Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 4963–4974
Marseille, 20-25 June 2022
© European Language Resources Association (ELRA), licensed under CC-BY-NC-4.0

MLQE-PE: A Multilingual Quality Estimation and Post-Editing Dataset

**Marina Fomicheva,[1*] Shuo Sun,[2*] Erick Fonseca,[3†] Chrysoula Zerva,[3]
Frédéric Blain,[1,4] Vishrav Chaudhary,[9§] Francisco Guzmán,[5] Nina Lopatina,[6‡]
André F. T. Martins,[3,8] Lucia Specia[1,7]**

[1]University of Sheffield, [2]Johns Hopkins University, [3]Instituto de Telecomunicações,
[4]University of Wolverhampton, [5]Meta AI, [6]IQT Labs, [7]Imperial College London, [8]Unbabel, [9]Microsoft
{m.fomicheva,l.specia}@sheffield.ac.uk, ssun32@jhu.edu, nina.lopatina@getspectrum.io
erick.fonseca@lx.it.pt, f.blain@wlv.ac.uk, vchaudhary@microsoft.com, fguzman@fb.com,
chrysoula.zerva@tecnico.ulisboa.pt, andre.martins@unbabel.com

Abstract

We present MLQE-PE, a new dataset for Machine Translation (MT) Quality Estimation (QE) and Automatic Post-Editing (APE). The dataset contains annotations for eleven language pairs, including both high- and low-resource languages. Specifically, it is annotated for translation quality with human labels for up to 10,000 translations per language pair in the following formats: sentence-level direct assessments and post-editing effort, and word-level binary good/bad labels. Apart from the quality-related scores, each source-translation sentence pair is accompanied by the corresponding post-edited sentence, as well as titles of the articles where the sentences were extracted from, and information on the neural MT models used to translate the text. We provide a thorough description of the data collection and annotation process as well as an analysis of the annotation distribution for each language pair. We also report the performance of baseline systems trained on the MLQE-PE dataset. The dataset is freely available and has already been used for several WMT shared tasks.

Keywords: Machine Translation, quality estimation, evaluation, direct assessments, post-edits

1. Introduction

Translation quality estimation (QE) is the task of evaluating a translation system's quality without access to reference translations (Blatz et al., 2004; Specia et al., 2018b). This task has numerous applications: deciding if a sentence or document that has been automatically translated is ready to be sent to the final user or if it needs to be post-edited by a human, flagging passages with potentially critical mistakes, using it as a metric for translation quality when a human reference is not available, or in the context of computer-aided translation interfaces, highlighting text that needs human revision and estimating the required human effort.

Due to its high relevance, QE has been the subject of evaluation campaigns in the Conference for Machine Translation (WMT) since 2014 (Bojar et al., 2014; Specia et al., 2018a; Fonseca et al., 2019; Specia et al., 2020), where datasets in various language pairs have been created containing source sentences, their automatic translations, and human post-edited text. However, the currently existing data has several shortcomings. First, the MT system used to produce the translations is not publicly available, which makes it impossible to develop the so-called glass-box approaches to QE and exploit model confidence (or conversely, uncertainty) of the MT system or look into its internal states. Second, the quality assessments have been either produced based on the difference between the MT output and the post-edited text (e.g., through the human translation error rate metric, HTER, or by marking individual words with OK or BAD labels), or by direct human assessments, but not both—which raises the question of how much these two quality assessments correlate. Third, most datasets have focused exclusively on high-resource language pairs, where it is often the case that many sentences are correctly translated; however, medium and low-resource settings are the ones where QE would be particularly useful, since it is where MT currently presents serious challenges. Finally, most of these datasets focus on a specific domain, such as IT or life sciences, where translations are generated by a domain-specific MT model, which also tends to result in most sentences being translated with high-quality.

To overcome the limitations stated above, we introduce MLQE-PE, the first multilingual quality estimation and post-editing dataset combining the following features:

- It includes access to the state-of-the-art neural MT (NMT) models built with an open-source toolkit (`fairseq`; Ott et al. (2019)), that were used to produce the translations in the dataset. This opens the door to uncertainty-based and glass-box approaches to QE. Moreover, it provides multiple independent annotator scores per segment, to motivate methods using aleatoric uncertainty.
- It combines both direct assessments (DA) of MT quality and post-edits. This allows combining two sorts of quality assessments: how good a transla-

[*]Equal contribution.

[†]Author's affiliation at the time of publication is Kaufland e-commerce. Work was done while the author was at Instituto de Telecomunicações.

[‡]Author's affiliation at the time of publication is Spectrum Labs. Work was done while the author was at IQT Labs.

[§]Work done while author was at Meta AI.

tion is and how much effort is necessary to correct it. Moreover, the post-edited sentences can be used for training and evaluating automatic post-editing systems, another important task considered in WMT campaigns (Chatterjee et al., 2019).

- It contains the titles of the Wikipedia articles where the original sentences were extracted from, thus allowing to take document-level context into account when predicting sentence-level or word-level MT quality.
- It includes 11 language pairs, mixing high-resource language pairs (English-German – En-De and English-Chinese – En-Zh, and Russian-English – Ru-En), medium-resource (Romanian-English – Ro-En, and Estonian-English – Et-En, English-Japanese – En-Ja) and low-resource ones (Nepali-English – Ne-En, Sinhala-English – Si-En, Pashto-English – Ps-En, Khmer-English – Km-En, and English-Czech – En-Cs). We aspire to keep extending the dataset with additional translations and language pairs to better support and inspire further work in the field.

