (~600M parameters). They continued pre-training all models on the WMT dataset and utilised task adaptive pre-training to further boost the models’ performance. The output of different models was combined using a weighting scheme to get final predictions. For Task 2, an ensemble of 10 transformer-based predictor-estimator models was used, with multi-task training for the word-level tasks. Each single model contains 10M parameters. They also submitted an ensemble result of multilingual-bert and xlm-roberta-base for sentence-level scoring tasks. For Task 3, they used an ensemble of 8 predictor-estimator models and multi-task training for the word-level subtask. The single model contains 150M parameters. For the scoring subtask, they explored an ensemble of linear regression models and pre-trained models. They also used WMT 2014 English-French dataset for fine-tuning.

NJUNLP (T2): This system is an ensemble using NuQE and QUETCH models (Kepler et al., 2019b), as well as the QE Brain model (Fan et al., 2019). In addition to these pre-existing models, the ensemble also uses a masked version of the QE Brain, where some tokens in the translation are masked during training, and a masked language model (Devlin et al., 2018). For sentence-level, the different models are used as feature extractors, which are used as inputs of a dense layer to produce the predictions. For word-level, they use majority voting to ensemble the different models.

Papago (T1, T3): Papago’s submission for Task 1

En-De is an ensemble of three models based on pre-trained contextualised representations: multilingual BERT (mBERT), XLM-Masked-Language-Modelling (XLM-MLM), and XLM-Causal-Language-Modelling (XLM-CLM). Three scores were produced from these models: an extension of BERTScore using the multilingual BERT model, SentenceBERT score (Reimers and Gurevych, 2019), and target (German) language model score using a pre-trained GPT-2 model. Additionally, the scores were computed for synthetic data created using WMT News translation data by randomly performing different methods, including swapping word order, omitting words or repeating phrases. The three models are pre-trained from these data in a multi-task regression setting. Lastly, these pre-trained models are fine-tuned using the QE corpus. For Task 3, the submitted system uses an ensemble of four models leveraging either multilingual BERT or XLM. The training scheme is very task-oriented: erroneous sentence pairs and their pseudo-MQM scores are generated from Europarl and this QE task’s training corpus.

RTM (T1, T2): For Task 1 and Task 2’s sentence-level prediction, the RTM model treats QE as a parallel semantic similarity prediction task within machine translation performance prediction (MTPP) or a monolingual semantic similarity when the source or the target language are unknown or have scarce resources. En-De and Ru-En were modelled as parallel MTPP and the rest as monolingual MTPP using only the English side of the training and development datasets. Machine learning algorithms including ridge regression, SVR, and regression trees were used and the submissions were constrained to the resources provided. RTM selects a subset of parallel and monolingual text for each translation direction.

TMUOU (T1): TMUOU proposes an ensemble of four regression models based on XLM-R large: model 1 uses the final hidden vector of the CLS token; model 2 concatenates the feature of model 1 with the mean of the final hidden vector of each token; models 3 and 4 are based on models 1 and 2, respectively, but

adds a special token for language identification at the beginning of each sentence. The ensemble model is a gradient boosting regressor that features the predictions of these four models, the sentence probability of the target translation system, and one-hot vectors that indicate both the source and target languages.

Tencent (T2): Tencent-TTL’s submission for the sentence-level Task 2 use a predictor-estimator model. They use two predictors as feature extractors: a transformer trained with WMT provided parallel corpus and a fine-tuned cross-lingual language model (XLM). For the XLM-based predictor, it produces two kinds of contextual token representations, i.e., masked representations and non-masked representations. For transformer-based predictor, only the non-masked representation is produced. The estimator was trained with LSTM or Transformer. Finally, they ensembled the systems with different models and the same model with different parameters using logistic regression to produce a single sentence-level prediction.

TransQuest (T1, T2): TransQuest proposes two architectures: MONOTRANSQUEST and SIAMESETRANSQUEST, both using pre-trained XLM-R large transformer model. The MONOTRANSQUEST architecture uses the computed CLS token pooled representation from a single transformer model and uses it as input of a softmax layer that predicts the quality score of the translation. The SIAMESETRANSQUEST architecture encodes both the source sentence and its translation with two separate XLM-R transformer models. For each transformer model, it computes the mean of all output vectors of the input words, and applies the cosine similarity measure between the two outputs. The final submission is an ensemble of the two architectures.

WL Research (T1): WL’s NUBIA method has three modules: a neural feature extractor, an aggregator and a calibrator. The feature extractor consists of different transformer-based architectures fine-tuned on relevant tasks of language evaluation such as semantic similarity (RoBERTa model fine-tuned on STS-B), logical inference (RoBERTa fine-tuned on

MNLI) and sentence likelihood (GPT2 perplexity score). The aggregator uses the features extracted by the transformers as well as non-neural features such as hypothesis sentence length and is trained to predict the quality of the hypothesis sentence. These features are then used to train a 10 hidden layer neural network. Given that NUBIA takes as input sentences in English, as pre-processing step, Google Translate was used to translate either the non-English candidate or source to have both in English.

XC (T1): This was a multilingual system trained using TransQuest (with BERT embeddings bert-base-multilingual-cased) and data for all language pairs concatenated. An attempt was also made to use project the BERT source and target sentence embeddings into a space where they are highly correlated using CCA (Canonical Correlation Analysis) followed by an MLP regressor trained to predict the quality score, but this did not perform as well as a vanilla TransQuest.

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 5 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 11, 12, 13, 14, 15, 16, 17 and 18 provide the detailed results for all language pairs and the multilingual variant, ranking participants by their performance for each of these cases. The detailed tables show a striking difference in performance by Pearson scores versus MAE/RMSE, especially for the top systems. This requires further investigation.

