---

**Proceedings Paper:**
Gajbhiye, A., Fomicheva, M., Alva-Manchego, F. et al. (4 more authors) (2021) Knowledge distillation for quality estimation. In: Zong, C., Xia, F., Li, W. and Navigli, R., (eds.) Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. The Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, 01-06 Aug 2021, Bangkok, Thailand (virtual conference). Association for Computational Linguistics (ACL) , pp. 5091-5099. ISBN 9781954085541

10.18653/v1/2021.findings-acl.452

---

# Knowledge Distillation for Quality Estimation

**Amit Gajbhiye,[1] Marina Fomicheva,[1] Fernando Alva-Manchego,[1] Frédéric Blain,[1,2]**
**Abiola Obamuyide,[1] Nikolaos Aletras,[1] Lucia Specia[1,3]**
[1]University of Sheffield, [2]University of Wolverhampton, [3]Imperial College London
{a.gajbhiye,m.fomicheva,f.alva,a.obamuyide,n.aletras}@shef.ac.uk
f.blain@wlv.ac.uk
l.specia@imperial.ac.uk

## Abstract

Quality Estimation (QE) is the task of automatically predicting Machine Translation quality in the absence of reference translations, making it applicable in real-time settings, such as translating online social media conversations. Recent success in QE stems from the use of multilingual pre-trained representations, where very large models lead to impressive results. However, the inference time, disk and memory requirements of such models do not allow for wide usage in the real world. Models trained on distilled pre-trained representations remain prohibitively large for many usage scenarios. We instead propose to directly transfer knowledge from a strong QE teacher model to a much smaller model with a different, shallower architecture. We show that this approach, in combination with data augmentation, leads to light-weight QE models that perform competitively with distilled pre-trained representations with 8x fewer parameters.

## 1 Introduction

Quality Estimation (QE) aims to predict the quality of the output of Machine Translation (MT) systems when no gold-standard translations are available. It can make MT useful in real-world applications by informing end-users on the translation quality. We focus on sentence-level QE, usually formulated as a regression task where quality is required to be predicted on an continuous scale, e.g. 0-100.

The high performances achieved in the most recent shared task on sentence-level QE (Specia et al., 2020) have been attributed to the use of strong pre-trained language models, namely BERT (Devlin et al., 2018) and its multilingual variants, especially XLM-Roberta (Conneau et al., 2020a). These models have an extremely large number of parameters and, since they are required at training and inference time, they are very disk and RAM-hungry, also making inference slow. This poses challenges for real-time inference, and prohibits deployment on client machines with limited resources.

Making models based on pre-trained representations smaller and more usable in practice is an active area of research. One approach is Knowledge Distillation (KD), aiming to extract knowledge from a top-performing large model (the *teacher*) into a smaller (in terms of memory print, computational power and prediction latency) yet well-performing model (the *student*) (Hinton et al., 2015; Gou et al., 2020). KD techniques have been used to make BERT and similar models smaller. For example, DistilBERT (Sanh et al., 2019) and TinyBERT (Jiao et al., 2020) follow the same general architecture as the teacher BERT, but with a reduced number of layers. However, these student models are also based on Transformers and, as such, they still have too large memory and disk footprints. For instance, the number of parameters in the multilingual DistilBERT-based TransQuest model for QE (Ranasinghe et al., 2020) is 135M.

In this paper, we propose **to distill the QE model directly**, where the student architecture can be completely different from that of the teacher. Namely, we distill a large and powerful QE model based on XLM-Roberta into a small RNN-based model. Existing work along these lines has applied KD mainly to classification tasks (Tang et al., 2019; Sun et al., 2019). We instead explore this approach in the context of **regression**. In contrast to classification, where KD provides useful information on the output distribution of incorrect classes, for regression the teacher predictions are point-based estimates, and as such have the same properties as gold labels. Therefore, it is not obvious whether teacher-student learning can be beneficial. The few existing works on KD for regression (Chen et al., 2017; Takamoto et al., 2020) use the teacher loss to minimise the impact of noise in the teacher



*Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 5091–5099

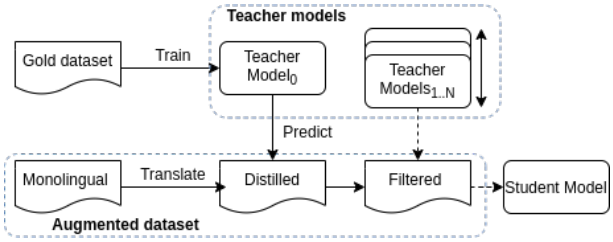


Figure 1: KD with data augmentation and noise filtering based on teacher uncertainty.

| Name | #params | Inference Speed (secs.) | Inference RAM (MiB) | Disk (M) |
|---|---|---|---|---|
| $TQ_{XLM-R-Large}$ | 561M | 0.82 | 9,263.5 | 2140 |
| $TQ_{DistilBERT}$ | 135M | 1.09 | 1,979.2 | 517 |
| BiRNN | 18M | 0.39 | 155.6 | 132 |

Table 1: Efficiency. Inference speed and RAM for prediction are for 1 sentence on CPU (Intel Xeon Silver 4114 CPU @ 2.20GHz).

predictions on the student training. However, this approach requires access to gold labelled examples to train the student, which in our case are very limited in number.

Our approach allows for much larger **unlabelled student training datasets**, built only from source-MT pairs and labelled by the teacher model. We study the performance of student models under different training data regimes: standard training with gold labels, training with teacher predictions on the same data, training with teacher predictions on augmented in-domain and out-of-domain data, as well as augmented data filtered based on uncertainty of teacher predictions. Interestingly, we find that (i) training with teacher predictions results in better performance than training with gold labels; and (ii) student models trained with augmented data perform competitively with DistilBERT-based TransQuest predictors with 8x fewer parameters.

## 2 Approach

Figure 1 summarises our approach, with the following main components:

**Teacher-student training.** We use predictions from a SoTA QE model to train a light-weight student with a different architecture. Specifically, as the teacher model we use the recently proposed TransQuest QE system (Ranasinghe et al., 2020) that fine-tunes multilingual pre-trained representations from XLM-Roberta-Large (Conneau et al., 2020a) to predict a continuous sentence-level quality score. For the student model, we rely on the BiRNN QE architecture proposed by Ive et al. (2018).[1] The BiRNN model encodes both source and translation sentences independently using two bi-directional Recurrent Neural Networks (RNNs). The two resulting sentence representations are concatenated as the weighted sum of their word vectors, generated by an attention mechanism. For predictions at sentence-level, the weighted representation of the two input sentences is passed through a dense layer with sigmoid activation to generate the quality estimates. Table 1 shows the number of parameters, memory and disk space requirements, as well as inference speed for the teacher model ($TQ_{XLM-R-Large}$), student model (BiRNN) and TransQuest system built on DistilBERT ($TQ_{DistilBERT}$). We refer the reader to Appendix A for the details on the architecture and implementation for these models.

In classification tasks, KD benefits learning as it uses information on the output distribution and has an effect akin to label smoothing (Tang et al., 2020). In regression, teacher labels are instead point-wise estimates just like the gold labels. Existing work on KD for regression uses teacher loss to minimise the impact of noise in the teacher predictions on student training (Takamoto et al., 2020). However, this approach is not suitable for QE as we have access to a very limited number of gold-labelled examples. We propose a simple strategy that relies directly on teacher predictions for training the student model.

**Data augmentation.** The power of QE models based on pre-trained representations is due to the rich knowledge that comes from training Transformer-based language models on very large amounts of data. Typically, much smaller datasets are available for downstream tasks, which suffice for fine-tuning but that are hardly suitable for training a neural model from scratch. We exploit the teacher-student framework to produce additional training data. Specifically, we first generate MT outputs for a set of sentences in the source language and domain of interest using the same MT system that was used for generating the test data. Second, we use the teacher model described above to pro-

[1] The implementation of our student models is available at https://github.com/sheffieldnlp/deepQuest-py.

duce predictions. These predictions are then used as labels for training the student.

**Noise filtering.** The benefits of data augmentation can be hampered by noise in teacher predictions. In a classification setting, where the student loss is computed with respect to the output distribution of the teacher model, this issue is ameliorated by the example re-weighting effect where teacher predictions with higher confidence have an overall higher impact on learning (Furlanello et al., 2018). Previous work has used teacher loss to address this issue for regression (Chen et al., 2017). However, this strategy is not suitable for data augmentation as it requires both gold labels and teacher predictions.