This dataset was created with contributions from different institutions: Facebook, University of Sheffield and Imperial College selected the Wikipedia articles and sentences, built the NMT models, prepared and outsourced data for DA annotation in 10 language pairs (En-De, En-Zh, Ro-En, Et-En, Ne-En, Si-En, Ps-En, Km-En, En-Ja, En-Cs). IQT Labs led the same efforts for collecting and DA-annotating the Ru-En data. Facebook, University of Sheffield and Imperial College also outsourced data for all language pairs except En-De and En-Zh for post-editing, and created reference translations for Et-En. Unbabel and Instituto de Telecomunicações outsourced the post-editing of En-De and En-Zh sentences and prepared the baseline QE models. The current version of MLQE-PE is publicly available at `https://github.com/sheffieldnlp/mlqe-pe`

2. Data Collection and Statistics

We briefly describe the data collection and preparation process. Table 1 presents some statistics about the MLQE-PE dataset. As shown in Table 1, we collected 10K sentences split into train, dev and two test partitions (test20 and test21) for nine language pairs. The test partitions have been released separately as they were used as test sets for two consecutive WMT QE shared tasks, in years 2020 and 2021 respectively. We maintain this distinction in the paper to facilitate comparisons and cross-referencing with results mentioned in publications related to those tasks. In addition, we collected 2K sentences for 4 language pairs, which are meant to be used for testing QE in a zero-shot setting where no training or development data is provided [1].

[1] 1K of these sentences will are kept as a blind test set and will be released as part of the WMT QE 2022 test sets.

2.1. Data collection

For the most part, the dataset is derived from Wikipedia articles (with exception of Russian-English, described below). The source sentences were collected from Wikipedia articles following the sampling process outlined in FLORES (Guzmán et al., 2019). First, we sampled documents from Wikipedia for English, Estonian, Romanian, Sinhala, Nepali, Khmer and Pashto. Second, we selected the top 100 documents containing the largest number of sentences that are: (i) in the intended source language according to a language-id classifier[2] and (ii) have the length between 50 and 150 characters. In addition, we filtered out sentences that have been released as part of recent Wikipedia parallel corpora (Schwenk et al., 2019), ensuring that our dataset is not part of parallel data commonly used for NMT training. For every language, we randomly selected the required number of sentences from the sampled documents and then translated them using SOTA NMT models (see below). For German and Chinese, we followed an additional procedure in order to ensure sufficient representation of high- and low-quality translations for these high-resource language pairs. We selected the sentences with minimal lexical overlap with the NMT training data. Specifically, we extracted content words for each sentence in the data used for training the NMT models and in the Wikipedia data. We then computed perplexity scores for the Wikipedia sentences given the NMT training data, and sampled 20K from available Wikipedia sentences weighted by the perplexity scores. In addition, we collected human reference translations for a 1K subset of Estonian-English dev/test data. Two reference translations were generated independently by two professional translators. This part of the dataset allows for comparing reference-free MT evaluation with reference-based approaches (see Fomicheva et al. (2020a) for details).

The Russian-English data collection followed a slightly different set up collected by collaborators from IQT Labs.[3] The original sentences were collected from multiple sources in order to gather a varied sample of data in different domains that are still challenging for current NMT systems. Data sources include: Russian proverbs and Reddit data from various subreddits, particularly those focused on topics of politics and religion. We included Reddit data since colloquial text is a challenge for MT. We included Russian proverbs from WikiQuotes to test MT on short sentences with unconventional grammar. We used the Reddit API and queried the most recent 1000 posts at the time, and the most recent 1000 comments in each of the selected subreddits. We automatically split the posts into sentences and then reviewed these manually. Markdown was removed and HTML unencoded. We removed sentences shorter than 15 characters or longer than 500 charac-

[2] `https://fasttext.cc`

[3] We note that Facebook was not involved in the collection of the Russian-English data.

Lang.	Sentences					Tokens (approx.)					DA	PE
	Train	Dev	Test20	Test21	Test22	Train	Dev	Test20	Test21	Test22		
En-De	7K	1K	1K	1K	-	115K	17K	16K	17K	-	✓	✓
En-Zh	7K	1K	1K	1K	-	116K	16K	17K	17K	-	✓	✓
Ru-En	7K	1K	1K	1K	-	82K	12K	12K	12K	-	✓	✓
Ro-En	7K	1K	1K	1K	-	120K	17K	17K	17K	-	✓	✓
Et-En	7K	1K	1K	1K	-	98K	14K	14K	14K	-	✓	✓
Ne-En	7K	1K	1K	1K	-	105K	15K	15K	15K	-	✓	✓
Si-En	7K	1K	1K	1K	-	110K	16K	16K	16K	-	✓	✓
Ps-En	-	-	-	1K	1K	-	-	-	27K	26K	✓	✓
Km-En	-	-	-	1K	1K	-	-	-	22K	22K	✓	✓
En-Ja	-	-	-	1K	1K	-	-	-	21K	21K	✓	✓
En-Cs	-	-	-	1K	1K	-	-	-	20K	20K	✓	✓

Table 1: Statistics of the MLQE-PE dataset. The numbers of sentences and tokens are shown for train, development and two test partitions (test20 and test21), respectively for En-De, En-Zh, Ru-En, Ro-En, Et-En, Ne-En and Si-En, and for the test partition for Ps-En, Km-En, En-Ja and En-Cs. The number of tokens refers to the source side.