Best performers The two top performing systems, TransQuest and Bergamot-LATTE (black-

⁸<https://github.com/ygraham/mt-ge-eval>

Model	Si-En	Ne-En	Et-En	Ro-En	En-De	En-Zh	Ru-En	Multi
TransQuest	0.68	0.82	0.82	0.91	0.55	0.54	0.81	0.72
Bergamot-LATTE (black-box)	0.68	0.81	0.83	0.91	0.54	0.53	0.80	0.72
IST and Unbabel (Kiwi-glass-box-ensemble)	0.64	0.79	0.77	0.89	0.52	0.49	0.77	0.67
TMUOU	0.67	0.78	0.79	0.90	0.48	0.44	0.78	0.69
XC	0.63	0.78	0.76	0.88	0.47	0.47	0.78	-
WL Research	0.58	0.69	0.64	0.82	0.25	0.30	0.60	0.55
Bergamot	0.56	0.66	0.68	0.80	0.48	0.43	-	-
Bergamot-LATTE (glass-box)	0.51	0.60	0.64	0.69	0.26	0.32	-	0.49
IST and Unbabel (OpenKiwi-base)	0.56	0.60	0.69	0.71	0.27	0.35	-	0.58
BASELINE	0.37	0.39	0.48	0.68	0.15	0.19	0.55	0.38
FVCRC	0.39	0.49	-	0.65	0.11	0.08	0.40	-
RTM	0.54	-	0.61	0.70	-	0.26	-	-
Shrangin †	-	-	-	0.85	-	-	-	-
Mak	-	-	-	-	-	-	0.54	-
Papago	-	-	-	-	0.50	-	-	-
aj54 †	-	-	-	-	-	0.44	-	-
JXNU-CCLQ	-	-	-	-	-	0.43	-	-
jackielo †	-	-	-	-	-	-	0.41	0.46
zhanghuimeng †	-	-	-	-	0.39	-	-	-
DexinWang †	-	-	-	-	0.25	-	-	-
Hancheng-Deng †	-	-	-	-	0.17	-	-	-
NiuTrans †	0.70	0.83	0.83	0.92	0.56	0.55	0.82	0.73

Table 5: Pearson correlation with direct assessments for the submissions to WMT20 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 teams that have been identified as having submitted more systems than the allowed limit to the leaderboard; ‡ indicates Codalab username of participants from whom we have not received further information.

box) perform the same or very closely for all language pairs. Both make use of the XML-R large pre-trained representations, and ensembles. This is clearly a booster, as these systems achieve almost double the correlation of the baseline. Note that the baseline also uses pre-trained word embeddings, but these are obtained using much smaller datasets: those used to train the NMT models for each respective language pair.

Except for a few systems for some language pairs, the vast majority of submissions outperform the baseline system, often by a large margin, except for Russian-English which had fewer submissions and where 1/3 of the systems are below the baseline. It is hard to make any conclusions about this difference across languages as Russian-English systems that are below the baseline did not submit systems for other languages. In relative terms, the improvement over the baseline for top systems in this language is similar to the other language pairs. The range of performances is remarkably wide, with the winning systems often doubling the Pearson correlation of the bottom pack, notably for English-Chinese and English-German.

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. Note the remarkable difference in correlation between the baseline and the top performers across languages. In the figures, we can visualise the substantial gains are achieved, largely due to the use of strong pre-trained representations.

High-resource performance MT quality for the high-resource language pairs, in particular English-German, was the most challenging to predict. As discussed in Fomicheva et al. (2020a), the MT outputs for this language pair have little variability in terms of perceived MT quality. The vast majority of translations were assigned high scores during DA evaluation, which makes it difficult to capture any meaningful variation between the DA scores. We observe that the results for HTER prediction for this language pair are more positive, a difference which we discuss in Section 6.

Low-resource performance Interestingly, the results for the low-resource language pairs, Sinhala-English and Nepali-English, are comparable with the rest. The fact that the performance of the winning approaches based on multilingual pre-trained representations does not degrade for the low-resource language pairs is worth noticing. It could indicate that: (i) the source language does

