# Welocalize-ARC/NKUA's Submission to the WMT 2022 Quality Estimation Shared Task

**Eirini Zafeiridou**[1,2], **Sokratis Sofianopoulos**[3]
[1]Welocalize Inc / Frederick, MD, United States
[2]National and Kapodistrian University of Athens / Athens, Greece
[3]Institute for Language and Speech Processing - Athena RC / Athens, Greece
eirini.zafeiridou@welocalize.com, s_sofian@athenarc.gr

## Abstract

This paper presents our submission to the WMT 2022 quality estimation shared task and more specifically to the quality prediction sentence-level direct assessment (DA) subtask. We build a multilingual system based on the predictor–estimator architecture by using the XLM-RoBERTa transformer for feature extraction and a regression head on top of the final model to estimate the $z$-standardized DA labels. Furthermore, we use pretrained models to extract useful knowledge that reflect various criteria of quality assessment and demonstrate good correlation with human judgements. We optimize the performance of our model by incorporating this information as additional external features in the input data and by applying Monte Carlo dropout during both training and inference.

## 1 Introduction

Machine translation quality estimation (MTQE) is the task of automatically estimating the quality of the MT output without using reference translations or any other human input (Blatz et al., 2004; Specia et al., 2009, 2018). MTQE has many use cases and can be applied in various settings (Specia and Shah, 2018). It can be used to estimate the post-editing effort, to rank and compare outputs of different MT systems or to classify the segments that need post-editing. It can also be used to estimate the quality of the final translations as well as to filter out noisy segments from translation memories or training datasets. MTQE techniques usually have multiple granularity levels and can be applied to a word, phrase, sentence or even to an entire document. Such systems are highly efficient when a vast amount of machine translated segments need to be evaluated in less time, with less effort and lower costs compared to traditional evaluation techniques.

The WMT 2022 quality estimation shared task includes the following separate tasks: quality prediction, explainable QE and critical error detection. Our team participated in the quality prediction sentence-level direct assessment (DA) subtask with a multilingual MTQE system.

Specifically, we developed a cross-lingual MTQE system following the predictor–estimator architecture (Kim and Lee, 2016; Kim et al., 2017). We used the large-scale pretrained XLM-RoBERTa (XLM-R)[1] model (Conneau et al., 2020) for feature extraction, similarly to Chen et al. (2021). We combined the model's output with additional external features that demonstrate good correlation with the target variable. We then used the concatenated vector as input to our final MTQE regression model. Our regressor is a feed-forward neural network with a linear output layer used to estimate the $z$-standardized DA labels.

## 2 Quality prediction: sentence-level direct assessment

The quality prediction task of the WMT 2022 quality estimation shared task consists of a sentence-level and a word-level subtask. Using the provided annotated training data, the objective of the sentence-level direct assessment subtask is to develop a system that automatically estimates a quality score for each provided sentence pair which is highly correlated with human-generated $z$-standardized DA values.

### 2.1 Data

According to the instructions, for each language pair, participants can use all the annotations offered for the quality estimation shared tasks of the preceding year(s) that are accessible through the MLQE-PE GitHub page.[2]

MLQE-PE is a multilingual dataset for quality estimation which includes 11 language combinations covering low, medium and high re-

[1]https://huggingface.co/xlm-roberta-large
[2]https://github.com/sheffieldnlp/mlqe-pe



*Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 653–660
Abu Dhabi, December 7–8, 2022. 

source languages (Fomicheva et al., 2020a,c). The dataset is mainly created by translating sentences from Wikipedia articles using cutting-edge transformer NMT models, and by having expert linguists annotate the translations based on a modified version of DA ratings. Each sentence is annotated individually using the FLORES setup (Guzmán et al., 2019), in which three qualified translators provide evaluations on a scale of 0–100 based on their perceived translation quality. Raw DA scores are then standardized and transformed into $z$-scores by using the mean and standard deviation of every single annotator. The $z$-standardized per-annotator values are then averaged in order to get one final score for every translation.