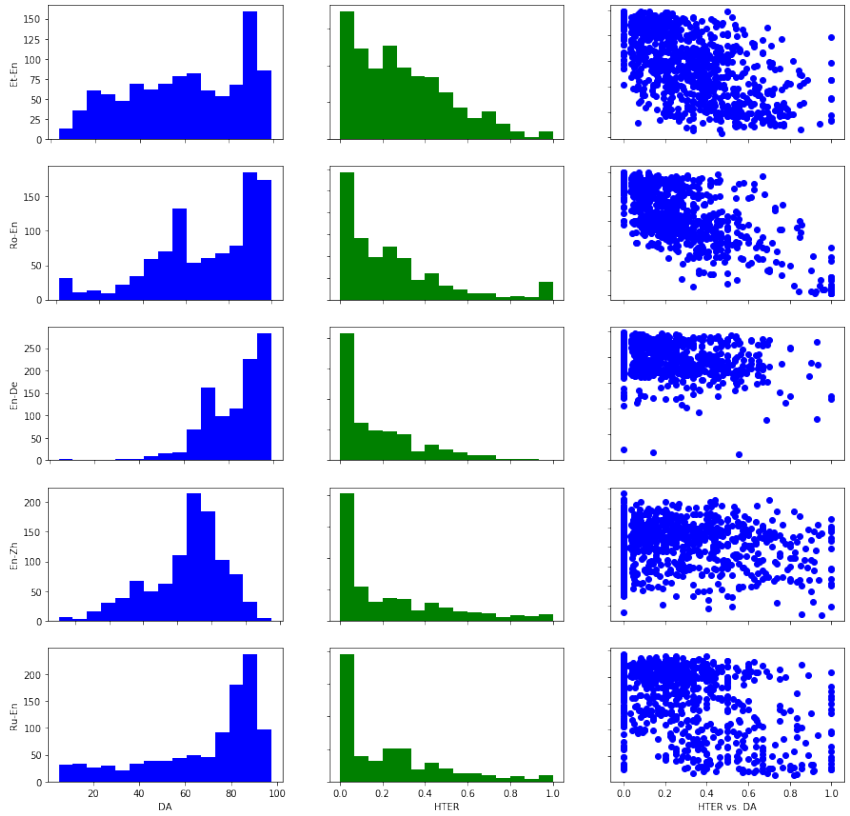

Figure 1: Distribution of direct assessment scores (DA), HTER scores and their scatter plots for the test21 partition of the dataset, for Et-En, Ro-En, En-De, En-Zh and Ru-En language pairs.

ters. We also removed sentences that did not have a source link. Table 2 shows the number of segments corresponding to each data source and the corresponding average direct assessment score.

2.2. NMT models

Transformer-based (Vaswani et al., 2017) NMT models were trained for all languages using the `fairseq`[4] toolkit. For **Et-En**, **Ro-En**, **En-De** and **En-Zh** we trained the MT models based on the standard Transformer architecture following the implementation details described in Ott et al. (2018). We used publicly available MT datasets such as Paracrawl (Esplà et al., 2019) and Europarl (Koehn, 2005). For **Ru-En**, translations were produced with the already existing Transformer-based NMT model described in Ng et al.

[4] `https://github.com/pytorch/fairseq`

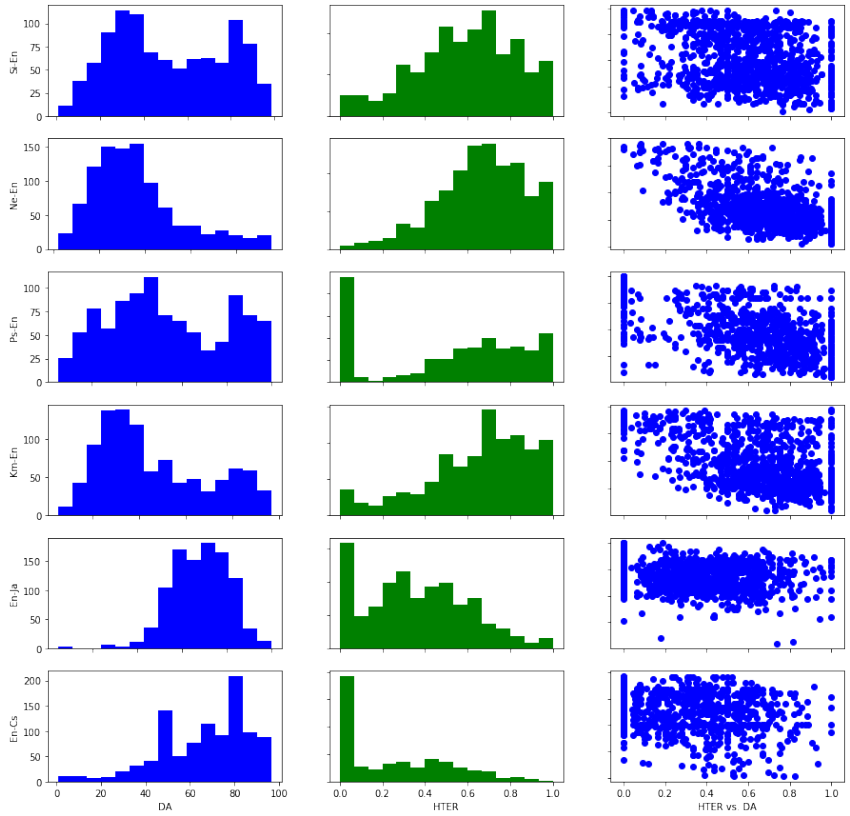

Figure 2: Distribution of direct assessments scores (DA), HTER scores and their scatter plots for the test21 partition of the dataset, for Si-En, Ne-En, Ps-En, Km-En, En-Ja and En-Cs language pairs.

	Count	DA
www.reddit.com/r/antireligious	2,155	75.6
www.reddit.com/r/PikabuPolitics	1,753	77.7
www.reddit.com/r/rupolitika	1,422	80.1
www.reddit.com/r/ru	2,171	74.0
wikiquote.org/wiki	2,499	41.1

Table 2: Number of sentences and average absolute direct assessment (DA) score for each data source in the Ru-En dataset

(2019).[5] **Si-En** and **Ne-En** MT systems were trained based on Big-Transformer architecture as defined in Vaswani et al. (2017). For these low-resource language pairs, the models were trained following the FLORES semi-supervised setting (Guzmán et al., 2019),[6] which involves two iterations of backtranslation using the source and the target monolingual data. For **Ps-En**, **Km-En**, **En-Cs** and **En-Ja** we use multilingual MT models described in Tang et al. (2020).[7]

[5] https://github.com/pytorch/fairseq/tree/master/examples/wmt19

[6] https://github.com/facebookresearch/flores/blob/master/reproduce.sh

[7] Instructions for training and using the NMT models are at https://github.com/pytorch/fairseq/tree/master/examples/multilingual.

