IST-Unbabel 2021 Submission for the Quality Estimation Shared Task

Chrysoula Zerva[1,2,*] Daan van Stigt[3,*] Ricardo Rei[2,3,4,*] Ana C. Farinha[3]
Pedro G. Ramos[3] José G. C. de Souza[3] Taisiya Glushkova[1,2] Miguel Vera[3]
Fábio Kepler[3] André F. T. Martins[1,2,3]
[1]Instituto Superior Técnico [2]Instituto de Telecomunicações [3]Unbabel [4]INESC-ID
[2,3,4]Lisbon, Portugal
{chrysoula.zerva, ricardo.rei, taisiya.glushkova, andre.t.martins}@tecnico.ulisboa.pt
{daan.stigt, catarina.farinha, pedro.ramos, jose.souza, miguel.vera, fabio.kepler}@unbabel.com

Abstract

We present the joint contribution of IST and Unbabel to the WMT 2021 Shared Task on Quality Estimation. Our team participated on two tasks: Direct Assessment and Post-Editing Effort, encompassing a total of 35 submissions. For all submissions, our efforts focused on training multilingual models on top of OpenKiwi predictor-estimator architecture, using pre-trained multilingual encoders combined with adapters. We further experiment with and uncertainty-related objectives and features as well as training on out-of-domain direct assessment data.

1 Introduction

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., 2018). This paper describes the joint contribution of Instituto Superior Técnico (IST) and Unbabel to the WMT21 Quality Estimation shared task (Specia et al., 2021), where systems were submitted to two tasks: 1) sentence-level direct assessment; 2) word- and sentence-level post-editing effort.

This year's submission combines several ideas built on top of the OpenKiwi framework. Motivated by the mixture of *blind* and *seen* language pairs in the test sets, we experimented with extensions that would allow us to train multilingual models that maintain good generalization ability and are robust to the presence of epistemic and aleatoric uncertainty.

For both tasks we trained and submitted an ensemble of multilingual models. All submitted models follow the predictor-estimator architecture (Kim and Lee, 2016; Kim et al., 2017) and use pre-trained models for feature extraction. Also, we fine-tune all models on the provided QE data using stacked adapter layers (Pfeiffer et al., 2020). We show that we can thus achieve comparable performance across language pairs while minimising the number of trainable parameters (see Table 1). Furthermore, we experimented with different types of uncertainty-related information to leverage it's benefits, improving performance and robustness of the submitted systems (see §3.1.1). All related code extensions will be publicly available.

Our main contributions are:

- We build on our OpenKiwi architecture by exploring adapter layers (Houlsby et al., 2019; Pfeiffer et al., 2020) for quality estimation as these demonstrated to be less amenable to overfitting while presenting the same or superior quality performance than fine-tuning the whole base pre-trained model for different NLP tasks (He et al., 2021).
- We incorporate different types of uncertainty into our architectures. We make use of the glass-box features (Fomicheva et al., 2020) extracted from the NMT models, the *aleatoric* (data) uncertainty derived from the human annotations and the *epistemic* (model) uncertainty (Hora, 1996; Kiureghian and Ditlevsen, 2009; Huellermeier and Waegeman, 2021) that originates from the QE model.
- We show that training the QE models on additional out-of-domain direct assessment (DA) data gives considerable gains in performance for the new language pairs from the *blind* test sets.

2 Quality Estimation Tasks

In this year's shared task edition we submitted models for the first two tasks:

1. Task 1: sentence-level direct assessment
2. Task 2: word- and sentence level post-editing effort, comprising of two subtasks: a) predicting the HTER score of the translated sentence

[*] The first three authors have equal contribution.

Proceedings of the Sixth Conference on Machine Translation (WMT), pages 961–972
November 10–11, 2021. ©2021 Association for Computational Linguistics

(hypothesis); and b) predicting OK/BAD tags for the words and gaps (both in source and translation)

We note that this year, both tasks 1 and 2 provided additional *blind* test sets with language pairs that were not included in the data made available for training/development, providing an interesting challenge and motivating multilingual and generalisable approaches.

3 Implemented Systems

3.1 Task 1

For Task 1 our final submission consisted of an ensemble of two different multilingual models, that differ in the way they process the input source (original sentence) and hypothesis (machine translation). Both models are based on the predictor-estimator architecture, using different pre-trained models to extract features and different training approaches to optimise for the QE task.