The organizers also provide additional train, development, and test sets for the English–Marathi language pair that is not included in the MLQE-PE dataset.

| language pair | Train | Dev. | Test |
|---|---|---|---|
| en–mr | 26000 | 1000 | 1000 |
| en–cs | – | 1000 | 1000 |
| en–ja | – | 1000 | 1000 |
| km–en | – | 1000 | 1000 |
| ps–en | – | 1000 | 1000 |
| en–de | 9000 | 1000 | – |
| en–zh | 9000 | 1000 | – |
| et–en | 9000 | 1000 | – |
| ne–en | 9000 | 1000 | – |
| ro–en | 9000 | 1000 | – |
| ru–en | 9000 | 1000 | – |
| si–en | 9000 | 1000 | – |
| en–yo | – | – | 1000 |
| **total** | 89000 | 12000 | 6000 |

Table 1: Size of the provided train, development and test sets per language (in sentences)

The data for the sentence-level quality prediction subtask can be downloaded from the task's GitHub page.[3] The number of the available sentences per language is illustrated in the Table 1.

According to the instructions, it is also feasible to use the DA annotations that were generated for the metrics shared tasks in previous years.

For the training of our models, we use only the training part of the data provided by the organizers. For the English–Japanese language pair, we also use the training data of the 2020 metrics shared task.

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

### 2.2 Evaluation

This year, the primary evaluation metric for the sentence-level DA subtask is the *Spearman*'s rank correlation coefficient which is used to reflect the correlation between the predicted scores and the human annotated $z$-standardized DA labels. Secondary metrics also include MAE, RMSE, and *Pearson*'s correlation coefficient.

## 3 Method

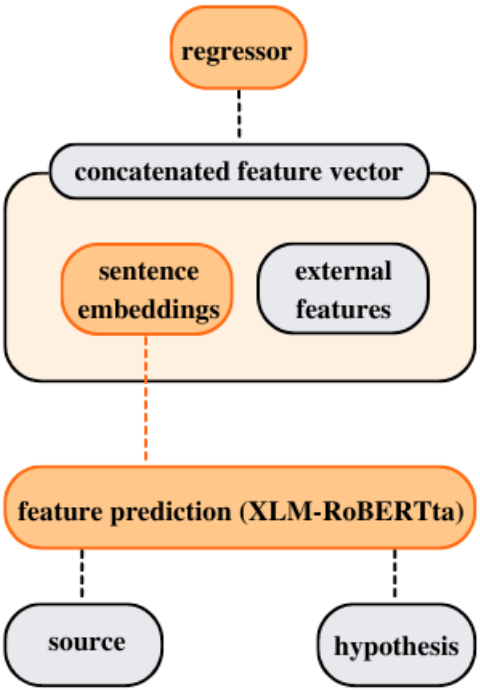


Figure 1: Model architecture

For the sentence-level direct assessment subtask we build and use a system based on the predictor–estimator architecture (Kim and Lee, 2016; Kim et al., 2017). Following similar state-of-the-art approaches (Fomicheva et al., 2020b; Moura et al., 2020; Rei et al., 2020; Zerva et al., 2021; Chen et al., 2021; Wang et al., 2021) we choose the pre-trained XLM-RoBERTa[1] model (Conneau et al., 2020) to encode the input sequences and predict our features. We keep the XLM-R encoder frozen during training and we use it to generate cross-lingual representations over the source sentences and their corresponding translations. We then concatenate the output with additional external features and we feed the final feature vector to a feed-forward layer to finally estimate the continuous

$z$-standardized DA scores. We also employ Monte Carlo dropout during both training and inference to optimize the performance of our model. We use the mean-squared-error loss function and the AdamW (Loshchilov and Hutter, 2019) optimizer with a learning rate of $10^{-5}$. The architecture of our model is illustrated in the Figure 1.

### 3.1 Cross-lingual Representations

In order to extract cross-lingual representations for each sentence pair, we start by encoding each source sentence and its hypothesis separately. From the output vectors, we extract the <s> classification token (equivalent to the [CLS]) that corresponds to the representation of the whole sequence (Ranasinghe et al., 2020). Then, similarly to the methodology proposed in RUSE (Shimanaka et al., 2018), we use the following sentence embeddings:

- Source embedding representation: $\vec{s}$
- Hypothesis embedding representation: $\vec{h}$
- Element-wise product: $\vec{s} \circ \vec{h}$
- Element-wise absolute difference: $| \vec{s} - \vec{h} |$

Motivated by the implementation of Rei et al. (2020), we concatenate the above representations into a single vector. Furthermore, we enrich the vector with additional external features $\vec{f}$ resulting in a final feature vector $\vec{x} = \left[\vec{f}; \vec{h}; \vec{s}; \vec{s} \circ \vec{h}; |\vec{s} - \vec{h}|\right]$, which is used as input to the output layer of our model.