The data used for training the NMT models is available from http://www.statmt.org/wmt20/quality-estimation-task.html. We provide access to the information from the NMT model used to generate the translations: model score for the sentence and log probabilities for words, as well as the NMT systems themselves.

2.3. Direct assessments.

To collect human quality judgments, we followed the FLORES setup (Guzmán et al., 2019) inspired by the work of Graham et al. (2013). Specifically, the annotators were asked to rate translation quality for each sentence on a 0–100 scale, where the 0–10 range represents an incorrect translation; 11–29, a translation that contains a 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.

Each segment was evaluated independently by three professional translators from a single language service provider. To improve annotation consistency, any evaluation in which the range of scores among the raters was above 30 points was rejected, and an additional

rater was requested to replace the most diverging translation rating until convergence was achieved. To further increase the reliability of the test and development partitions of the dataset, we requested an additional set of three annotations from a different group of annotators (i.e., from another language service provider) following the same annotation protocol, thus resulting in a total of six annotations per segment.

Raw human scores were converted into z-scores, that is, standardized according to each individual annotator's overall mean and standard deviation. The scores collected for each segment were averaged to obtain the final score. Such setting allows for the fact that annotators may genuinely disagree on some aspects of quality.

2.4. Human post-editing.

For all language pairs, the translated sentences have been post-edited by human translators. For En-De and En-Zh, we used paid editors from the Unbabel community. For all other languages, we used professional translators subcontracted by Facebook. The human translators performing post-editing had no access to the direct assessments scores.

Table 3 shows average translation quality for all language pairs based on direct assessment annotation (DA) and post-editing (HTER) for the test21 partition of the dataset. Figures 1 and 2 show the distribution of the corresponding sentence-level scores, as well as the scatter plot of DA against HTER scores.

First, we note that the distribution of direct assessment scores is very different across language pairs. This illustrates the variety of the collected data in terms of MT output quality. For low-resource language pairs there are more sentences with low direct assessment scores, whereas in the case of high-resource language pairs the vast majority of translations received a high score. In particular, En-De has a very peaked distribution with very little variability in quality.

Second, we note that higher DA score often corresponds to lower translation edit rate in Table 3. Thus, on average direct assessment and post-editing effort produce consistent results as an indication of overall translation quality per language pair. However, sentence-level DA and HTER scores for the same data behave quite differently. Table 4 shows the correlation between direct assessments and HTER scores for all the language pairs on the test21 partition of the dataset. As illustrated in Table 4 and in the scatter plots on Figures 1 and 2 for most of the language pairs there is a weak negative correlation between the two types of quality scores.

Direct quality assessment and post-editing give two different perspectives on MT quality. Table 5 shows an example where direct assessment and HTER lead to a different interpretation of quality. Direct assessment score is low as the MT output contains a serious error that distorts the meaning of the sentence: "bars" (as in "metal bars") is translated as "pub". However the sentence is easy to post-edit as the error involves only one word to be replaced, resulting in a low HTER score. Table 6 illustrates the opposite: MT output was assigned a high direct assessment score, but the HTER score is also high, indicating that substantial changes were introduced during post-editing. The post-edited version is more fluent, whereas the MT output is a more literal rendering of the source sentence, but the meaning is preserved and, therefore, it received a high direct assessment score.

	Average DA ↑	Average HTER ↓
En-De	82.61	0.18
Ro-En	69.18	0.24
En-Ja	67.96	0.36
En-Cs	66.94	0.26
En-Zh	62.86	0.23
Et-En	60.09	0.29
Ps-En	53.53	0.53
Si-En	51.42	0.59
Km-En	46.58	0.65
Ne-En	36.51	0.66
Ru-En	68.67	0.23

Table 3: Average MT quality in terms of DA scores (higher is better) and HTER scores (lower is better) on the test21 partition of the dataset.

	Pearson	Spearman
En-De	-0.42	-0.48
Ro-En	-0.76	-0.71
En-Ja	-0.14	-0.11
En-Cs	-0.41	-0.46
En-Zh	-0.21	-0.16
Et-En	-0.61	-0.63
Ps-En	-0.71	-0.67
Si-En	-0.29	-0.28
Km-En	-0.49	-0.43
Ne-En	-0.54	-0.49
Ru-En	-0.51	-0.47

Table 4: Pearson and Spearman correlation between DA and HTER scores for the test21 partition of the dataset.

2.5. Word-level labels

In the datasets containing post-edit annotation, we also obtained word-level labels for fine-grained post-editing effort estimation. Both the source and MT sides have them.

In order to generate them, we first align source and MT outputs using SimAlign [8]. We follow the findings of Sabet et al. (2020) and use *Argmax* matching for high resource languages that are close to english (En-De, En-Cs) and *Itermax* for the rest of the language pairs.

[8] https://github.com/cisnlp/simalign

Type	Text	Scores
Source	He wakes up in a cage, and enjoys rubbing the rusted bars.	
MT	他在笼子里醒来, 喜欢擦生锈的酒吧.	DA = 33
PE	他在笼子里醒来, 喜欢摩擦生锈的铁条。	HTER = 0.33
MT gloss	He wakes up in a cage, and enjoys rubbing the rusted **pub**.	
PE gloss	He wakes up in a cage, and enjoys rubbing the rusted **metal bar**.	

Table 5: Example of the discrepancy between HTER and DA annotation tasks: low DA score (low quality) but low HTER score (minimal post-editing).