The key idea explored with our first model (denoted by M1 variations in the experiments), revolved around pursuing highly generalisable multilingual models, robust to overfitting. To this end, we train a cross-lingual transformer (XLM-RoBERTa (Conneau et al., 2020)) on large, multilingual data with direct assessments and then use adapters (Houlsby et al., 2019; Pfeiffer et al., 2020) to adapt to the domain specific data of the QE task with minimal training effort. In line with our efforts for good generalisation, we use only task-specific adapters and refrain from using specific adapters for each language pair. For these experiments we build on the OpenKiwi architecture (Kepler et al., 2019), using a pre-trained xlm-roberta-large encoder as a feature predictor. The source and hypothesis sentences are jointly encoded with hypothesis first. Then, source and hypothesis features are generated using average pooling over the hypothesis embeddings and forwarded to the estimator module which corresponds to a feed-forward layer. Figure 1 provides the general architecture[1]

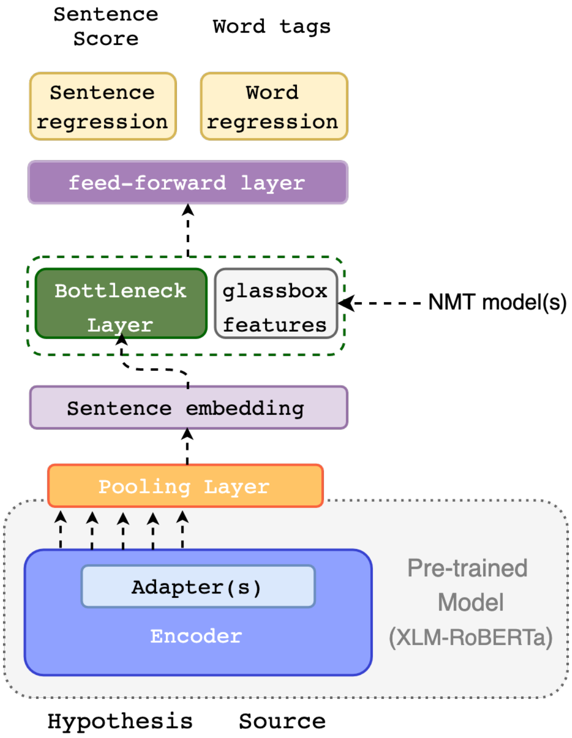

Figure 1: General architecture of M1 model variations. Word tag prediction is used only for Task 2.

The model was first trained on the direct assessment data provided in the Metrics shared tasks (Mathur et al., 2020), as described in §3.1.2. Upon training, the XML-R encoder is frozen and the the model is fine-tuned on sentence regression with the task-specific data, using stacked adapters. We hence manage to maintain a low number of trainable parameters during fine-tuning and minimize training time while learning to predict task-specific sentence scores.

For the second model (denoted by M2-KL-G-MCD) we aimed to explore the potential of a large pre-trained multilingual model (trained with MT objectives). We use the mBART (Liu et al., 2020) encoder-decoder architecture to encode the source and force-decode the hypothesis. We specifically use the mBART50 model (Tang et al., 2020) which is trained with multilingual finetuning on 50 languages, including all languages of interest for the QE 2021 task. We obtain the features by averaging the decoder embeddings and concatenating with the <eos> token of the sequence. The estimator part of the model consists of a *bottleneck* feed-forward layer that reduces the dimensionality of the decoder output, and is concatenated with a vector with additional glass-box features from the NMT models (see §3.1.1). The combined vector is then forwarded to a feed-forward estimator and the full model is fine-tuned on the task specific QE data. Apart from the glass-box features we experimented further with methods that allow the model to be

[1] Note that glass-box features are integrated but not used in this submission as they did not significantly improve performance.

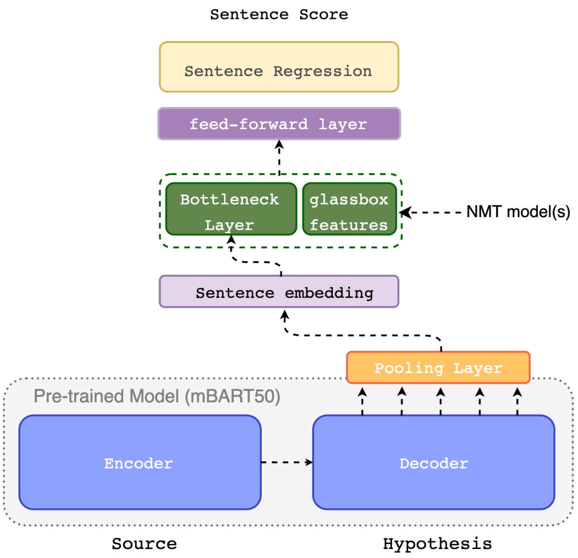

Figure 2: General architecture of M2 model variations.

more robust towards the underlying uncertainty of its predictions. We elaborate that in the next section. Figure 2 provides a general architecture of the M2 model variations.

3.1.1 Learning from uncertainty

Multiple neural models are involved in the process of obtaining and scoring machine translations, which naturally leads to several sources of uncertainty. These sources can be very informative and useful for MT evaluation. In this work we try to consider three types of uncertainty: (1) uncertainty of the NMT models used to obtain the *hypotheses*, (2) data (aleatoric) uncertainty for which we use the inter-annotator disagreement as a proxy, and (3) uncertainty of the MT evaluation model itself.