### 3.2 Additional external features

Fomicheva et al. (2020c) suggested the use of glass-box features to predict the quality of the NMT outputs. Specifically, they proposed methods to quantify the model's uncertainty in unsupervised QE scenarios. Moura et al. (2020) and Zerva et al. (2021) also used such glass-box features as an effective strategy for the development of their QE systems. In our approach, we use the sentence-level NMT model scores included in the MLQE-PE dataset (Fomicheva et al., 2020a,c) and we further explore additional characteristics that can be effectively used in similar QE settings. We suggest a set of external features that reflect various criteria of translation quality assessment and exhibit good correlation with human judgements, as illustrated in Table 2.

**Masked Language Model scores**
(features: src_ppl, hyp_ppl, diff_ppl)

According to Lau et al. (2017), language models (LMs) can be effectively used to estimate linguistic acceptability judgements. Salazar et al. (2020) showed that pseudo-log-likelihood scores (PLLs) and their corresponding pseudo-perplexities (PPPLs) derived from masked language models (MLMs) can help to distinguish linguistically acceptable from unacceptable sentences in an unsupervised way with comparable performance to large unidirectional autoregressive LMs. Based on the above observation, our primary objective is to derive scores at the sentence level that reflect the overall likelihood that the model gives to an entire sentence. We choose the multilingual XLM-RoBERTa[1] model and compute PLL scores by iteratively masking all tokens of the sequence. We generate PLL scores for both the source and the hypothesis and then we also calculate their absolute difference.

**NMT Model scores**
(features: model_scores)

According to Fomicheva et al. (2020c), seq2seq NMT models can provide meaningful insights for measuring the model's uncertainty that can be effectively used to estimate translation quality. At each timestep, the NMT system returns the probability distribution for every token in the sequence by applying a softmax function over the target language vocabulary. The token-level probabilities are then used to compute a sentence-level log-likelihood score. In our implementation we extracted this feature directly from the MLQE-PE dataset (Fomicheva et al., 2020a,c). Even if this information is already included in the provided dataset, we also decided to build another model, similar to the one described in this paper, that predicts these specific values when there is no access to the NMT system used to produce the translations.

**Independent NMT Model scores**
(features: M2M100_loss)

We use the pretrained M2M100 multilingual seq-to-seq model (Fan et al., 2020) to re-score the provided NMT outputs for each sentence pair. Our objective is not to generate a new hypothesis for each source sentence, but to compare every given hypothesis with the prediction produced by another multilingual translation system. The final score corresponds to the calculated cross-entropy loss

when comparing the generated prediction of the M2M100 model to the provided NMT hypothesis.

**Semantic Textual Similarity scores**

(features: cos_sim)

Sentence similarity corresponds to the task of automatically identifying how similar or dissimilar two texts are. Neural models compare sentences by initially transforming them into semantic vectors, also known as sentence embeddings. We use the LaBSE[4] (Feng et al., 2022) pretrained model through the sentence transformers library (Reimers and Gurevych, 2019, 2020) to obtain a vector representation for every source sentence and its hypothesis. Then we compare their embeddings and get a cosine similarity score at sentence level.

**COMET scores**

(features: COMET_qe)

COMET (Rei et al., 2020) is a multilingual MT quality evaluation framework that demonstrates high correlation with human judgements. In our implementation, we use the reference-free wmt21-comet-qe-mqm[5] model (Rei et al., 2021), pretrained based on the MQM benchmark, which can be computed automatically without having available any reference translation. In this way, we are able to get one predicted score for every sentence and use this value as an additional feature during the training of our model.