Type	Text	Scores
Source	The two battled to a standstill and eventually rendered one another comatose.	
MT	这两个人的战斗陷入停顿, 最后彼此昏迷不已.	DA = 73
PE	两人对战陷入僵局，最后双双昏倒。	HTER = 1.00
MT gloss	The two people's battle fell into a standstill, finally both were in a coma.	
PE gloss	The two people battled to a standstill and both fell into a coma.	

Table 6: Example of the discrepancy between HTER and DA annotation tasks: high DA score (high quality) but high HTER score (substantial post-editing).

We then compute the shortest edit distances between MT and post-edited texts with Tercom[9]; this effectively informs us which words were deleted, inserted or replaced. Then, any word w_s in the source aligned to a word w_m in MT that was kept in the post-edit receives a tag OK; if w_s is not aligned with any other word in MT or if w_m was deleted in the post-edit, it is tagged BAD. Thus, BAD tags in the source side indicate which words caused MT errors.

For the MT side, we tag both words and the gaps between them, indicating whether a missing additional word should have been there. Any word w_m aligned to another word w_p in the post-edit receives a tag OK; words deleted or replaced are tagged BAD. Any gap g between words in the MT output, before the first word or after the last one receives a tag OK if no word w_p is inserted in there, and BAD otherwise [10]. Figure 3 shows an example from thr En-De language pair, that demonstrates all possible error types: mistranslated source tokens, annotated with **BAD** on the source; missing words from the MT (deletions: annotated as BAD gap tags); wrong words in MT, annotated with **BAD** on the MT. It also shows the human-edited post-edit as well as the automatically generated alignment between source and target. Further statistics for word-level tags are shown in Table 10 in the Appendix.

3. Baseline performance

As one of the main goals of this dataset is to support the development of better QE models, we report the performance of baseline systems trained on the presented MLQE-PE data. We present baselines trained on the different annotations schemes: a sentence-level QE model trained on the DA scores, and a multi-tasking model trained simultaneously on both the sentence-level HTER scores and the word-level OK/BAD labels.

Both baseline models follow the predictor-estimator architecture (Kim et al., 2017), a two-stage neural model that uses multilevel task learning for translation quality estimation. The *predictor* part receives a source-target pair and learns a feature representation of the translation (target). These representations are used as input vectors for the *estimator* part that is trained to predict translation qualities at sentence and/or word level.

We built on the OpenKiwi framework (Kepler et al., 2019) for the implementation, and train both parts of the models simultaneously. For the predictor, we use transformer-based architectures; specifically we employ pre-trained, multilingual XLM-RoBERTa encoders (Conneau et al., 2020). For both baselines the huggingface implementation of the XLM-RoBERTa `base` model is used [11]. We find that fine-tuning the encoders on the train data before training the full architecture significantly improves performance. Hence, the `xlm-roberta-base` encoder is first fine-tuned on the concatenated source and target sentences from the train and development partitions of all language pairs (see Table 1). The fine-tuning uses a masked language modeling (MLM) loss and follows the universal language model fine-tuning approach (Howard and Ruder, 2018) [12]. The fine-tuned model is then used to jointly encode the source and target sentences, concatenated with target first. The predictor features are generated using average pooling over the encoder output and are forwarded to the estimator module which corresponds to a feed-forward layer.

The first baseline model, $B1$, is trained on the combined train splits ($7,000$ sentence pairs for each language pair) using the DA annotations and no word-

[9] `http://www.cs.umd.edu/~snover/tercom/`

[10] The code to reproduce the word tagging and HTER calculation for the MLQE-PE data can be found in `https://github.com/deep-spin/qe-corpus-builder`

[11] `https://huggingface.co/transformers/pretrained_models.html`

[12] We use a script based on: `https://github.com/huggingface/transformers/blob/master/examples/legacy/run_language_modeling.py` with `per_machine_train_batch_size` set to 16 and `block_size` set to 512.

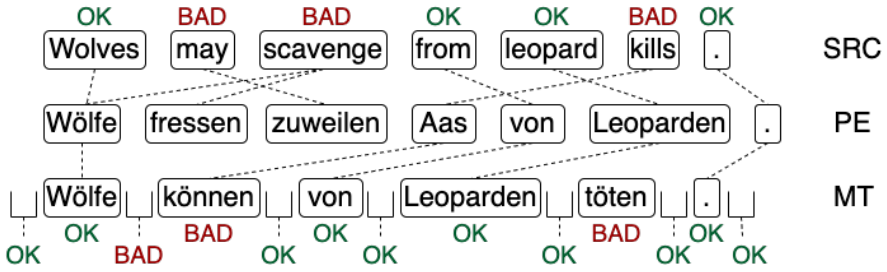

Figure 3: Example of word-level binary tags on source (SRC) and target (MT) level; the gap tags on the target side are represented by the ⊔ symbol. Aligning edges between SRC and PE are produced using SimAlign.

level information. The second baseline model, $B2$, is trained in a multi-tasking fashion using both the HTER and word-level scores of the the combined train splits. Both baselines use the combined development splits ($1,000$ sentence pairs for each language pair) for early-stopping. The hyper-parameters used to train the baseline models are provided in Table 11 in Appendix B.

Tables 7, 8 and 9 present the performance of our baseline systems for each label and language pair, for sentence- and word-level predictions respectively.

Languages	Pearson r ↑	MAE ↓	RMSE ↓
	Direct Assessment		
En-De	0.403	0.629	0.433
En-Zh	0.525	0.683	0.534
Ru-En	0.677	0.702	0.492
Ro-En	0.818	0.556	0.408
Et-En	0.660	0.700	0.543
Ne-En	0.738	0.657	0.524
Si-En	0.513	0.797	0.626
En-Cs	0.352	0.845	0.686
En-Ja	0.230	0.816	0.617
Km-En	0.562	0.788	0.614
Ps-En	0.476	0.852	0.711
AVG	0.541	0.729	0.562

Table 7: Performance at **sentence-level** of Predictor-Estimator baseline models for each label and language pair of the MLQE-PE dataset wrt. direct assessment (DA) scores.