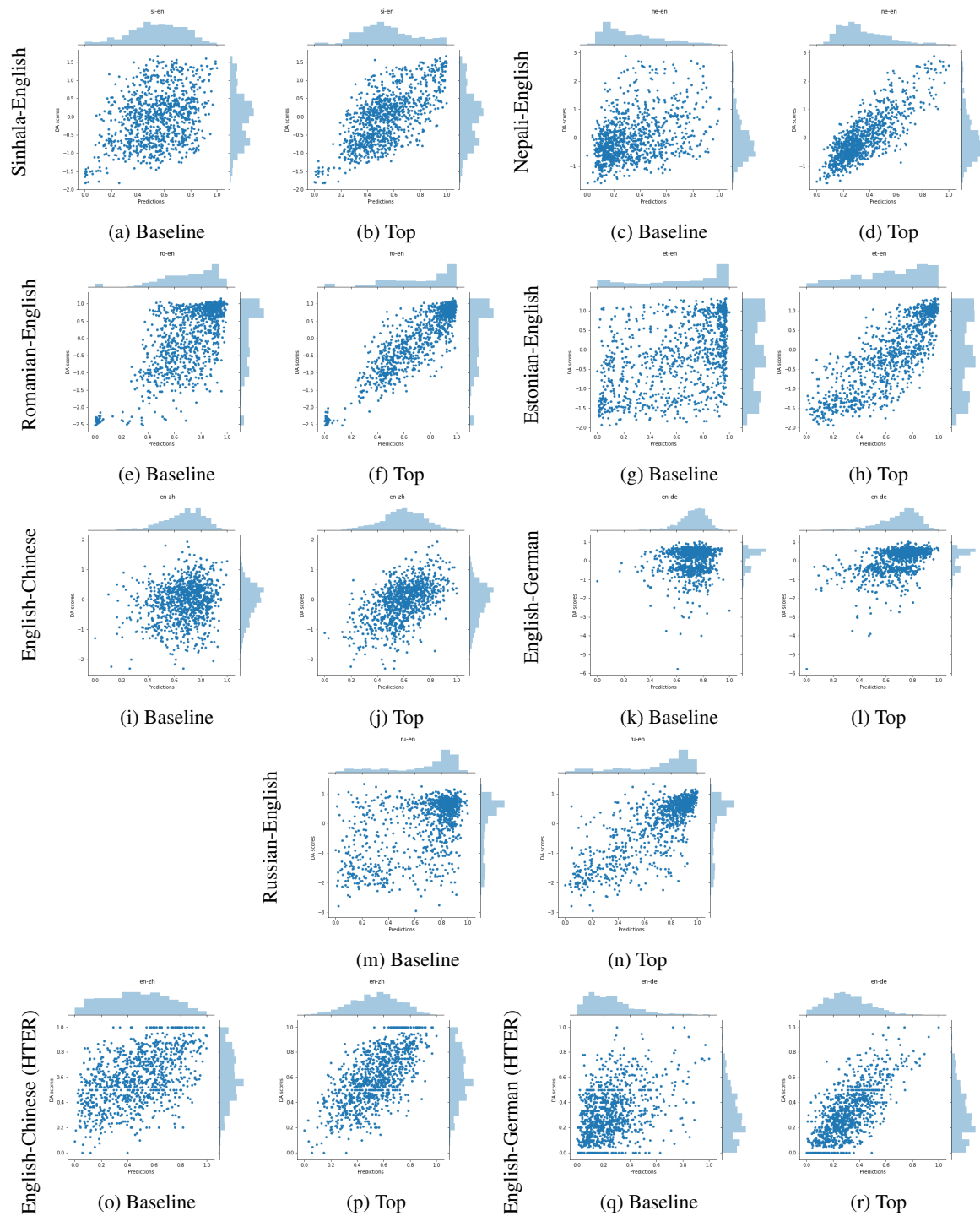


Figure 2: Scatter plots for the predictions against true scores for the baseline and top-performing systems. Sub-figures (a) through (n) show systems trained on direct assessment, while sub-figures (o) through (r) show systems trained on HTER. Predictions are scaled to $[0..1]$.

not have as high an impact on QE as the target language, which has been previously observed as a problem for QE with *partial-input* experiments (Sun et al., 2020); (ii) shared supervision on the target side is beneficial, and thus having more training data into English inherently benefits multiple languages; or (iii) the distribution of scores is more balanced for low- and medium-resource languages, which makes the task easier. To shed more light on this, for future shared tasks, we recommend having more low resource languages as the target language for QE, or more data with shared languages on the source.

High correlations Finally, for some language pairs the performance of the top system is very high, particularly for Romanian–English (Pearson $r = 0.91$). As shown in Figure 2f, there is a number of very low-quality sentences that the QE systems are able to successfully detect. By inspecting those cases, we find that they correspond to ‘hallucinated’ outputs from the Romanian–English MT system 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. This highlights a possible issue with using Pearson correlation to evaluate the performance of QE systems: strong correlations can be achieved by having an over-representation extreme values (i.e. really bad or really good translations), and bad correlations can be an artefact of the lack of representation of extreme values (as in the case of English–German).

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. Statistical significance on Pearson r was computed using the William’s test.

Table 6 summarises the results for English–German and English–Chinese tracks, ranking systems by their average performance for the two language pairs (with 0 as Pearson score for languages without systems). In the Appendix, Tables 19 and 20 show the detailed evaluation results for the two language pairs, ranking participating systems best to worst using Pearson’s r correlation as primary key.

For English–German, the two top performing systems, HW-TSC and Bering Lab, are substan-

tially ahead of the other participants’ systems, with a considerable advantage for HW-TSC, which is the top system with statistical significance. For English–Chinese, Tencent and IST/Unbabel glass-box system were the top performing systems and neither outperforms the other; for this language pair, the range of Pearson scores achieved by participants’ systems is much narrower than for English–German. Finally, for both language pairs, we see that all submissions outperform the baseline system by a large margin, most prominently for English–German.

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 2019 edition, we evaluate separately the source and target side, with the latter including predictions on actual target words as well as gaps. The word-level results for Task 2 are summarised in Tables 7 and 8, ordered by the MCC metric on target errors.

The number of submissions per language pair was different, which limits any conclusions that can be made with respect to general rankings of systems. For English–German, the findings are similar to the sentence-level task: the Bering Lab and HW-TSC teams are the top performing systems by a great margin, with the former better at predicting source side errors and the latter slightly better at predicting target side errors. For English–Chinese, the range of scores is narrower, with HW-TSC, NICT Kyoto, and IST/Unbabel all performing very closely (with HW-TSC on top). For both language pairs, all systems performed above the baseline, and we also see that the scores for the source side are substantially lower than the target side.

5.3 Task 3

MQM score estimation For the document-level estimation task, submissions are evaluated in terms of Pearson’s correlation r , as in Tasks 1 and 2, between the true and predicted document-level scores. Participants results are shown in Table 9. This task attracted fewer participants than the other two, probably because it is more complex. Papago has the best results, with a considerable gap to the IST/Unbabel, which in turn also were well ahead of the baseline.