**HTER scores**

(features: hter_scores)

The translation edit rate (TER) (Snover et al., 2006) calculates the editing operations needed to transform an MT output into a version that exactly matches at least one candidate translation among a list of gold-standard reference texts. The human-targeted translation edit rate (HTER) (Snover et al., 2006) is another version of the TER metric which incorporates the human factor in the process and requires human post-edits of the MT output. Even if this information is already included in the MLQE-PE dataset, we use the available HTER annotations in order to train another model, similar to the one described in this paper, that estimates the post-editing effort by predicting HTER scores for each source sentence and its translation. We finally use this information as an additional external feature for our final model.

The *Spearman* and *Pearson* correlation between all the aforementioned features and the target variable can be found in the Table 2. Based on these values, it seems that the features with the highest correlation are the cosine similarity (cos_sim), the COMET qe (COMET_qe), and the human-targeted translation edit rate (hter_scores). The NMT model scores (model_scores) and the independent NMT model scores (M2M100_loss) also demonstrate a moderate correlation with the $z$-standardized DA scores, while the masked language model scores (src_ppl, hyp_ppl, diff_ppl) have quite lower correlation comparing to the rest.

| features | *Spearman* $r$ | *Pearson* $r$ |
|---|---|---|
| src_ppl | −0.15 | −0.16 |
| hyp_ppl | −0.14 | −0.14 |
| diff_ppl | −0.11 | −0.13 |
| M2M100_loss | −0.29 | −0.25 |
| cos_sim | 0.34 | 0.40 |
| COMET_qe | 0.42 | 0.41 |
| model_scores | 0.25 | 0.30 |
| hter_scores | −0.37 | −0.37 |

Table 2: *Spearman* and *Pearson* correlation between the external selected features and the $z$-standardized DA scores. Features are described one by one in section 3.2.

### 3.3 Monte Carlo dropout

Dropout refers to randomly dropping nodes while training a neural network (Srivastava et al., 2014) and it is an effective strategy to prevent a model from overfitting. During training we use Monte Carlo dropout with a rate of 0.1 to mask random neurons of the model. Likewise, during inference we perform numerous iterations for each test instance and in this way we obtain a different score each time for the same instance by applying Monte Carlo dropout. Then, we use all the model's estimates to get an average score for every single sentence.

## 4 Experimental Results

In this section we present the performance of our model on the provided test dataset for the WMT 2022 shared task on quality evaluation for the prediction of sentence-level direct assessments. In particular, our model outperformed the baseline system in terms of *Spearman* and *Pearson* correlation in all the multilingual and bilingual tasks, in which we participated, as illustrated in the Tables 3 and 4 respectively. In the multilingual (full) sub-

[4] https://huggingface.co/sentence-transformers/LaBSE

[5] https://github.com/Unbabel/COMET/blob/master/METRICS.md

| Model | Multi (full) | Multi (w/o en–yo) | en–cs | en–ja | en–mr | km–en |
|---|---|---|---|---|---|---|
| our model | 0.448 | 0.506 | 0.563 | 0.276 | 0.444 | 0.623 |
| baseline model | 0.415 | 0.497 | 0.560 | 0.272 | 0.436 | 0.579 |

Table 3: *Spearman*'s correlations of the 2022 sentence-level DA subtask

| Model | Multi (full) | Multi (w/o en–yo) | en–cs | en–ja | en–mr | km–en |
|---|---|---|---|---|---|---|
| our model | 0.455 | 0.535 | 0.592 | 0.281 | 0.586 | 0.618 |
| baseline model | 0.393 | 0.511 | 0.576 | 0.273 | 0.525 | 0.568 |

Table 4: *Pearson*'s correlations of the 2022 sentence-level DA subtask

task we were ranked 3rd while in the multilingual (w/o en–yo) we got the 4th place.

Based on the official results, it seems that the lowest performing language pair, for both our model and the baseline, is English–Japanese while the highest performing one is Khmer–English. We did not further examine the reasons of this pattern, as we considered this exercise out of the scope of our study. In a future work, it would be useful to investigate whether certain factors contribute to this pattern (such as the source and target language complexity, the writing script, the performance of the pretrained models used to generate the features for each language, or even the content of the test dataset).

It is also worth mentioning that for most language pairs of the test sets, as illustrated in the Table 1, we did not have available training data. If our model had explicitly seen all of these languages during training, we would expect its performance to be improved.