As an alternative, we propose a mechanism to filter-out noisy examples in the augmented dataset based on uncertainty quantification. Recent work has shown that ensembles produce accurate uncertainty estimates (Lakshminarayanan et al., 2017). We exploit this idea by training a set of additional teacher models independently on the same training data using random initialisation, and using the variance of their predictions as an indicator of predictive uncertainty.[2] Intuitively, examples with very high variance would correspond to noisy teacher predictions. We filter out from the student training data the instances where the variance is more than one standard deviation away from its mean value. This is expected to have a higher impact on the results in the out-of-domain setting where the performance of the teacher model is less stable and teacher predictions can contain more noise.

## 3 Experiments

**MLQE Dataset.** For training the teacher and for evaluation, we use the MLQE dataset (Fomicheva et al., 2020), same as in the WMT2020 QE Shared Task (Specia et al., 2020). This dataset contains sentences extracted from Wikipedia translated to and from English for a total of six language pairs: English–German (En-De),[3] English–Chinese (En-Zh), Romanian–English (Ro-En), Estonian–English (Et-En), Sinhala–English (Si-En) and Nepali–English (Ne-En). Each translation was produced with a SoTA Transformer-based NMT model and manually annotated for quality using an annotation scheme inspired by the Direct Assessment methodology (Graham et al., 2013). The scores are produced on a continuous scale indicating perceived translation quality in 0-100. For each language pair, this dataset contains partitions for training (7K), dev (1K), and test (1K).

| Language | Sentences |
|---|---|
| Estonian | 25,176 |
| Romanian | 372,690 |
| Sinhala | 139,406 |
| Nepalese | 85,343 |
| English | 1,563,519 |

Table 2: Number of sentences extracted from Wikipedia for data augmentation.

**Distilled dataset.** Monolingual data for data augmentation was sampled from Wikipedia following the procedure described in Fomicheva et al. (2020) to preserve the domain of the MLQE dataset. Specifically, we sampled documents from Wikipedia for English, Estonian, Romanian, Sinhalese and Nepalese and 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 and (ii) have the length between 50 and 150 characters. Table 2 shows the total amount of sentences in the monolingual Wikipedia dataset collected for data augmentation.

To test the impact of data domain on the performance of the student QE models, we also collect out-of-domain data for the Et-En language pair. The out-of-domain data is sampled from Common Crawl. We use the version of Common Crawl distributed by the WMT2018 News Translation Task[4]. The total amount of sentences in this dataset is 100,779,314.

To translate the data, we used the same MT models that generated the test data, built with fairseq (Ott et al., 2019) and made available by the WMT2020 QE Shared Task organisers.[5] Sentences that were part of the training data for the MT models or part of the MLQE dataset were excluded. We generate quality predictions for the remaining sentences using the teacher models, as

[2]Here we use ensemble only as a way of estimating the error in the predictions and leave distillation based on ensemble predictions to future work.

[3]We skip this language pair as the performance of the teacher model for it is too weak.

[4]http://www.statmt.org/wmt18/translation-task.html

[5]https://github.com/facebookresearch/mlqe/tree/master/nmt_models.

| Name | Training data | Et-En | Ro-En | Si-En | Ne-En | En-Zh |
|---|---|---|---|---|---|---|
| $\text{TQ}_{\text{TEACHER}}$ | MLQE-gold | 0.77 | 0.88 | 0.60 | 0.75 | 0.44 |
| $\text{BiRNN}_{\text{STUDENT}}$ | MLQE-dist | 0.45 | 0.62 | 0.44 | 0.46 | **0.18** |
| $\text{BiRNN}_{\text{STUDENT+AUG}}$ | Wiki-dist | **0.50** | **0.69** | **0.45** | **0.54** | 0.17 |
| BiRNN | MLQE-gold | 0.37 | 0.60 | 0.40 | 0.42 | 0.15 |
| Predictor-Estimator | MLQE-gold | 0.48 | 0.69 | 0.37 | 0.39 | 0.19 |
| $\text{TQ}_{\text{DistilBERT}}$ | MLQE-gold | 0.62 | 0.78 | 0.51 | 0.61 | 0.36 |