Fine-grained annotations Fine-grained annotations are evaluated as follows. For each error anno-

Model	En-De	En-Zh
IST and Unbabel (Kiwi-glass-box)	0.633	0.651
NJUNLP	0.618	0.642
NICT Kyoto	0.615	0.643
Bergamot	0.613	0.613
IST and Unbabel (OpenKiwi-base)	0.531	0.593
TransQuest	0.499	0.612
BASELINE	0.392	0.506
HW-TSC	0.758	-
Bering Lab	0.723	-
Tencent Inc.	-	0.664
niuniuniu †	-	0.569
aj54 †	-	0.552
zhanghuimeng †	0.494	-
DexinWang †	0.402	-
NiuTrans †	0.649	0.675

Table 6: Pearson correlation with direct assessments for the submissions to WMT20 Quality Estimation Task 2. 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 teams that have been identified as having submitted more systems than the allowed limit to the leaderboard; ‡ indicates Codalab username of participants from whom we have not received further information.

Model	Target Side			Source Side		
	MCC	F ₁ -BAD	F ₁ -OK	MCC	F ₁ -BAD	F ₁ -OK
Bering Lab	0.597	0.662	0.935	0.454	0.609	0.818
HW-TSC	0.583	0.644	0.938	0.523	0.649	0.875
NICT Kyoto	0.485	0.568	0.916	0.353	0.537	0.806
IST and Unbabel (Kiwi-glass-box)	0.465	0.550	0.916	0.349	0.535	0.801
NJUNLP	0.451	0.498	0.929	-	-	-
IST and Unbabel (OpenKiwi-base)	0.432	0.522	0.909	0.324	0.516	0.799
Elturco.AI	0.423	0.520	0.887	-	-	-
BASELINE	0.358	0.468	0.879	0.266	0.477	0.779
NuiTrans †	0.500	0.581	0.916	0.347	0.532	0.806

Table 7: Official results of the WMT20 Quality Estimation Task 2 word-level for the **English-German** dataset. Baseline systems are highlighted in grey; † indicates teams have been identified as having submitted more systems than the allowed limit to the leaderboard.

Model	Target Side			Source Side		
	MCC	F ₁ -BAD	F ₁ -OK	MCC	F ₁ -BAD	F ₁ -OK
HW-TSC	0.587	0.714	0.866	-	-	-
NICT Kyoto	0.582	0.704	0.878	0.336	0.668	0.669
IST and Unbabel (OpenKiwi-base)	0.575	0.706	0.850	0.287	0.705	0.410
IST and Unbabel (Kiwi-glass-box)	0.567	0.701	0.842	0.287	0.705	0.403
NJUNLP	0.551	0.672	0.877	-	-	-
BASELINE	0.509	0.658	0.849	0.270	0.682	0.547
NuiTrans †	0.610	0.723	0.887	0.308	0.666	0.639

Table 8: Official results of the WMT20 Quality Estimation Task 2 word-level for the **English-Chinese** dataset. Baseline systems are highlighted in grey; † indicates teams have been identified as having submitted more systems than the allowed limit to the leaderboard.

tation a_i^s in the system output, we look for the gold annotation a_j^g with the highest overlap in number of characters. The precision of a_i^s is defined by the ratio of the overlap size to the annotation length; or 0 if there was no overlapping gold annotation. Conversely, we compute the recall of each gold annotation a_j^g considering the best matching annotation a_k^s in the system output,⁹ or 0 if there was no overlapping annotation. The document precision and recall are computed as the average of all annotation precision in the corresponding system output and recalls in the gold output; and therewith we compute the document F_1 . The final score is the unweighted average of the F_1 for all documents.

The annotation scores are shown in Table 10. Only one participant, IST/Unbabel submitted valid results, but still better than the baseline.

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. Overall, participating systems achieved very promising results, 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. Like in other NLP tasks, for QE it had already been shown to substantially improve the results over models that do not use such representations, with heavier pre-trained embeddings contributing substantially more (Kepler et al., 2019a). Strong pre-trained embeddings such as XLM-R were used by most submissions this year.

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. Reporting the number of parameters could be a good practice for the future.

Comparison to previous years submissions are not possible as they use very different datasets, except for Task 3, where a new test set was collected from the same initial larger dataset, but the training data is virtually the same. For the fine-grained

⁹Notice that if a gold annotation a_j^g has the highest overlap with a system annotation a_i^s , it does not necessarily mean that a_i^s has the highest overlap with a_j^g .

version of the task, results are on par with last year (0.48 F_1), while for the scoring variant the results this year are more encouraging: while the baseline remains similar (Pearson = 0.39 this year and 0.35 last year), the top system is significantly better this year: 0.57 Pearson instead of 0.37 last year.

Unfortunately, the document-level task still attracts very few participants, being naturally more difficult to model. However, document-level translation quality is a growing concern in the MT community, and we believe it is interesting that this task continues to exist, possibly with a different dataset and format, in the next editions.

Comparison between HTER and DA. Compared to the results from the previous editions of this shared task, participating systems show overall higher correlation with DA labels. Besides the QE systems getting much stronger, DA labels might be easier to predict, as HTER is a semi-automatic metric and may suffer from the same issues as TER, as it does not capture to what extent the overall quality of the sentence is affected by MT errors. We should note, however, that for the language pairs selected for post-editing this year (English–German and English–Chinese) the correlation is higher for HTER. A possible reason is a very skewed output distribution of the DA scores for these particular language pairs.