NMT model uncertainty The idea of extracting uncertainty-related features from the MT systems in order to estimate the quality of their predictions, was originally introduced by Fomicheva et al. (2020). This glass-box approach to QE is mostly focusing on capturing epistemic uncertainty, and the proposed features are extracted either using Monte Carlo (MC) dropout on the NMT or using the output probability distributions obtained from a standard deterministic MT system. In our last year's submission (Moura et al., 2020) the integration of such features proved to be effective, thus we decided to incorporate it into our new model as well. We list the extracted features below:

- `TP` sentence average of word translation probability
- `Softmax-Ent` sentence average of softmax output distribution entropy
- `Sent-Std` sentence standard deviation of word probabilities
- `D-TP` average TP across N(N = 30) stochastic forward-passes
- `D-Var` variance of TP across N stochastic forward-passes
- `D-Combo` combination of D-TP and D-Var defined by $1 - D - TP/D - Var$
- `D-Lex-Sim` lexical similarity - measured by METEOR score (Lavie and Denkowski, 2009) - of MT output generated in different stochastic passes.

Aleatoric uncertainty The noise and complexity of the training data is a source of predictive uncertainty in itself, referred to as data or aleatoric uncertainty (Kiureghian and Ditlevsen, 2009). This uncertainty is often reflected in the disagreement between human annotations for the same *source-hypothesis* segment (Cohn and Specia, 2013; Fornaciari et al., 2021). We hypothesize that the direct assessments can be better modelled as normally distributed scores rather than a single score, and that a model trained to predict this distribution (mean and standard deviation) could provide better quality estimates [2]. We formalise this as a KL divergence objective, using the closed form solution to estimate the KL divergence between the target distribution $p(x) = N(\mu_1, \sigma_1)$ and the predicted distribution $q(x) = N(\mu_2, \sigma_2)$, as shown in Eq. 1.

$$KL(p||q) = \log \frac{\sigma_2}{\sigma_1} + \frac{\sigma_1^2 + (\mu_1 - \mu_2)^2}{2\sigma_2^2} - \frac{1}{2} \quad (1)$$

where we take the mean and standard deviation (std) of the direct assessment z_scores as the target (ground truth proxy) values p. This way, we account for the annotator disagreement (reflected in the std value) during learning.

QE epistemic uncertainty We use MC dropout (Gal and Ghahramani, 2016) to account for the uncertainty of the QE model. Specifically, we enable dropout during inference and run multiple forward runs over each test instance. Thus we obtain a distribution of quality predictions for each instance

[2] Note that for this task's data we only had access to 3 scores per segment so the mean and std values are calculated over these numbers.

instead of a single point estimate. We use the estimated mean of the distribution as our predicted quality estimate. MC dropout has been shown to improve predictive accuracy and perform on par or even better compared to deep ensembles for MT evaluation tasks (Glushkova et al., 2021). It thus allows us to simulate ensembling in a cheap and effective way, without the need to train multiple checkpoints.

3.1.2 Out-of-domain direct assessment data

The QE data is relatively limited, making it harder to train multilingual models with a large number of parameters without over-fitting. Thus, as explained in §3.1 we aimed to investigate whether we could obtain models that generalise better and are more robust to noise and out-of-distribution data by training the XLM-RoBERTa model first on a larger –yet noisier and out-of-domain dataset. To that end we leverage the data provided for the past Metrics shared tasks, which covers the language pairs used in this year's QE task, including the blind tests for which we had no in-domain data available. Altogether, it encompasses 30 language pairs from the news domain (versus 7 in the QE dataset). We provide more detailed statistics for each language pair of the Metrics data in Appendix C. We refer to experiments using the model initially trained on the Metrics data as M1M-. We also show that using the trained XLM-RoBERTa encoder from the M1M model can prove beneficial for the predictions on post-edited data of Task 2 (see Table 3).

3.2 Task 2

For Task 2 we submitted an ensemble of two variations of the first model (M1-ADAPT and M1M-ADAPT) presented for Task 1 (see §3.1). In both cases, we use multi-task training and a feed-forward for each output types: hypothesis word tags, hypothesis gap tags, source word tags, and sentence regression (on HTER scores). Both variations use a pre-trained XLM-RoBERTa (`large`) encoder to extract features as described for Task 1, but differ in the training of the encoder. In the first case we use the pre-trained model [3] and fine-tune on the QE data using stacked adapters. In the second variation we swap the original pre-trained model with the XLM-RoBERTa model that has been trained on the Metrics data as described in §3.1.2. We note that the two variations favor different language pairs, hence we combine multiple checkpoints from each variation (ranging training steps). We use the `test-20` split of the data to optimise the hyper-parameters and following this approach we use the estimated top-3 checkpoints from each variation using the combined dataset [4] and the top checkpoint for the non-augmented model trained exclusively on the train set, resulting in total 7 checkpoints in our final ensemble.