Table 3: Pearson correlation with human judgments on the MLQE test set. MLQE-gold: training partition of MLQE dataset; MLQE-dist: distilled version of the MLQE training set with teacher predictions used as labels; Wiki-dist: the Wikipedia dataset produced by data augmentation. Boldface results indicate our best student models.

described in Section 2. We used a random subset of 100K sentences from Wikipedia to train the student model for each of the language pairs except for Et-En where the total amount of collected in-domain monolingual data is 25K.

**Models.** As teachers, we use pre-trained models from TransQuest ($\text{TQ}_{\text{TEACHER}}$), one of the winning submissions in the WMT2020 QE Shared Task, which we fine-tuned on the MLQE dataset. For noise filtering, we train five teacher models with random initialisation. As students, we use BiRNN models from DeepQuest (Ive et al., 2018). We also compare our results against the Predictor-Estimator model (Kim et al., 2017; Kepler et al., 2019), the baseline at the WMT2020 QE Shared Task, and TransQuest models using multilingual DistilBERT.[6]

## 4 Results

Table 3 shows the Pearson correlation with human judgments on the test partition of the MLQE dataset for different models and specifies the type of data used for training.[7] The correlation for the student models ($\text{BiRNN}_{\text{STUDENT}*}$) does not reach the performance of $\text{TQ}_{\text{TEACHER}}$. Smaller models may lack representation power for modeling cross-lingual tasks such as QE. Also, distillation for regression is more challenging, as discussed in Section 2. However, **training on the in-domain distilled data** ($\text{BiRNN}_{\text{STUDENT+AUG}}$) **allows to obtain performances comparable to DistilBERT** ($\text{TQ}_{\text{DistilBERT}}$) **with much lighter models** (see Table 1).[8] Furthermore, this approach results in a substantial improvement over shallow models trained on gold data (BiRNN and Predictor-Estimator) for all of the language pairs. The student performance for each language pair is strongly related to the performance of the teacher. Thus, the Ro-En student achieves the highest correlation results, whereas correlation for En-Zh is weak.

| | Ro-En | Si-En | Ne-En | En-Zh |
|---|---|---|---|---|
| 10K | 0.56 ±.00 | 0.36 ±.00 | 0.41 ±.00 | 0.09 ±.01 |
| 50K | 0.64 ±.00 | 0.45 ±.01 | 0.53 ±.00 | 0.20 ±.03 |
| 70K | 0.66 ±.00 | 0.46 ±.01 | 0.54 ±.00 | 0.19 ±.02 |
| 100K | 0.69 ±.00 | 0.47 ±.02 | 0.54 ±.00 | 0.17 ±.02 |

Table 4: Pearson correlation on the test partition of the MLQE dataset for BiRNN student models trained with different amounts of distilled Wikipedia data.

We further analyse what is the impact of different data selection strategies on the results. **First**, we sample random subsets of training instances from the Wikipedia distilled dataset and evaluate the performance of the student model trained with this data. We run the training 3 times with different random splits for training and validation and report the mean and confidence intervals. Table 4 shows the results for all languages where we have enough Wikipedia data (for Et-En we only have 25K in total). The largest boost in correlation is observed when going from 10K to 50K.

**Second**, we compare student models trained on these subsets of distilled data of different sizes, i.e. using data extracted from Wikipedia (in-domain), against data splits of the same size extracted from Common Crawl (out-of-domain). For the out-of-domain data we apply the noise filtering strategy described in Section 2. Figure 2 shows the results

[6] Multilingual DistilBERT is available at https://huggingface.co/distilbert-base-multilingual-cased. We follow the same training procedure as for the teacher model described in detail in Appendix A.

[7] $\text{TQ}_{\text{TEACHER}}$, $\text{TQ}_{\text{DistilBERT}}$ and Predictor-Estimator use contextual representations trained on large amounts of additional data, which are then fine-tuned for the QE task.