For the sentence-level predictions we use three evaluation metrics: (1) Pearson correlation coefficient which we use as the main performance indicator, (2) mean averaged error (MAE) and (3) root mean squared error (RMSE) between the ground-truth and predicted score. All systems achieve meaningful correlations (Tables 7 and 8), but sentence-level performance seems to be lower for high resource language pairs (especially En-De). This is rather intuitive, since the MT systems for such language pairs provide higher quality predictions that are harder to score (see also the distributions for En-De in Fig. 1). Focusing on the HTER results (Table 8), we can see that the language pairs for which the data

Languages	Pearson r ↑	MAE ↓	RMSE ↓
	HTER		
En-De	0.529	0.183	0.129
En-Zh	0.282	0.287	0.246
Ru-En	0.448	0.255	0.188
Ro-En	0.862	0.144	0.111
Et-En	0.714	0.195	0.149
Ne-En	0.626	0.205	0.160
Si-En	0.607	0.204	0.159
En-Cs	0.306	0.262	0.206
En-Ja	0.098	0.279	0.232
Km-En	0.576	0.241	0.196
Ps-En	0.503	0.333	0.290
AVG	0.502	0.235	0.188

Table 8: Performance at **sentence-level** of Predictor-Estimator baseline models for each label and language pair of the MLQE-PE dataset wrt. HTER scores.

is skewed and has a high proportion of perfect translations (HTER score is zero in Figs 1 and 2) are harder for the baseline model and yield lower correlation with human scores. Regarding the zero-shot language-pairs, the MAE and RMSE metrics seem to be more affected by the lack of training data, implying that the models over- or underestimate the quality. On the contrary, the pearson correlation is mostly affected when translating out of English (En-XX), while the into English translations still show high correlations.

For the word-level scores we use Matthews correlation coefficient (MCC, (Matthews, 1975)) as the primary metric and report the F_1-scores for the OK and BAD classes as well. As we can see in Table 10, word-level errors on the target/MT side are easier for the models to predict resulting in better performance across metrics. In terms of individual language pairs, the word-level predictions demonstrate the same pattern as the sentence-level ones in that performance for high resource language pairs is lower compared to low-resource ones and zero-shot performance drops especially for out-of-English translations.

It should be noted that the baseline scores are indicative of generic quality estimation patterns but not of best

	Words in MT				Words in SRC			
Languages	MCC ↑	F_1-BAD ↑	F_1-OK ↑	F_1-Multi ↑	MCC ↑	F_1-BAD ↑	F_1-OK ↑	F_1-Multi ↑
En-De	0.370	0.455	0.911	0.415	0.322	0.393	0.924	0.363
En-Zh	0.247	0.426	0.723	0.308	0.241	0.394	0.751	0.295
Ru-En	0.256	0.360	0.889	0.319	0.251	0.326	0.893	0.292
Ro-En	0.536	0.642	0.862	0.553	0.511	0.618	0.871	0.539
Et-En	0.461	0.589	0.869	0.512	0.405	0.522	0.879	0.459
Ne-En	0.440	0.828	0.583	0.483	0.390	0.768	0.570	0.438
Si-En	0.425	0.793	0.574	0.456	0.335	0.698	0.544	0.379
En-Cs	0.273	0.454	0.819	0.372	0.224	0.362	0.862	0.312
En-Ja	0.131	0.437	0.497	0.217	0.175	0.393	0.693	0.272
Km-En	0.351	0.766	0.534	0.409	0.279	0.644	0.552	0.355
Ps-En	0.313	0.674	0.631	0.425	0.249	0.501	0.720	0.361
AVG	0.346	0.579	0.717	0.402	0.307	0.511	0.751	0.370

Table 9: Performance at **word-level** of Predictor-Estimator baseline models for each label and language pair of the MLQE-PE dataset.

achieved performance, since it is limited by the small encoder model and simple training approach. Systems submitted for the WMT QE shared tasks (Specia et al., 2020; Specia et al., 2021) were more robust and demonstrated better performance, showing there is ample room for further improvement and experimentation.

4. Current and future use

We aspire for the MLQE-PE dataset to support a diverse set of tasks related to the improvement, interpretation, correction and quality estimation of MT. Since the initial publication of the dataset it has already been used as the main dataset for various shared tasks and peer-reviewed publications. Specifically, WMT Quality Estimation Shared tasks used MLQE-PE data to provide train, development and test sets for sentence and word-level quality estimation in the 2020 and 2021 editions (Specia et al., 2020; Specia et al., 2021). Additionally, the Et-En and Ro-En language pairs were used in the Explainable Quality Estimation shared task, organised as part of the Eval4NLP workshop (Fomicheva et al., 2021). The goal was to invite system submissions that explain sentence-level scores with word-level annotations, and the MLQE-PE sentence and word-level annotations were used for training and development purposes respectively.

Since the dataset contains human-edited post-edits, it is also a suitable resource for Automatic Post Editing (APE) tasks. The En-De and En-Zh parts of the data have been used to provide training and test data for the WMT-APE shared task (2020 and 2021 editions: Chatterjee et al. (2020; Akhbardeh et al. (2021)). It could also be used to identify specific error patterns of MT systems, to facilitate and automate the detection of catastrophic errors or to promote research into model confidence and active learning approaches. The provision of document information could hopefully foster more context-aware approaches in the future, while the provision of NMT models and multi-annotator scores promotes glass-box and uncertainty-aware approaches (Fomicheva et al., 2020b; Wang et al., 2021; Zerva et al., 2021).

5. Conclusion

We introduced MLQE-PE, a new dataset that was mainly created to be used for the tasks of quality estimation (sentence and word-level prediction) and automatic post-editing. It currently contains data in eleven language pairs, direct assessment and post-editing-based sentence-level labels, as well as binary OK/BAD word-level labels. In addition, a subset of the data contains independently created reference translations, which can be used, for example, for machine translation evaluation.