4 Experimental Results

We present the performance of the implemented models on the `test-20` dataset.

4.1 Task 1

The results can be seen in Tables 1 and 2. In line with the shared task guidelines we treat Pearson r as the primary performance metric and select the submitted models accordingly. We can observe, that while on average the M1 model and its variations outperform the M2 model, their performance is comparable, and M2-KL-G-MCD can even outperform M1M-ADAPT for specific language pairs, hence it made sense to combine them in the final ensemble. We can also see that fine-tuning the M1 model on the Metrics data, results in performance gains for the majority of the language pairs. Specifically, even applying the M1M directly, without further fine-tuning on QE data, achieves competitive performance for most pairs, which further improves upon fine-tuning. It helps in increasing the performance on the *blind* sets (denoted as *zero-shot* in the Appendix B). The performance gains concern mostly the correlation performance indicators (Pearson and Spearman correlations), since especially for M1 the error-based indicators (MAE and RMSE) seem to favor the versions of the model that have not seen the Metrics data. One possible explanation for this discrepancy could lie in the differences between the range and distribution of DA scores for the two datasets. Indicatively, the range of scores on the `train-dev-test-20` concatenation of the QE data is $[-7.542, 3.178]$ and for the Metrics data $[-8.624, 4.332]$. The target DA scores in both datasets are calculated via standardizing (taking the z score) the direct assessments for each annotator and then averaging all standardized

[3] `https://huggingface.co/transformers/model_doc/xlmroberta.html`

[4] The combined dataset in this case refers to the concatenation of the train/dev/test20 annotated data splits provided for the shared task

		Pears↑	Spear↑	MAE↓	RMSE↓
EN-DE	M1 BASE	0.4534	0.4532	0.4482	0.6371
	M1-ADAPT	0.5092	0.4825	0.4868	0.6288
	M1M	0.5288	0.4872	0.4485	0.6327
	M1M-ADAPT	0.5695	0.5131	0.4127	0.6095
EN-ZH	M1 BASE	0.4429	0.4362	0.5364	0.6867
	M1-ADAPT	0.4723	0.4755	0.5228	0.6714
	M1M	0.4447	0.4400	0.4772	0.6110
	M1M-ADAPT	0.4815	0.4872	0.5502	0.7017
ET-EN	M1 BASE	0.7939	0.8076	0.5388	0.6928
	M1-ADAPT	0.7948	0.8061	0.4518	0.5810
	M1M	0.7580	0.7611	0.5820	0.7134
	M1M-ADAPT	0.7956	0.8110	0.5358	0.6921
NE-EN	M1 BASE	0.7805	0.7592	0.4278	0.5461
	M1-ADAPT	0.7609	0.7475	0.4075	0.5393
	M1M	0.7477	0.7324	0.4499	0.6161
	M1M-ADAPT	0.7888	0.7556	0.4192	0.5332
RO-EN	M1 BASE	0.8718	0.8360	0.3598	0.4878
	M1-ADAPT	0.8923	0.8533	0.3068	0.4201
	M1M	0.8345	0.8132	0.4585	0.5863
	M1M-ADAPT	0.8889	0.8488	0.3142	0.4437
RU-EN	M1 BASE	0.7587	0.6919	0.4885	0.6949
	M1-ADAPT	0.7736	0.7142	0.4138	0.6082
	M1M	0.6703	0.6535	0.5606	0.7583
	M1M-ADAPT	0.7425	0.7159	0.4989	0.7250
SI-EN	M1 BASE	0.6456	0.6112	0.5060	0.6481
	M1-ADAPT	0.6613	0.6172	0.4742	0.5939
	M1M	0.6308	0.6535	0.4742	0.5786
	M1M-ADAPT	0.6649	0.6225	0.4863	0.6064
ML	M1 BASE	0.6781	0.6565	0.4722	0.6276
	M1-ADAPT	0.6949	0.6709	0.4377	0.5775
	M1M	0.6593	0.5131	0.4127	0.6095
	M1M-ADAPT	**0.7045**	**0.6791**	**0.4596**	**0.6160**

Table 1: Results for Task 1 with the M1 predictor-estimator (XLM-RoBERTa) and different training/fine-tuning approaches. M1M is the M1 model trained on the Metrics dataset and M#-ADAPT signifies a model fine-tuned on the QE data with adapters. ML stands for MULTILINGUAL, showing the performance averaged over all language pairs. Underlined numbers indicate the best result for each language pair and evaluation metric. **Bold** systems were selected for the final ensemble.

assessments for each segment. Thus, the difference in target score range and distribution could affect the magnitude of predicted scores and the distance to the ground truth values, which is reflected in the MAE and RMSE metrics. These findings, further supported by the results on Task 2, is a first step in exploring the underlying connection and bridging the gap between the Metrics and Quality Estimation shared tasks.