[8] This is true for all language pairs except Et-En and En-Zh. For Et-En we have a considerably smaller amount of in-domain data available for training, whereas for En-Zh the teacher model appears to be too weak to be useful for KD.

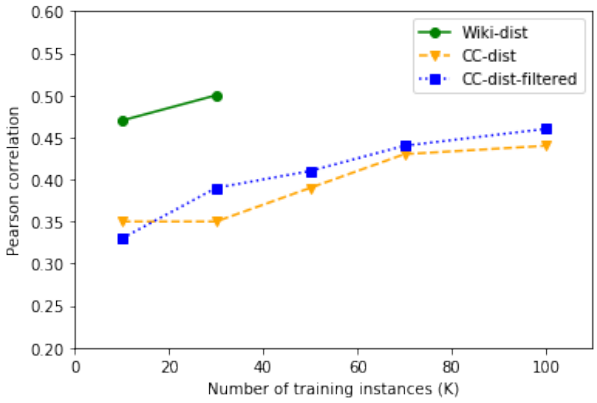


Figure 2: Pearson correlation results on the MLQE test set for the student models trained with different amounts of distilled data in-domain (Wiki-dist), out-of-domain (CC-dist) and out-of-domain with noise filtering (CC-dist-filtered), for Et-En.

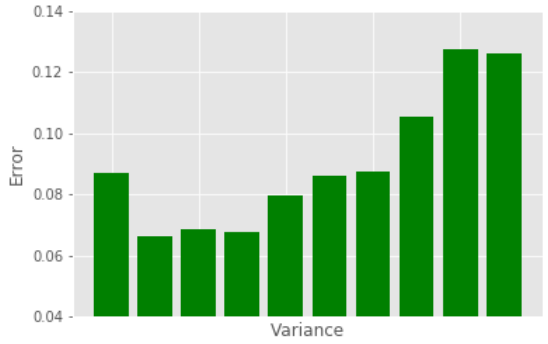


Figure 3: Variance in the predictions of 6 teacher models trained with different random seed against predictions error on the test partition of Ro-En MLQE dataset.

for Et-En, where our largest in-domain set has 25K sentences. We observe that using in-domain data appears to be much more effective than sampling larger amounts of generic data. Noise filtering gives some improvement in the results but its effect appears to be marginal compared to the effect of training with in-domain data.

Figure 3 provides an illustration of the relation between the variance in the predictions of multiple teacher models and prediction error for Ro-En language pair on the in-domain data. We group the sentences in the test partitions of MLQE dataset in 10 bins according to the variance between the predictions of the different teacher models in the ensemble. We then calculate the average prediction error in each bin, where the error is the absolute difference between model predictions and human judgements. As shown in Figure 3, higher variance

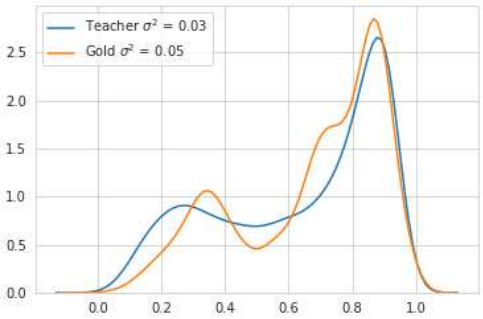


Figure 4: Distribution of teacher scores (blue) and gold labels (orange) on the training partitions of Et-En MLQE dataset.

in the predictions indeed corresponds to larger prediction error.

Interestingly, from Table 3 we see that **training with distilled data brings benefits even without data augmentation** for some of the language pairs. The correlation for Et-En improves from 0.37 to 0.45 by training on teacher predictions (BiRNN$_{\text{STUDENT}}$) instead of gold labels (BiRNN) on the same MLQE dataset. To gain an intuition for this improvement, Figure 4 shows the distribution of teacher predictions and human scores on the train partition of MLQE dataset. We hypothesize that teacher predictions having a smoother distribution with reduced variance makes learning easier. As shown in Appendix B, we observe this trend for all language pairs in the dataset.