HTER and DA annotation capture different aspects of translation quality. In fact, as shown in Fomicheva et al. (2020a), the correlation between the two types of scores is fairly low. An interesting question is whether the approaches that perform best for predicting DA also achieve the best results for HTER. Figure 3 plots sentence-level Pearson correlation with HTER and direct assessments for the systems that participated in both tasks. While the systems with the highest and the lowest ranks are the same, results change considerably for the systems in the middle. Specifically, TransQuest is one of the winning submissions for the prediction of DA, but is outperformed by the submissions that use glass-box features, i.e. Bergamot and IST and Unbabel (Kiwi-glass-box) for the HTER task.

Multilingual approaches. Most of the participating approaches rely on pre-trained multilingual representations and use the provided data annotated with quality labels for fine-tuning. This shows the potential for multilingual prediction in these systems making them much more appealing in prac-

Model	Pearson r	MAE	RMSE
Papago	0.573	15.611	23.327
IST and Unbabel (Kiwi-doc-iob)	0.475	17.127	25.530
BASELINE	0.389	19.939	26.608
NiuTrans †	0.494	20.607	24.258

Table 9: Official results of the WMT20 Quality Estimation Task 3 scoring for the **English–French** dataset. Baseline systems are highlighted in grey; † indicates teams have been identified as having submitted more systems than the allowed limit to the leaderboard.

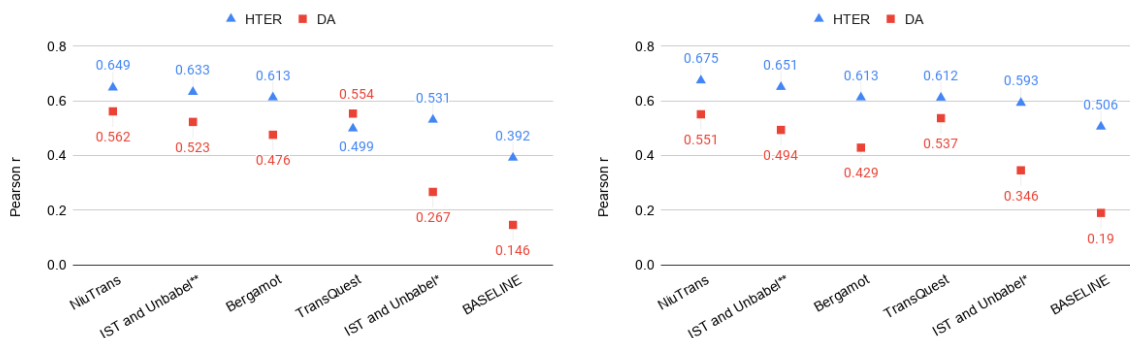


Figure 3: Pearson correlation for the systems that participated in both Task 1 and Task 2 at sentence level for English-German (left) and English-Chinese (right). OpenKiwi-base and Kiwi-glass-box submissions are marked with * and ** respectively.

Model	F1
IST and Unbabel (Kiwi-doc)	0.472
BASELINE	0.416
NiuTrans †	0.418

Table 10: Official results of the WMT20 Quality Estimation Task 3 annotation for the **English–French** dataset. Baseline systems are highlighted in grey; † indicates teams have been identified as having submitted more systems than the allowed limit to the leaderboard.

tice where having dedicated systems for each language pair may be infeasible. However, in the task most submissions built models specific to each language pair, and then submitted their predictions to the multilingual task. A notable exception is the Bergamot-LATTE team, where a single prediction model was trained for all languages.

Influence of source-language document-level context. To investigate the utility of document-level information, we offered to participants the title of the Wikipedia article where the sentences were extracted for Tasks 1 and 2. However, no participating system requested these additional labels, and therefore this remains an open question.

Applicability of NMT model information.

Multiple submissions use glass-box features based on the information extracted from the NMT system in an unsupervised manner (Bergamot-LATTE), in a regression setting (Bergamot) or in combination with pre-trained representations (IST and Unbabel). Results show the potential of this approach. Although substantially outperformed by the top submissions that use pre-trained representations trained with very large amounts of data, glass-box approaches beat the baseline, which use the same amount of training data as the NMT system, by a large margin. These approaches might offer a better trade-off between accuracy and efficiency for cases where the NMT model is accessible.

New publicly available benchmarks. Creating the multi-language, multi-label dataset for this year’s edition was a significant joint effort from various institutions, and we hope it will be useful for researchers in QE as well as in related areas. For example, Task 2 data was also used for the WMT20 Automatic Post-Editing task. We hope to continue adding data to this collection following the same principles, and that others will also contribute by adding other languages to it in the future. We made all submissions to the task available for

those interested in further analysing the results, investigating approaches for prediction ensembling, among others.

7 Conclusions

This year’s edition of the QE Shared Task introduced a number of new elements: the largest number of languages ever, new types of annotation (direct assessment, in addition to labels derived from post-editing and manual error tagging), and number of samples annotated overall. It also attracted the largest number of teams and submissions. We believe the current set of tasks covers a broad enough range of challenges that are far from solved, such as improving performance for languages with skewed distributions, addressing low resource languages, predicting source words that lead to errors, multilingual or language-independent models, etc. In future editions, we hope to keep pushing for progress in these areas.

Acknowledgments

Marina Fomicheva, Frédéric Blain and Lucia Specia were supported by funding from the Bergamot project (EU H2020 Grant No. 825303). André Martins and Erick Fonseca 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 Camila Pohlmann and the Unbabel community team for monitoring the post-editing process. We thank IQT Labs for providing the Russian-English dataset for Task 1.

References

Yujin Baek, Zae Myung Kim, Jihyung Moon, Hyunjoong Kim, and Eunjeong Park. 2020. Patquest: Page translation quality estimation. In *Proceedings of the Fifth Conference on Machine Translation, Volume 2: Shared Tasks Papers*, Online. Association for Computational Linguistics.