The results of the test set from the official leaderboard[6] for each language pair, in which we participated, can be found in the Official results of the WMT 2022 QE Task 1 – Sentence-level Direct Assessment. In these tables, our proposed model is compared to the baseline system in terms of RMSE, MAE, *Spearman* and *Pearson* correlation coefficient.

[6] https://www.statmt.org/wmt22/quality-estimation-task_results.html

## 5 Conclusions

This paper presents our submission to the WMT 2022 quality estimation Task 1 on sentence-level direct assessment. We introduce a model trained based on the predictor–estimator architecture using the XLM-RoBERTa[1] for feature prediction and a regression head to finally estimate the z-standardized DA values. We suggest the use of additional external features that reflect different criteria of human judgements and multiple levels of translation quality. These features exhibit good correlation with the target variable and consequently with human annotations. Our approach is applicable in multilingual settings even with languages or writing scripts not explicitly seen during the training of the MTQE model. Our system demonstrates competitive results and a strong correlation with human judgements of quality assessment outperforming the baseline system in terms of both *Spearman* and *Pearson* correlation coefficient.

## 6 Acknowledgements

The work reported in this paper was performed as part of the final master's thesis titled "Quality Estimation for Neural Machine Translation" for the completion of the MSc in Language Technology, a degree co-organized by the National Kapodistrian University of Athens and the Institute for Language and Speech Processing of the Athens Research Center. We would like to thank Welocalize and especially David Clarke for greatly supporting the development of this MSc thesis.

## References


John Blatz, Erin Fitzgerald, George Foster, Simona Gandrabur, Cyril Goutte, Alex Kulesza, Alberto Sanchis, and Nicola Ueffing. 2004. Confidence Estimation for Machine Translation. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, pages 315–321, Geneva, Switzerland. COLING.

Yimeng Chen, Chang Su, Yingtao Zhang, Yuxia Wang, Xiang Geng, Hao Yang, Shimin Tao, Guo Jiaxin, Wang Minghan, Min Zhang, Yujia Liu, and Shujian Huang. 2021. HW-TSC's Participation at WMT 2021 Quality Estimation Shared Task. In *Proceedings of the Sixth Conference on Machine Translation*, pages 890–896, Online. Association for Computational Linguistics.

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.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2020. Beyond English-Centric Multilingual Machine Translation.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT Sentence Embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

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*, pages 1010–1017, 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*, 8:539–555.

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.

Hyun Kim and Jong-Hyeok Lee. 2016. A Recurrent Neural Networks Approach for Estimating the Quality of Machine Translation Output. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 494–498, San Diego, California. Association for Computational Linguistics.

Hyun Kim, Jong-Hyeok Lee, and Seung-Hoon Na. 2017. Predictor-Estimator using Multilevel Task Learning with Stack Propagation for Neural Quality Estimation. In *Proceedings of the Second Conference on Machine Translation*, pages 562–568, Copenhagen, Denmark. Association for Computational Linguistics.

Jey Han Lau, Alexander Clark, and Shalom Lappin. 2017. Grammaticality, Acceptability, and Probability: A Probabilistic View of Linguistic Knowledge. *Cognitive science*, pages 1202–1241.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations (ICLR)*. arXiv.

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*, pages 1029–1036, Online. Association for Computational Linguistics.

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

Ricardo Rei, Ana C Farinha, Chrysoula Zerva, Daan van Stigt, Craig Stewart, Pedro Ramos, Taisiya Glushkova, André F. T. Martins, and Alon Lavie. 2021. Are References Really Needed? Unbabel-IST 2021 Submission for the Metrics Shared Task. In *Proceedings of the Sixth Conference on Machine Translation*, pages 1030–1040, Online. Association for Computational Linguistics.

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

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. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2020. Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Julian Salazar, Davis Liang, Toan Q. Nguyen, and Katrin Kirchhoff. 2020. Masked Language Model Scoring. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2699–2712, Online. Association for Computational Linguistics.

Hiroki Shimanaka, Tomoyuki Kajiwara, and Mamoru Komachi. 2018. RUSE: Regressor Using Sentence Embeddings for Automatic Machine Translation Evaluation. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 751–758, Belgium, Brussels. Association for Computational Linguistics.

Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. A Study of Translation Edit Rate with Targeted Human Annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas.