## 5 Conclusions

In this paper, we showed that knowledge distillation, through a teacher-student approach that directly distills QE predictions, can be effective in building a light-weight QE model with similar performance to a SoTA architecture trained on distilled yet large pre-trained representations. We also introduced a noise filtering approach that leverages the uncertainty of an ensemble of teacher models to determine which training instances should be discarded when training the student models, which can be beneficial especially for data augmentation from out-of-domain sources. This results in QE models 4x smaller in disk space with 8x fewer parameters, and 3x faster in inference speed.

## Acknowledgements

This work was supported by funding from the Bergamot project (EU H2020 Grant No. 825303).

## References


Guobin Chen, Wongun Choi, Xiang Yu, Tony X. Han, and Manmohan Chandraker. 2017. Learning efficient object detection models with knowledge distillation. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 742–751.

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

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020b. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 8440–8451. 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*.

Marina Fomicheva, Shuo Sun, Lisa Yankovskaya, Frédéric Blain, Francisco Guzmán, Mark Fishel, Nikolaos Aletras, Vishrav Chaudhary, and Lucia Specia. 2020. Unsupervised quality estimation for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:539–555.

Tommaso Furlanello, Zachary Lipton, Michael Tschannen, Laurent Itti, and Anima Anandkumar. 2018. Born again neural networks. In *International Conference on Machine Learning*, pages 1607–1616. PMLR.

Jianping Gou, Baosheng Yu, Stephen John Maybank, and Dacheng Tao. 2020. Knowledge distillation: A survey. *arXiv preprint arXiv:2006.05525*.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 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.

Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. 2015. Distilling the knowledge in a neural network. *CoRR*, abs/1503.02531.

Julia Ive, Frédéric Blain, and Lucia Specia. 2018. DeepQuest: a framework for neural-based quality estimation. In *Proceedings of COLING 2018, the 27th International Conference on Computational Linguistics: Technical Papers*, Santa Fe, new Mexico.

Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. 2020. Tinybert: Distilling BERT for natural language understanding. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings, EMNLP 2020, Online Event, 16-20 November 2020*, pages 4163–4174. Association for Computational Linguistics.

Fábio Kepler, Jonay Trénous, Marcos Treviso, Miguel Vera, and André F. T. Martins. 2019. OpenKiwi: An open source framework for quality estimation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics–System Demonstrations*, pages 117–122, Florence, Italy. Association for Computational Linguistics.

Hyun Kim, Jong-Hyeok Lee, and Seung-Hoon Na. 2017. Predictor-estimator using multilevel task learning with stack propagation for neural quality estimation. In *Proceedings of the Second Conference on Machine Translation*, pages 562–568.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. 2017. Simple and Scalable Predictive Uncertainty Estimation Using Deep Ensembles. In *Advances in Neural Information Processing Systems*, pages 6402–6413.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*.

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*.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

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

Siqi Sun, Yu Cheng, Zhe Gan, and Jingjing Liu. 2019. Patient knowledge distillation for BERT model compression. 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 2019, Hong Kong, China, November 3-7, 2019*, pages 4322–4331. Association for Computational Linguistics.

Makoto Takamoto, Yusuke Morishita, and Hitoshi Imaoka. 2020. An efficient method of training small models for regression problems with knowledge distillation. In *3rd IEEE Conference on Multimedia Information Processing and Retrieval, MIPR 2020, Shenzhen, China, August 6-8, 2020*, pages 67–72. IEEE.

Jiaxi Tang, Rakesh Shivanna, Zhe Zhao, Dong Lin, Anima Singh, Ed H Chi, and Sagar Jain. 2020. Understanding and improving knowledge distillation. *arXiv preprint arXiv:2002.03532*.

Raphael Tang, Yao Lu, Linqing Liu, Lili Mou, Olga Vechtomova, and Jimmy Lin. 2019. Distilling task-specific knowledge from BERT into simple neural networks. *CoRR*, abs/1903.12136.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