4.2 Task 2

The results can be seen in Table 3. Similarly to Task 1, the primary evaluation metric for the sentence level sub-task of Task 2 is the Pearson r coefficient, while the word level sub-task is evaluated using the Matthews correlation coefficient (MCC, (Matthews, 1975)) as the primary performance indicator.

		Pears↑	Spear↑	MAE↓	RMSE↓
EN-DE	M2 BASE	0.4889	0.4645	0.4608	0.6180
	M2-KL	0.4971	0.4769	0.4549	0.6191
	M2-KL-G	0.5110	0.4738	0.4396	0.6133
	M2-KL-G-MCD	0.5093	0.4754	0.4495	0.6128
EN-ZH	M2 BASE	0.4484	0.4355	0.4940	0.6374
	M2-KL	0.4574	0.4471	0.5042	0.6485
	M2-KL-G	0.4566	0.4543	0.5278	0.6751
	M2-KL-G-MCD	0.4628	0.4584	0.4973	0.6390
ET-EN	M2 BASE	0.7792	0.7842	0.4581	0.5624
	M2-KL	0.7833	0.7896	0.4684	0.5824
	M2-KL-G	0.7847	0.7962	0.4643	0.5924
	M2-KL-G-MCD	0.7868	0.7951	0.4539	0.5674
NE-EN	M2 BASE	0.7333	0.7154	0.4347	0.5531
	M2-KL	0.7638	0.7393	0.4040	0.5247
	M2-KL-G	0.7529	0.7228	0.4194	0.5353
	M2-KL-G-MCD	0.7596	0.7269	0.4125	0.5313
RO-EN	M2 BASE	0.8780	0.8407	0.3403	0.4514
	M2-KL	0.8826	0.8406	0.3199	0.4305
	M2-KL-G	0.8728	0.8397	0.3314	0.4635
	M2-KL-G-MCD	0.8777	0.8429	0.3209	0.4426
RU-EN	M2 BASE	0.7406	0.6874	0.4696	0.6381
	M2-KL	0.7532	0.7123	0.4558	0.6299
	M2-KL-G	0.7485	0.7191	0.4630	0.6612
	M2-KL-G-MCD	0.7509	0.7204	0.4492	0.6358
SI-EN	M2 BASE	0.6243	0.5899	0.4709	0.5939
	M2-KL	0.6373	0.6000	0.4572	0.5726
	M2-KL-G	0.6506	0.6168	0.4586	0.5796
	M2-KL-G-MCD	0.6545	0.6199	0.4495	0.5697
ML	M2 BASE	0.6704	0.6454	0.4469	0.5792
	M2-KL	0.6821	0.6580	0.4378	0.5725
	M2-KL-G	0.6825	0.6604	0.4434	0.5886
	M2-KL-G-MCD	**0.6859**	**0.6627**	**0.4333**	**0.5712**

Table 2: Results for Task 1 with the M2 predictor-estimator (mBART) and different uncertainty handling additions. "KL" signifies the incorporation of KL loss, "G"the incorporation of glass-box features and MCD the addition of MC dropout. ML stands for MULTILINGUAL, showing the performance averaged over all language pairs. Underlined numbers indicate the best result for each language pair and evaluation metric. **Bold** systems were selected for the final ensemble.

We can see that while HTER scores do not always correlate highly with DAs (see Table 4), the use of the M1M model encoder that was trained on large data with direct assessments can still prove useful. Indeed, when fine-tuning on the Task2 data, the model using the M1M encoder (M1M-ADAPT in the table 3) provides a performance boost for the Pearson correlation in most language pairs, and competitive performance for the rest. Based on these results, we deem it worthwhile to include checkpoints trained with this configuration in the ensemble estimating that they will contribute in higher performance, especially on the blind test sets. This can be further confirmed when

		Pearson↑	SRC-MCC↑	TGT-MCC↑
En-De	M1 base	0.5256	0.3331	0.4092
	M1-adapt	0.5573	0.4211	0.36454
	M1M-adapt	0.5499	0.3647	0.4239
En-Zh	M1 base	0.3786	0.3253	0.3589
	M1-adapt	0.3711	0.4346	0.3288
	M1M-adapt	0.3721	0.4255	0.3643
Et-En	M1 base	0.7319	0.4537	0.5110
	M1-adapt	0.7360	0.5545	0.4978
	M1M-adapt	0.7498	0.4929	0.5513
Ne-En	M1 base	0.5898	0.5198	0.4386
	M1-adapt	0.5987	0.6884	0.5426
	M1M-adapt	0.6252	0.4244	0.4682
Ro-En	M1 base	0.8531	0.5727	0.6190
	M1-adapt	0.8282	0.5984	0.5653
	M1M-adapt	0.8280	0.5682	0.5813
Ru-En	M1 base	0.4899	0.2766	0.3213
	M1-adapt	0.4811	0.341	0.3071
	M1M-adapt	0.5060	0.2927	0.3421
Si-En	M1 base	0.6659	0.4653	0.4776
	M1-adapt	0.6698	0.6776	0.5057
	M1M-adapt	0.6935	0.3872	0.4937
ML	M1 base	0.6050	0.4209	0.4479
	M1-adapt	**0.6061**	**0.5323**	**0.4445**
	M1M adapt	**0.6178**	**0.4222**	**0.4607**