Ergun Biçici. 2020. Rtm ensemble learning results at quality estimation task. In *Proceedings of the Fifth Conference on Machine Translation, Volume 2: Shared Tasks Papers*, Online. Association for Computational Linguistics.

Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems 32*, pages 7059–7069.

Qu Cui, Xiang Geng, Shujian Huang, and Jiajun Chen. 2020. Nju’s submission for wmt2020 qe shared task. In *Proceedings of the Fifth Conference on Machine Translation, Volume 2: Shared Tasks Papers*, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Chris Dyer, Victor Chahuneau, and Noah A Smith. 2013. A simple, fast, and effective reparameterization of ibm model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648.

Kai Fan, Jiayi Wang, Bo Li, Fengming Zhou, Boxing Chen, and Luo Si. 2019. “bilingual expert” can find translation errors. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6367–6374.

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

Marina Fomicheva, Shuo Sun, Lisa Yankovskaya, Frédéric Blain, Vishrav Chaudhary, Mark Fishel, Francisco Guzmán, and Lucia Specia. 2020b. Bergamot-latte submissions for the wmt20 quality estimation shared task. In *Proceedings of the Fifth Conference on Machine Translation, Volume 2: Shared Tasks Papers*, Online. Association for Computational Linguistics.

Marina Fomicheva, Shuo Sun, Lisa Yankovskaya, Frédéric Blain, Francisco Guzmán, Mark Fishel, Nikolaos Aletras, Vishrav Chaudhary, and Lucia Specia. 2020c. Unsupervised quality estimation for neural machine translation. *Transactions of the Association for Computational Linguistics*.

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

- Francisco Guzmán, Peng-Jen Chen, Myle Ott, Juan Pino, Guillaume Lample, Philipp Koehn, Vishrav Chaudhary, and Marc’Aurelio Ranzato. 2019. The FLORES evaluation datasets for low-resource machine translation: Nepali–English and Sinhala–English. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6098–6111, Hong Kong, China. Association for Computational Linguistics.
- Ruining He and Julian McAuley. 2016. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In *proceedings of the 25th international conference on world wide web*, pages 507–517. International World Wide Web Conferences Steering Committee.
- Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. *arXiv preprint arXiv:1902.00751*.
- Chi Hu, Hui Liu, Kai Feng, Chen Xu, Nuo Xu, Zefan Zhou, Shiqin Yan, Yingfeng Luo, Chenglong Wang, Xia Meng, Tong Xiao, and Jingbo Zhu. 2020. The niutrans system for the wmt20 quality estimation shared task. In *Proceedings of the Fifth Conference on Machine Translation, Volume 2: Shared Tasks Papers*, Online. Association for Computational Linguistics.
- Hassan Kane, Muhammed Yusuf Kocyigit, Ali Abdalla, Pelkins Ajano, and Mohamed Coulibali. 2020. NUBIA: Neural based interchangeability assessor for text generation.
- Fábio Kepler, Jonay Trénous, Marcos Treviso, Miguel Vera, António Góis, M Amin Farajian, António V Lopes, and André FT Martins. 2019a. Unbabel’s participation in the wmt19 translation quality estimation shared task. *arXiv preprint arXiv:1907.10352*.
- Fábio Kepler, Jonay Trénous, Marcos Treviso, Miguel Vera, and André F. T. Martins. 2019b. 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.
- Dongjun Lee. 2020. Two-phase cross-lingual language model fine-tuning for machine translation quality estimation. In *Proceedings of the Fifth Conference on Machine Translation, Volume 2: 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.
- Julian McAuley, Christopher Targett, Qinfeng Shi, and Anton Van Den Hengel. 2015. Image-based recommendations on styles and substitutes. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 43–52. ACM.
- João Moura, Miguel Vera, Daan van Stigt, Fabio Kepler, and André F. T. Martins. 2020. Ist-unbabel participation in the wmt20 quality estimation shared task. In *Proceedings of the Fifth Conference on Machine Translation, Volume 2: Shared Tasks Papers*, Online. Association for Computational Linguistics.
- Akifumi Nakamachi, Hiroki Shimanaka, Tomoyuki Kajiwara, and Mamoru Komachi. 2020. Tmuou submission for wmt20 quality estimation shared task. In *Proceedings of the Fifth Conference on Machine Translation, Volume 2: Shared Tasks Papers*, Online. Association for Computational Linguistics.
- Nathan Ng, Kyra Yee, Alexei Baevski, Myle Ott, Michael Auli, and Sergey Edunov. 2019. Facebook fair wmt19 news translation task submission. In *Proc. of WMT*, pages 1–4.
- 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 at wmt2020: Sentence-level direct assessment. In *Proceedings of the Fifth Conference on Machine Translation, Volume 2: Shared Tasks Papers*, Online. Association for Computational Linguistics.
- Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. 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 3982–3992, Hong Kong, China. Association for Computational Linguistics.
- Raphael Rubino. 2020. Nict kyoto submission for the wmt’20 quality estimation task: Intermediate training for domain and task adaptation. In *Proceedings of the Fifth Conference on Machine Translation, Volume 2: Shared Tasks Papers*, Online. Association for Computational Linguistics.
- Marina Sanchez-Torron and Philipp Koehn. 2016. Machine translation quality and post-editor productivity. *AMTA 2016, Vol.*, page 16.
- Lucia Specia, Gustavo Paetzold, and Carolina Scarton. 2015. Multi-level translation quality prediction with quest++. In *ACL-IJCNLP 2015 System Demonstrations*, pages 115–120, Beijing, China.