## A Teacher and student models

**Teacher model** For the teacher model, we use the transformer-based MonoTransQuest (Ranasinghe et al., 2020) architecture of the TransQuest framework with XLMR-large (Conneau et al., 2020b) as the underlying pre-trained representation model. XLMR is Transformer (Vaswani et al., 2017) based masked language model (with 24 Transformer blocks and the vocabulary size of 250K) trained on one hundred languages using approximately two terabytes of CommonCrawl data. MonoTransQuest takes as input the concatenation of the original sentence and its translation separated by the the special $[SEP]$ token. The learned representation of the special $[CLS]$ token is considered as the joint representation of the original and translated sentence. The joint representation is then fed to the final $softmax$ layer to predict the quality score of the translation. For distillation we use the models that were made available for download by the authors.[9] For training the additional teacher models for data filtering we follow the training settings in Ranasinghe et al. (2020): we used a batch size of 8, Adam optimiser with learning rate 2e-5, and a linear learning rate warm-up over 10% of the training data. During the training process, the parameters of XLM-R model, as well as the parameters of the subsequent layers, are updated. All the models were trained for 3 epochs.

**Student model** For the student model, we rely on the BiRNN QE architecture proposed by Ive et al. (2018). Our implementation of this architecture is available for download.[10] The light-weight architecture (15 layers) of this model is as follows: both source and target sentences are independently encoded by a dedicated embedding layer followed by a bi-directional Recurrent Neural Network (RNN). The two resulting sentence representations are then concatenated as a weighted sum of their word vectors, generated by an attention mechanism. The resulting representation is then passed through an output dense layer with sigmoid activation to generate the quality estimates. We use the BiRNN model in its default configuration: both source and target embeddings are of size 300, each encoder has a hidden size of 50. The vocabulary size is limited to the 30k most common words. The model is trained with early stopping with a patience of 5.

[9] https://tharindudr.github.io/TransQuest/pretrained/#available-models

[10] https://github.com/sheffieldnlp/deepQuest-py

## B Output Distribution for Teacher Models

Figure 5 shows the distribution of teacher scores (blue) and gold labels (orange) on the training partitions for the language pairs in MLQE dataset.

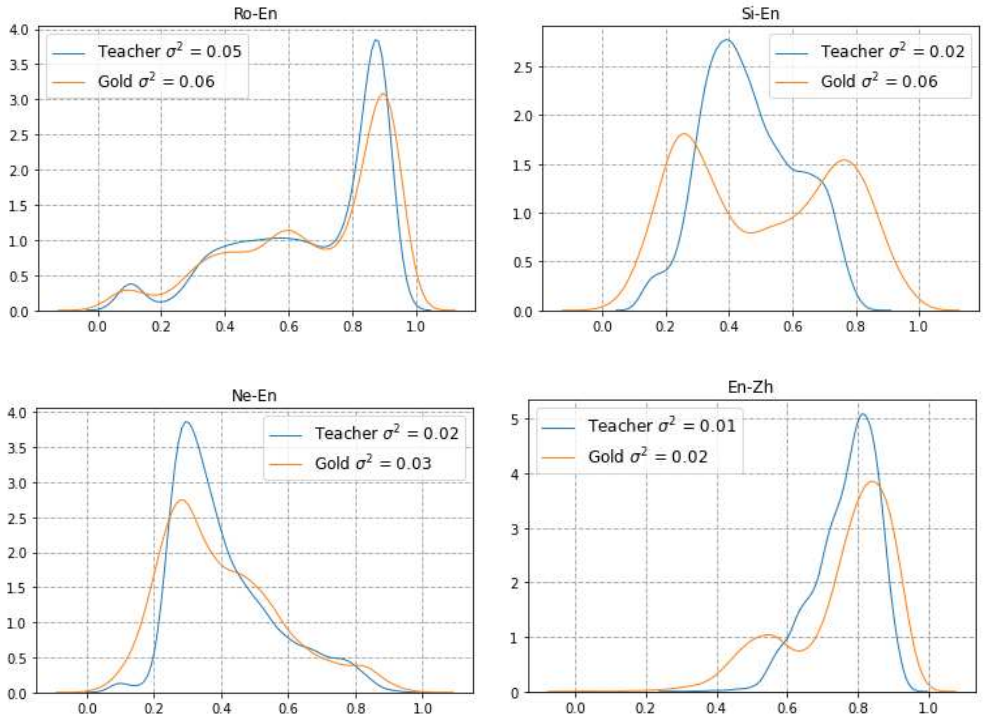


Figure 5: Distribution of teacher scores (blue) and gold labels (orange) on the training partitions for different language pairs in the MLQE dataset.