Table 3: Results for Task 2 with the M1 predictor-estimator (XLM-RoBERTa) and different training/fine-tuning approaches. M1M is the M1 model trained on the Metrics dataset and M#-adapt signifies a model fine-tuned on the QE data with adapters. ML stands for Multilingual, showing the performance averaged over all language pairs. Underlined numbers indicate the best result for each language pair and evaluation metric. **Bold** systems were selected for the final ensemble.

inspecting the results for the blind sets (`en-cs`, `en-ja`, `km-en` and `ps-en`) in the official results on `test-21` as shown in Appendix B.

lp	train	dev	test-20
En-De	-0.1654	-0.4032	-0.3850
En-Zh	-0.2947	-0.1895	-0.1932
Et-En	-0.5464	-0.5850	-0.5995
Ne-En	-0.4527	-0.5004	-0.4558
Ro-En	-0.5887	-0.7932	-0.7880
Ru-En	-0.5358	-0.5055	-0.5152
Si-En	-0.3916	-0.4384	-0.4125

Table 4: Pearson correlation between the z_mean of the direct assessments for the QE Task 1 data and the HTER score for the post edits in QE Task 2 data.

5 Conclusions

We presented a joint contribution of IST and Unbabel to the WMT 2021 QE shared task. Our submissions are ensembles of multilingual checkpoints extending the `OpenKiwi` framework. We found adapter-tuning to be suitable for fine-tuning OpenKiwi on the QE tasks data and less prone to overfitting. We showed that pre-training on large, out-of-domain annotated data can prove beneficial both for the direct assessment and the post-editing QE tasks. We also demonstrated that handling uncertainty-related sources of information improves the performance when integrated into the QE system. For Task 2 we do multi-task training based on the models from the previous task and use multiple checkpoints to create the submitted ensemble.

Acknowledgements

We are grateful to Alon Lavie and Craig Stewart for their valuable feedback and discussions. This work was supported by the P2020 programs MAIA (contract 045909) and Unbabel4EU (contract 042671), by the European Research Council (ERC StG DeepSPIN 758969), and by the Fundação para a Ciência e Tecnologia through contract UIDB/50008/2020.

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.

Trevor Cohn and Lucia Specia. 2013. Modelling annotator bias with multi-task Gaussian processes: An application to machine translation quality estimation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 32–42, Sofia, Bulgaria. 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.

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 Fornaciari, Alexandra Uma, Silviu Paun, Barbara Plank, Dirk Hovy, and Massimo Poesio. 2021. Beyond black & white: Leveraging annotator disagreement via soft-label multi-task learning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2591–2597, Online. Association for Computational Linguistics.

Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1050–1059, New York, New York, USA. PMLR.

Taisiya Glushkova, Chrysoula Zerva, Ricardo Rei, and André F. T. Martins. 2021. Uncertainty-Aware Machine Translation Evaluation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, Online. Association for Computational Linguistics.

Ruidan He, Linlin Liu, Hai Ye, Qingyu Tan, Bosheng Ding, Liying Cheng, Jiawei Low, Lidong Bing, and Luo Si. 2021. On the effectiveness of adapter-based tuning for pretrained language model adaptation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2208–2222, Online. Association for Computational Linguistics.

Stephen C. Hora. 1996. Aleatory and epistemic uncertainty in probability elicitation with an example from hazardous waste management. *Reliability Engineering & System Safety*, 54(2):217–223. Treatment of Aleatory and Epistemic Uncertainty.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for NLP. In *International Conference on Machine Learning*, pages 2790–2799. PMLR.

Eyke Huellermeier and Willem Waegeman. 2021. Aleatoric and epistemic uncertainty in machine learning : an introduction to concepts and methods. *Machine Learning*, 110(3):457–506.

Fabio 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 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.

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.

Armen Der Kiureghian and Ove Ditlevsen. 2009. Aleatory or epistemic? does it matter? *Structural Safety*, 31(2):105–112. Risk Acceptance and Risk Communication.

Alon Lavie and Michael Denkowski. 2009. The meteor metric for automatic evaluation of machine translation. *Machine Translation*, 23:105–115.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Nitika Mathur, Johnny Wei, Markus Freitag, Qingsong Ma, and Ondřej Bojar. 2020. Results of the WMT20 metrics shared task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 688–725, Online. Association for Computational Linguistics.

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

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.