- Shuo Sun, Francisco Guzmán, and Lucia Specia. 2020. Are we estimating or guesstimating translation quality? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6262–6267, Online. Association for Computational Linguistics.
- Minghan Wang, Hao Yang, Hengchao Shang, Daimeng Wei, Jiaxin Guo, Lizhi Lei, Ying Qin, Shimin Tao, Shiliang Sun, Yimeng Chen, and Liangyou Li. 2020a. Hw-tsc’s participation at wmt 2020 quality estimation shared task. In *Proceedings of the Fifth Conference on Machine Translation, Volume 2: Shared Tasks Papers*, Online. Association for Computational Linguistics.
- Zixuan Wang, Haijiang Wu, Xiaoli Wang, Xinjie Wen, Ruichen Wang, and Qingsong Ma. 2020b. Tencent submission for wmt20 quality estimation shared task. In *Proceedings of the Fifth Conference on Machine Translation, Volume 2: Shared Tasks Papers*, Online. Association for Computational Linguistics.
- Evan J. Williams. 1959. *Regression Analysis*, volume 14. Wiley, New York, USA.
- Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.
- Lei Zhou, Liang Ding, and Koichi Takeda. 2020. Zero-shot translation quality estimation with explicit cross-lingual patterns. In *Proceedings of the Fifth Conference on Machine Translation, Volume 2: Shared Tasks Papers*, Online. Association for Computational Linguistics.

A Official Results of the WMT20 Quality Estimation Task 1

Tables 11, 12, 13, 14, 15, 16, 17 and 18 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	Pearson r	MAE	RMSE
TransQuest	0.722	0.480	0.596
Bergamot-LATTE (black-box)	0.718	0.408	0.527
TMUOU	0.686	0.418	0.543
IST and Unbabel (Kiwi-glass-box-ensemble)	0.673	0.433	0.569
IST and Unbabel (OpenKiwi-base)	0.583	0.547	0.719
WL Research	0.546	0.538	0.683
Bergamot-LATTE (glass-box)	0.489	0.895	1.062
jackielo ‡	0.462	0.918	1.141
BASELINE	0.376	0.788	0.999
NiuTrans †	0.732	0.529	0.653

Table 11: Official results of the WMT20 Quality Estimation Task 1 for the **Multilingual** variant. Baseline systems are highlighted in grey; † indicates teams that have been identified as having submitted more systems than the allowed limit to the leaderboard; ‡ indicates Codalab usernames of participants from whom we have not received further information.

Model	Pearson r	MAE	RMSE
• TransQuest	0.554	0.613	0.740
• Bergamot-LATTE (black-box)	0.544	0.451	0.616
IST and Unbabel (Kiwi-glass-box-ensemble)	0.523	0.470	0.635
Papago	0.498	0.454	0.637
TMUOU	0.482	0.455	0.625
Bergamot	0.476	0.483	0.636
XC	0.465	0.739	0.861
zhanghuimeng ‡	0.392	0.715	0.964
IST and Unbabel (OpenKiwi-base)	0.267	0.525	0.683
Bergamot-LATTE (glass-box)	0.259	0.819	0.940
WL Research	0.253	0.527	0.683
DexinWang ‡	0.246	0.503	0.680
Hancheng_Deng ‡	0.171	0.490	0.726
BASELINE	0.146	0.679	0.967
FVCRC	0.111	0.805	1.063
NiuTrans †	0.562	0.558	0.676

Table 12: Official results of the WMT20 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; † indicates teams that have been identified as having submitted more systems than the allowed limit to the leaderboard; ‡ indicates Codalab usernames of participants from whom we have not received further information.

Model	Pearson r	MAE	RMSE
• TransQuest	0.537	0.675	0.831
Bergamot-LATTE (black-box)	0.530	0.452	0.587
IST and Unbabel (Kiwi-glass-box-ensemble)	0.494	0.459	0.592
XC	0.465	0.782	0.944
aj54 ‡	0.444	1.020	1.170
TMUOU	0.438	0.585	0.739
Bergamot	0.429	0.467	0.612
JXNU-CCLQ	0.426	0.709	0.890
IST and Unbabel (OpenKiwi-base)	0.346	0.518	0.684
Bergamot-LATTE (glass-box)	0.321	1.094	1.228
WL Research	0.298	0.796	0.970
RTM	0.259	68.010	68.414
BASELINE	0.190	0.885	1.068
FVCRC	0.085	0.873	1.059
NiuTrans †	0.551	0.499	0.654

Table 13: Official results of the WMT20 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; † indicates teams that have been identified as having submitted more systems than the allowed limit to the leaderboard; ‡ indicates Codalab usernames of participants from whom we have not received further information.

Model	Pearson r	MAE	RMSE
• TransQuest	0.908	0.300	0.392
• Bergamot-LATTE (black-box)	0.906	0.281	0.388
TMUOU	0.896	0.294	0.414
IST and Unbabel (Kiwi-glass-box-ensemble)	0.891	0.398	0.530
XC	0.882	0.556	0.661
Shrangin ‡	0.846	0.727	1.009
WL Research	0.821	0.393	0.520
Bergamot	0.796	0.438	0.554
IST and Unbabel (OpenKiwi-base)	0.708	0.508	0.655
RTM	0.703	0.517	0.654
Bergamot-LATTE (glass-box)	0.693	0.994	1.132
BASELINE	0.685	0.760	1.052
FVCRC	0.650	0.840	1.174
NiuTrans †	0.917	0.583	0.691

Table 14: Official results of the WMT20 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; † indicates teams that have been identified as having submitted more systems than the allowed limit to the leaderboard; ‡ indicates Codalab usernames of participants from whom we have not received further information.