The dataset is freely available and was already used for several tasks. We hope that this data will foster further work on these and other tasks, such as uncertainty estimation and model calibration. We also hope it will sparkle interest from researchers who may want to contribute related resources, i.e., more data, different languages, etc.

6. Acknowledgements

Marina Fomicheva, Frédéric Blain and Lucia Specia were supported by funding from the Bergamot project (EU H2020 Grant No. 825303). André Martins, Chrysoula Zerva 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 Marina Sánchez-Torrón and Camila Pohlmann for monitoring the post-editing process. We also thank Mark Fishel from the University of Tartu for providing the Estonian reference translations.

7. Bibliographical References

Akhbardeh, F., Arkhangorodsky, A., Biesialska, M., Bojar, O., Chatterjee, R., Chaudhary, V., Costa-jussa, M. R., España-Bonet, C., Fan, A., Federmann, C., Freitag, M., Graham, Y., Grundkiewicz, R., Haddow, B., Harter, L., Heafield, K., Homan, C., Huck, M., Amponsah-Kaakyire, K., Kasai, J., Khashabi, D., Knight, K., Kocmi, T., Koehn, P., Lourie, N., Monz, C., Morishita, M., Nagata, M., Nagesh, A., Nakazawa, T., Negri, M., Pal, S., Tapo, A. A., Turchi, M., Vydrin, V., and Zampieri, M. (2021). Findings of the 2021 conference on machine translation (WMT21). In *Proceedings of the Sixth Conference on Machine Translation*, pages 1–88, Online, November. Association for Computational Linguistics.

Blatz, J., Fitzgerald, E., Foster, G., Gandrabur, S., Goutte, C., Kulesza, A., Sanchis, A., and Ueffing, N. (2004). Confidence Estimation for Machine Translation. In *Proc. of the International Conference on Computational Linguistics*, page 315.

Bojar, O., Buck, C., Federmann, C., Haddow, B., Koehn, P., Leveling, J., Monz, C., Pecina, P., Post, M., Saint-Amand, H., Soricut, R., Specia, L., and Tamchyna, A. (2014). Findings of the 2014 workshop on statistical machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 12–58, Baltimore, Maryland, USA, June. Association for Computational Linguistics.

Chatterjee, R., Federmann, C., Negri, M., and Turchi, M. (2019). Findings of the wmt 2019 shared task on automatic post-editing. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 11–28.

Chatterjee, R., Freitag, M., Negri, M., and Turchi, M. (2020). Findings of the WMT 2020 shared task on automatic post-editing. In *Proceedings of the Fifth Conference on Machine Translation*, pages 646–659, Online, November. Association for Computational Linguistics.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. (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, July. Association for Computational Linguistics.

Esplà, M., Forcada, M., Ramírez-Sánchez, G., and Hoang, H. (2019). ParaCrawl: Web-scale parallel corpora for the languages of the EU. In *Proceedings of Machine Translation Summit XVII Volume 2: Translator, Project and User Tracks*, pages 118–119, Dublin, Ireland, 19–23 August. European Association for Machine Translation.

Fomicheva, M., Specia, L., and Guzmán, F. (2020a). Multi-hypothesis machine translation evaluation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1218–1232.

Fomicheva, M., Sun, S., Yankovskaya, L., Blain, F., Guzmán, F., Fishel, M., Aletras, N., Chaudhary, V., and Specia, L. (2020b). Unsupervised quality estimation for neural machine translation. *Transactions of the Association for Computation Linguistics*.

Fomicheva, M., Lertvittayakumjorn, P., Zhao, W., Eger, S., and Gao, Y. (2021). The eval4nlp shared task on explainable quality estimation: Overview and results. In *Proceedings of the 2nd Workshop on Evaluation and Comparison of NLP Systems*, pages 165–178.

Fonseca, E., Yankovskaya, L., Martins, A. F. T., Fishel, M., and Federmann, C. (2019). Findings of the WMT 2019 shared tasks on quality estimation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 1–10, Florence, Italy, August. Association for Computational Linguistics.

Graham, Y., Baldwin, T., Moffat, A., and Zobel, J. (2013). Continuous measurement scales in human evaluation of machine translation. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 33–41.

Guzmán, F., Chen, P.-J., Ott, M., Pino, J., Lample, G., Koehn, P., Chaudhary, V., and Ranzato, M. (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 6097–6110, Hong Kong, China, November. Association for Computational Linguistics.

Howard, J. and Ruder, S. (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.

Kepler, F., Trénous, J., Treviso, M., Vera, M., and Martins, A. F. T. (2019). OpenKiwi: An open source framework for quality estimation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 117–122, Florence, Italy, July. Association for Computational Linguistics.

Kim, H., Lee, J.-H., and Na, S.-H. (2017). Predictor-Estimator using Multilevel Task Learning with Stack Propagation for Neural Quality Estimation. In *Conference on Machine Translation (WMT)*.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86.

Matthews, B. W. (1975). Comparison of the predicted and observed secondary structure of t4 phage

lysozyme. *Biochimica et Biophysica Acta (BBA)-Protein Structure*, 405(2):442–451.

Ng, N., Yee, K., Baevski, A., Ott, M., Auli, M., and Edunov, S. (2019). Facebook fair wmt19 news translation task submission. In *Proc. of WMT*, pages 1–4.

Ott, M., Edunov, S., Grangier, D., and Auli, M. (2018). Scaling neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 1–9.

Ott, M., Edunov, S., Baevski, A., Fan, A., Gross, S., Ng, N., Grangier, D., and Auli, M. (2019). fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53.

Sabet, M. J., Dufter, P., Yvon, F., and Schütze, H. (2020). Simalign: High quality word alignments without parallel training data using static and contextualized embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2020*.