Jonas Pfeiffer, Andreas Rücklé, Clifton Poth, Aishwarya Kamath, Ivan Vulić, Sebastian Ruder, Kyunghyun Cho, and Iryna Gurevych. 2020. AdapterHub: A framework for adapting transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 46–54, Online. Association for Computational Linguistics.

Lucia Specia, Frédéric Blain, Marina Fomicheva, Chrysoula Zerva, Zhenhao Li, Vishrav Chaudhary, and André F. T. Martins. 2021. Findings of the wmt 2021 shared task on quality estimation. In *Proceedings of the Sixth Conference on Machine Translation*, Online. Association for Computational Linguistics.

Lucia Specia, Carolina Scarton, and Gustavo Henrique Paetzold. 2018. Quality estimation for machine translation. *Synthesis Lectures on Human Language Technologies*, 11(1):1–162.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning. *arXiv preprint arXiv:2008.00401*.

A Hyperparameters

A.1 M1

In Table 5 is an excerpt of the training configuration used for training OpenKiwi for our M1 models. Note that the configurations follow the configuration file format of OpenKiwi and any additional configurations are identical to the ones proposed in the sample configuration file of the `github` repository[5].

System	
batch_size	2
Encoder	
hidden_size	1024
Decoder	
bottleneck_size	1024
dropout	0.05
hidden_size	1024
Optimizer	
class_name	adam
encoder_learning_rate	0.0001
learning_rate_decay	1.0
learning_rate_decay_start	0
learning_rate	0.0001
Trainer	
training_steps	2180
early_stop_patience	10
validation_steps	0.5
gradient_accumulation_steps	4
gradient_max_norm	1.0

Table 5: Hyperparameters for M1 models

A.2 M2

In Table 6 is an excerpt of the training configuration used for training the M2 models using the mBART encoder-decoder:

System	
bottleneck_size	256
dropout	0.1
hidden_size	2048
nr_frozen_epochs	0.333
Optimizer	
optimizer	adam
encoder_learning_rate	6.0e-06
learning_rate	1.0e-05
Trainer	
training_steps	5512
early_stopping_patience	2
save_top_k	3
batch_size	4
gradient_accumulation_steps	4

Table 6: Hyperparameters for M2 models

B Evaluation on test set of WMT21

We present the performance of the submitted ensembles on the TEST-21 dataset as calculated in the official QE results [6] for each task and sub-task. We also provide the comparison with the organisers' baseline.

B.1 Task 1: Direct Assessments prediction at sentence-level

The results for Task1 on TEST-21 are presented in Table 7.

B.2 Task 2: HTER prediction at sentence-level

The results for Task2 on TEST-21TEST-21 are presented in Table 8, showing the performance for the sentence level, HTER score predictions.

B.3 Task 2: Word-level prediction

The results for Task2 on TEST-21 are presented in Table 9, showing the performance for the word tag predictions.

C Statistics on the Metrics data

We present below (Tables 10 and 11) the statistics on the Metrics data used to train the M1M model on direct assessments.

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

[6] https://www.statmt.org/wmt21/quality-estimation-task_results.html

Method	Pearson r↑	MAE↓	RMSE↓
Multilingual			
IST-Unbabel	0.665	0.627	0.482
Baseline	0.541	0.729	0.562
En-De			
IST-Unbabel	0.579	0.567	0.393
Baseline	0.403	0.629	0.433
En-Zh			
IST-Unbabel	0.586	0.631	0.499
Baseline	0.525	0.683	0.534
Ro-En			
IST-Unbabel	0.899	0.393	0.289
Baseline	0.818	0.556	0.408
Et-En			
IST-Unbabel	0.796	0.519	0.404
Baseline	0.660	0.700	0.543
Ne-En			
IST-Unbabel	0.856	0.515	0.401
Baseline	0.738	0.657	0.524
Si-En			
IST-Unbabel	0.605	0.742	0.583
Baseline	0.513	0.797	0.626
Ru-En			
IST-Unbabel	0.792	0.583	0.412
Baseline	0.677	0.702	0.492
Zero-shot language pairs			
En-Cz			
IST-Unbabel	0.577	0.751	0.583
Baseline	0.352	0.845	0.686
En-Ja			
IST-Unbabel	0.355	0.764	0.566
Baseline	0.230	0.816	0.617
Ps-En			
IST-Unbabel	0.628	0.780	0.658
Baseline	0.476	0.852	0.711
Km-En			
IST-Unbabel	0.650	0.721	0.568
Baseline	0.562	0.788	0.614

Table 7: Results for Task 1 on the held-out evaluation set of WMT 2021.