Model	Pearson r	MAE	RMSE
• Bergamot-LATTE (black-box)	0.826	0.427	0.540
• TransQuest	0.824	0.485	0.604
TMUOU	0.792	0.493	0.636
IST and Unbabel (Kiwi-glass-box-ensemble)	0.770	0.740	0.919
XC	0.764	0.745	0.906
IST and Unbabel (OpenKiwi-base)	0.690	0.531	0.652
Bergamot	0.681	0.565	0.682
Bergamot-LATTE (glass-box)	0.642	0.918	1.096
WL Research	0.637	0.567	0.714
RTM	0.614	66.362	67.656
BASELINE	0.477	0.918	1.138
NiuTrans †	0.833	0.561	0.716

Table 15: Official results of the WMT20 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; † indicates teams that have been identified as having submitted more systems than the allowed limit to the leaderboard.

Model	Pearson r	MAE	RMSE
• TransQuest	0.822	0.372	0.474
Bergamot-LATTE (black-box)	0.814	0.368	0.475
IST and Unbabel (Kiwi-glass-box-ensemble)	0.792	0.433	0.549
TMUOU	0.785	0.397	0.511
XC	0.778	1.414	1.512
WL Research	0.687	0.452	0.594
Bergamot	0.662	0.486	0.612
IST and Unbabel (OpenKiwi-base)	0.604	0.497	0.648
Bergamot-LATTE (glass-box)	0.600	0.727	0.854
FVCRC	0.488	0.918	1.046
BASELINE	0.386	0.735	0.871
NiuTrans †	0.830	0.481	0.629

Table 16: Official results of the WMT20 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; † indicates teams that have been identified as having submitted more systems than the allowed limit to the leaderboard.

Model	Pearson r	MAE	RMSE
• TransQuest	0.685	0.436	0.534
• Bergamot-LATTE (black-box)	0.682	0.429	0.539
TMUOU	0.668	0.459	0.572
IST and Unbabel (Kiwi-glass-box-ensemble)	0.639	0.506	0.642
XC	0.626	0.879	1.021
WL Research	0.577	0.492	0.614
IST and Unbabel (OpenKiwi-base)	0.565	0.515	0.634
Bergamot	0.560	0.490	0.602
RTM	0.541	49.675	50.774
Bergamot-LATTE (glass-box)	0.513	0.673	0.819
FVCRC	0.388	0.694	0.848
BASELINE	0.374	0.752	0.898
NiuTrans †	0.698	0.445	0.543

Table 17: Official results of the WMT20 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, and † indicates teams have been identified as having submitted more systems than the allowed limit to the leaderboard.

Model	Pearson r	MAE	RMSE
• TransQuest	0.808	0.402	0.583
Bergamot-LATTE (black-box)	0.796	0.412	0.584
XC	0.784	0.603	0.759
TMUOU	0.781	0.433	0.622
IST and Unbabel (Kiwi-glass-box-ensemble)	0.767	0.428	0.613
WL Research	0.596	0.575	0.763
BASELINE	0.548	0.825	1.193
Mak	0.543	0.590	0.811
jackielo ‡	0.411	0.878	1.267
FVCRC	0.400	0.831	1.220
NiuTrans †	0.816	0.535	0.687

Table 18: Official results of the WMT20 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; † indicates teams that have been identified as having submitted more systems than the allowed limit to the leaderboard; ‡ indicates Codalab usernames of participants from whom we have not received further information.

B Official Results of the WMT20 Quality Estimation Task 2 (Sentence-level)

Tables 19 and 20 show the evaluation results for English-German and English-Chinese respectively, ranking participating systems best to worst using Pearson’s r correlation as primary key for each language pair.

Model	Pearson r	MAE	RMSE
• HW-TSC	0.758	0.099	0.133
Bering Lab	0.723	0.107	0.140
IST and Unbabel (Kiwi-glass-box)	0.633	0.137	0.178
NJUNLP	0.618	0.129	0.160
NICT Kyoto	0.615	0.151	0.197
Bergamot	0.613	0.130	0.160
IST and Unbabel (OpenKiwi-base)	0.531	0.138	0.180
TransQuest	0.499	0.149	0.184
zhanghuimeng ‡	0.494	0.163	0.198
DexinWang ‡	0.402	0.155	0.196
BASELINE	0.392	0.150	0.190
NiuTrans †	0.649	0.123	0.154

Table 19: Official results of the WMT20 Quality Estimation Task 2 sentence-level for the **English-German** dataset. Teams marked with “•” are the winners, as they are not significantly outperformed by any other system according to the Williams Significance Test (Williams, 1959). Baseline systems are highlighted in grey; † indicates teams that have been identified as having submitted more systems than the allowed limit to the leaderboard; ‡ indicates Codalab usernames of participants from whom we have not received further information.

Model	Pearson r	MAE	RMSE
• Tencent Inc.	0.664	0.129	0.160
• IST and Unbabel (Kiwi-glass-box)	0.651	0.135	0.171
NICT Kyoto	0.643	0.129	0.161
NJUNLP	0.642	0.129	0.161
Bergamot	0.613	0.136	0.169
TransQuest	0.612	0.135	0.168
IST and Unbabel (OpenKiwi-base)	0.593	0.143	0.175
niuniuniu ‡	0.569	0.142	0.177
aj54 ‡	0.552	0.145	0.176
BASELINE	0.506	0.147	0.181
NiuTrans †	0.675	0.125	0.156

Table 20: Official results of the WMT20 Quality Estimation Task 2 sentence-level 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; † indicates teams that have been identified as having submitted more systems than the allowed limit to the leaderboard; ‡ indicates Codalab usernames of participants from whom we have not received further information.