Lucia Specia, Frédéric Blain, Varvara Logacheva, Ramón F. Astudillo, and André F. T. Martins. 2018. 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, Belgium, Brussels. Association for Computational Linguistics.

Lucia Specia and Kashif Shah. 2018. *Machine Translation Quality Estimation: Applications and Future Perspectives*, pages 201–235. Springer International Publishing, Cham.

Lucia Specia, Marco Turchi, Nicola Cancedda, Nello Cristianini, and Marc Dymetman. 2009. Estimating the Sentence-Level Quality of Machine Translation Systems. In *Proceedings of the 13th Annual conference of the European Association for Machine Translation*, pages 28–35, Barcelona, Spain. European Association for Machine Translation.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research*, 15:1929–1958.

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

Chrysoula Zerva, Daan van Stigt, Ricardo Rei, Ana C Farinha, Pedro Ramos, José G. C. de Souza, Taisiya Glushkova, Miguel Vera, Fabio Kepler, and André F. T. Martins. 2021. IST-Unbabel 2021 Submission for the Quality Estimation Shared Task. In *Proceedings of the Sixth Conference on Machine Translation*, pages 961–972, Online. Association for Computational Linguistics.

## 7 Official results of the WMT 2022 QE Task 1 – Sentence-level Direct Assessment

| Model | *Spearman* $r$ | *Pearson* $r$ | RMSE | MAE | Disk footprint (bytes) | Model params. |
|---|---|---|---|---|---|---|
| baseline | 0.415 | 0.393 | 0.979 | 0.820 | 2,280,011,066 | 564,527,011 |
| our model | 0.448 | 0.455 | 0.794 | 0.632 | 2,307,101,417 | 576,733,248 |

Table 5: Evaluation of the **Multilingual models** in the 2022 DA subtask

| Model | *Spearman* $r$ | *Pearson* $r$ | RMSE | MAE | Disk footprint (bytes) | Model params. |
|---|---|---|---|---|---|---|
| baseline | 0.497 | 0.511 | 0.748 | 0.585 | 2,280,011,066 | 564,527,011 |
| our model | 0.506 | 0.535 | 0.733 | 0.571 | 2,307,068,585 | 576,725,041 |

Table 6: Evaluation of the **Multilingual models (without en–yo)** in the 2022 DA subtask

| Model | *Spearman* $r$ | *Pearson* $r$ | RMSE | MAE | Disk footprint (bytes) | Model params. |
|---|---|---|---|---|---|---|
| baseline | 0.560 | 0.576 | 0.804 | 0.608 | 2,280,011,066 | 564,527,011 |
| our model | 0.563 | 0.592 | 0.785 | 0.610 | 2,307,068,585 | 576,725,041 |

Table 7: Evaluation of the **en–cs models** in the 2022 DA subtask

| Model | *Spearman* $r$ | *Pearson* $r$ | RMSE | MAE | Disk footprint (bytes) | Model params. |
|---|---|---|---|---|---|---|
| baseline | 0.272 | 0.273 | 0.747 | 0.576 | 2,280,011,066 | 564,527,011 |
| our model | 0.276 | 0.281 | 0.755 | 0.579 | 2,307,068,585 | 576,725,041 |

Table 8: Evaluation of the **en–ja models** in the 2022 DA subtask

| Model | *Spearman* $r$ | *Pearson* $r$ | RMSE | MAE | Disk footprint (bytes) | Model params. |
|---|---|---|---|---|---|---|
| baseline | 0.436 | 0.525 | 0.628 | 0.461 | 2,280,011,066 | 564,527,011 |
| our model | 0.444 | 0.586 | 0.534 | 0.401 | 2,307,068,585 | 576,725,041 |

Table 9: Evaluation of the **en–mr models** in the 2022 DA subtask

| Model | *Spearman* $r$ | *Pearson* $r$ | RMSE | MAE | Disk footprint (bytes) | Model params. |
|---|---|---|---|---|---|---|
| baseline | 0.579 | 0.568 | 0.774 | 0.616 | 2,280,011,066 | 564,527,011 |
| our model | 0.623 | 0.618 | 0.794 | 0.619 | 2,307,068,585 | 576,725,041 |

Table 10: Evaluation of the **km–en models** in the 2022 DA subtask