Method	Pearson r↑	MAE↓	RMSE↓
Multilingual			
IST-Unbabel	0.597	0.219	0.171
Baseline	0.502	0.235	0.188
En-De			
IST-Unbabel	0.617	0.172	0.116
Baseline	0.529	0.183	0.129
En-Zh			
IST-Unbabel	0.290	0.266	0.220
Baseline	0.282	0.287	0.246
Ro-En			
IST-Unbabel	0.879	0.122	0.098
Baseline	0.831	0.142	0.115
Et-En			
IST-Unbabel	0.811	0.153	0.112
Baseline	0.714	0.195	0.149
Ne-En			
IST-Unbabel	0.718	0.161	0.126
Baseline	0.626	0.205	0.160
Si-En			
IST-Unbabel	0.710	0.178	0.136
Baseline	0.607	0.204	0.159
Ru-En			
IST-Unbabel	0.539	0.224	0.165
Baseline	0.448	0.255	0.188
Zero-shot language pairs			
En-Cz			
IST-Unbabel	0.529	0.271	0.200
Baseline	0.306	0.262	0.206
En-Ja			
IST-Unbabel	0.275	0.279	0.224
Baseline	0.098	0.279	0.232
Ps-En			
IST-Unbabel	0.555	0.328	0.284
Baseline	0.503	0.333	0.290
Km-En			
IST-Unbabel	0.655	0.243	0.199
Baseline	0.576	0.241	0.196

Table 8: Results for Task 2 sentence-level system on the held-out evaluation set of WMT 2021.

METHOD	SRC-MCC↑	TGT-MCC-WORDS↑	TGT-MCC-GAPS↑
		EN-DE	
IST-UNBABEL	0.404	0.466	0.183
BASELINE	0.322	0.370	0.116
		EN-ZH	
IST-UNBABEL	0.286	0.310	0.068
BASELINE	0.241	0.247	0.065
		RO-EN	
IST-UNBABEL	0.603	0.649	0.357
BASELINE	0.511	0.536	0.205
		ET-EN	
IST-UNBABEL	0.522	0.570	0.254
BASELINE	0.405	0.461	0.136
		NE-EN	
IST-UNBABEL	0.445	0.508	0.268
BASELINE	0.390	0.440	0.215
		SI-EN	
IST-UNBABEL	0.406	0.528	0.258
BASELINE	0.335	0.425	0.208
		RU-EN	
IST-UNBABEL	0.351	0.332	0.165
BASELINE	0.251	0.256	0.073
		ZERO-SHOT LANGUAGE PAIRS	
		EN-CZ	
IST-UNBABEL	0.294	0.376	0.125
BASELINE	0.224	0.273	0.039
		EN-JA	
IST-UNBABEL	0.175	0.169	0.025
BASELINE	0.175	0.131	0.036
		PS-EN	
IST-UNBABEL	0.294	0.370	0.177
BASELINE	0.249	0.313	0.134
		KM-EN	
IST-UNBABEL	0.345	0.448	0.259
BASELINE	0.279	0.351	0.175

Table 9: Results for Task 2 word-level system on the held-out evaluation set of WMT 2021.

	CS-EN	DE-EN	FI-EN	RU-EN	RO-EN	TR-EN	ZH-EN	ET-EN
	LT-EN	GU-EN	KK-EN	JA-EN	KM-EN	PL-EN	PS-EN	TA-EN
Total tuples	28887	91584	47205	61505	560	30746	71941	20496
	10315	9063	6789	8917	4722	11666	4611	7562
Avg. tokens (reference)	31.43	24.61	20.48	23.31	24.35	23.32	31.70	23.93
	26.84	17.73	20.65	28.64	19.49	21.93	19.87	19.91
Avg. tokens (source)	25.65	22.93	14.49	19.77	24.99	19.01	6.05	18.61
	20.61	15.13	16.47	3.27	29.91	18.55	21.87	15.31
Avg. tokens (MT)	29.99	24.19	19.95	23.51	24.42	22.97	30.60	24.06
	25.44	17.15	20.00	27.41	19.59	21.64	19.37	20.14

Table 10: Statistics for the WMT 15 to 20 Direct Assessments corpus into-English language pairs.

	EN-RU	EN-CS	EN-DE	EN-FI	EN-LV	EN-TR	EN-ZH
	EN-ET	EN-LT	EN-GU	EN-KK	EN-JA	EN-PL	EN-TA
Total tuples	63771	60905	55352	30924	5810	5171	66830
	13376	8959	6924	8219	9573	10506	7886
Avg. tokens (reference)	22.48	23.48	23.96	17.7	20.45	19.74	7.26
	18.83	20.61	22.07	19.21	1.4	24.54	19.84
Avg. tokens (source)	24.5	25.82	24	23.21	24.99	24.2	28.81
	24.23	24.09	24.3	24.13	25.2	25.33	25.15
Avg. tokens (MT)	22.14	23	23.84	17.81	21.18	19.24	7.53
	18.96	20.62	22.39	19.71	2.29	23.19	19.18

Table 11: Statistics for the WMT 15 to 20 Direct Assessments corpus from-English language pairs.