Schwenk, H., Chaudhary, V., Sun, S., Gong, H., and Guzmán, F. (2019). WikiMatrix: Mining 135M Parallel Sentences in 1620 Language Pairs from Wikipedia. *arXiv preprint arXiv:1907.05791*.

Specia, L., Blain, F., Logacheva, V., Astudillo, R., and Martins, A. F. (2018a). Findings of the wmt 2018 shared task on quality estimation. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 689–709.

Specia, L., Scarton, C., and Paetzold, G. H. (2018b). Quality Estimation for Machine Translation. *Synthesis Lectures on Human Language Technologies*, 11(1):1–162.

Specia, L., Blain, F., Fomicheva, M., Fonseca, E., Chaudhary, V., Guzmán, F., and Martins, A. F. T. (2020). Findings of the WMT 2020 shared task on quality estimation. In *Proceedings of the Fifth Conference on Machine Translation*, pages 743–764, Online, November. Association for Computational Linguistics.

Specia, L., Blain, F., Fomicheva, M., Zerva, C., Li, Z., Chaudhary, V., and Martins, A. F. T. (2021). Findings of the WMT 2021 shared task on quality estimation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 684–725, Online, November. Association for Computational Linguistics.

Tang, Y., Tran, C., Li, X., Chen, P., Goyal, N., Chaudhary, V., Gu, J., and Fan, A. (2020). Multilingual translation with extensible multilingual pretraining and finetuning. *ArXiv*, abs/2008.00401.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Wang, K., Shi, Y., Wang, J., Zhang, Y., Zhao, Y., and Zheng, X. (2021). Beyond glass-box features: Uncertainty quantification enhanced quality estimation for neural machine translation. *arXiv preprint arXiv:2109.07141*.

Zerva, C., van Stigt, D., Rei, R., Farinha, A. C., Ramos, P., de Souza, J. G., Glushkova, T., Vera, M., Kepler, F., and Martins, A. F. (2021). Ist-unbabel 2021 submission for the quality estimation shared task. In *Proceedings of the Sixth Conference on Machine Translation*, pages 961–972.

A. BAD tag distribution

We present here a breakdown of the proportion of BAD tags in each split and language pair in the MLQE-PE dataset. For the target-side annotations, we calculate the proportion of BAD tokens over the total amount of tokens in the target sentence. While we also count the BAD tags if they occur in the gaps, we do do not increase the total token number to avoid obtaining misleadingly low proportion of word-level errors.

We see that most sentences in the dataset have at least one BAD tag and this is more intense for the low resource language pairs, with Ne-En, Si-En and Km-En having at least one BAD tag in almost every sentence. The overall amount of BAD tags is also higher for the low- and mid-resource language pairs.

The target-side BAD tags appear consistently higher in proportion when compared to the source side ones, but this is partly a property of the annotation, since on the target side these tags account not only for erroneous tokens but also deletion and insertion errors.

B. Baseline Hyperparameters

We present the hyperparameters used to train the baseline models in Table 11. The configurations follow the configuration file format of OpenKiwi and any additional configurations not mentioned in the table are

		Source		Target	
		BAD tags	Sentences	BAD tags	Sentences
En-De	Train	12.64%	63.40%	18.97%	67.60%
	Dev	13.32%	66.10%	19.69%	70.10%
	Test20	11.82%	57.90%	17.82%	63.00%
	Test21	12.71%	61.40%	12.71%	61.40%
En-Zh	Train	33.00%	90.86%	44.77%	92.40%
	Dev	21.82%	72.90%	28.47%	74.20%
	Test20	25.37%	80.00%	34.06%	81.30%
	Test21	17.02%	63.60%	17.02%	63.60%
Et-En	Train	20.72%	80.46%	28.78%	85.70%
	Dev	21.73%	86.40%	30.17%	91.80%
	Test20	23.34%	85.60%	32.93%	91.70%
	Test21	21.98%	83.70%	21.98%	83.70%
Ne-En	Train	52.14%	97.97%	67.16%	98.39%
	Dev	55.75%	99.60%	71.52%	99.80%
	Test20	56.04%	99.10%	72.54%	99.30%
	Test21	55.95%	99.60%	55.95%	99.60%
Ro-En	Train	29.04%	88.90%	37.24%	89.70%
	Dev	17.34%	67.20%	21.32%	68.30%
	Test20	20.04%	73.30%	24.96%	74.90%
	Test21	20.50%	76.90%	20.50%	76.90%
Ru-En	Train	13.01%	48.67%	18.82%	51.64%
	Dev	12.72%	46.90%	16.39%	49.50%
	Test20	11.28%	43.40%	15.47%	46.10%
	Test21	18.17%	61.60%	18.17%	61.60%
Si-En	Train	47.93%	95.77%	66.69%	96.16%
	Dev	48.00%	96.10%	67.00%	96.50%
	Test20	48.21%	96.60%	67.08%	96.80%
	Test21	47.39%	97.20%	47.39%	97.20%
Zero-Shot					
En-Cs	Test21	17.63%	62.90%	17.63%	62.90%
En-Ja	Test21	20.82%	71.80%	20.82%	71.80%
Km-En	Test21	43.08%	96.16%	43.08%	96.16%
Ps-En	Test21	28.84%	76.40%	28.84%	76.40%

Table 10: Ratio of BAD tags in the word-level data for the different splits of the dataset (third and fifth columns), and ratio of sentences containing at least one such tag (fourth and sixth columns).

identical to the default ones on github [13].

Module	Parameter	Value
System	batch_size	2
Encoder	hidden_size	768
Decoder	dropout	0.1
	hidden_size	768
Trainer	early_stop_patience	5

Table 11: Hyper-parameters for the baseline models.

[13] https://github.com/Unbabel/OpenKiwi/blob/master/config/xlmroberta.yaml