TMUOU Submission for WMT20 Quality Estimation Shared Task

Akifumi Nakamachi
Osaka University
nakamachi.akifumi@ist.osaka-u.ac.jp

Hiroki Shimanaka
Tokyo Metropolitan University
shimanaka-hiroki@ed.tmu.ac.jp

Tomoyuki Kajiwara
Osaka University
kajiwara@ids.osaka-u.ac.jp

Mamoru Komachi
Tokyo Metropolitan University
komachi@tmu.ac.jp

Abstract

We introduce the TMUOU[1] submission for the WMT20 Quality Estimation Shared Task 1: Sentence-Level Direct Assessment. Our system is an ensemble model of four regression models based on XLM-RoBERTa with language tags. We ranked 4th in Pearson and 2nd in MAE and RMSE on a multilingual track.

1 Introduction

Quality Estimation (QE) is a task of estimating translation quality without reference sentences (Gandrabur and Foster, 2003; Blatz et al., 2004; Specia et al., 2018). Automatic evaluation metrics based on reference sentences, such as BLEU (Papineni et al., 2002), have contributed to improving translation quality on benchmark datasets. However, in situations where machine translation (MT) is actually used, these metrics are sometimes unable to assess the translation quality owing to the lack of reference sentences. The development of QE methods that are well correlated with manual evaluations enable users to decide whether to use the translation results as is, post-edit the results, or employ other machine translations.

At the Conference on Machine Translation (WMT), there have been conducted several QE-related competitions such as the QE task (Fonseca et al., 2019) for estimating post-edit rate HTER (Snover et al., 2006) and the QE as a Metric task (Ma et al., 2019) for relative evaluations of translation quality. This year, the WMT QE task held a new competition (Specia et al., 2020) on absolute evaluations of translation quality. In task 1, sentences are annotated with direct assessment (DA) scores as in the metrics task (Bojar et al., 2017).

We have been working on the metrics task with an approach that uses pre-trained sentence encoders (Shimanaka et al., 2018, 2019). Shimanaka et al. (2018) employed InferSent (Conneau et al., 2017), Quick-Thought (Logeswaran and Lee, 2018), and Universal Sentence Encoder (Cer et al., 2018) as encoders, and achieved the highest performance in all to-English language pairs of WMT18 metrics shared task (Ma et al., 2018). Subsequently, Shimanaka et al. (2019) employed BERT (Devlin et al., 2019) as an encoder to further improve the correlation with manual evaluations. In this study, we apply similar approaches to the QE task. However, to support both source and target languages, we employ XLM-RoBERTa[2] (Conneau et al., 2020), a pre-trained multilingual sentence encoder.

2 WMT20 QE Shared Task 1

In the WMT20 QE task 1 (Sentence-Level Direct Assessment), participants predict translation quality at the sentence level from pairs of source and MT output sentences. This task provides datasets for seven language pairs and sets up a multilingual track for a language-independent approach.

2.1 Datasets

Source sentences have been collected from Wikipedia for six language pairs: English–German (En-De), English–Chinese (Eh-Zh), Romanian–English (Ro-En), Estonian–English (Et-En), Nepalese–English (Ne-En), and Sinhala–English(Si-En). In addition, a combination of 75% Reddit data and 25% Wikipedia data for the Russian–English (Ru-En) language pair is provided. Organizers trained state-of-the-art neural MT models on each dataset using the fairseq toolkit (Ott et al., 2019) and generated MT output sentences.

[1] **T**okyo **M**etropolitan **U**niversity and **O**saka **U**niversity

[2] https://github.com/facebookresearch/XLM

Proceedings of the 5th Conference on Machine Translation (WMT), pages 1037–1041

Source	MT output	QE score
Its ferocious winds defoliated nearly all vegetation, splintering or uprooting thousands of trees and decimating the island's lush rainforests.	Seine wilden Winde entblätterten fast die gesamte Vegetation, zersplitterten oder entwurzelten Tausende von Bäumen und dezimierten die üppigen Regenwälder der Insel.	1.267
The Cubs tied it in the third on a triple by Ben Zobrist to knock in Daniel Murphy.	Die Cubs band es in der dritten auf einem Triple von Ben Zobrist in Daniel Murphy klopfen.	−3.760

Table 1: Examples of English-German dataset.

Three or more professional translators annotated DA scores in the range of 0-100 points for each pair of source and MT output sentences. These annotations are following the FLORES setup (Guzmán et al., 2019). The dataset consists of pairs of source and MT output sentences, z-standardized DA scores, and MT model score (log probabilities for words). Table 1 shows examples of the dataset. For each language pair, 7,000 training sets, 1,000 development sets, and 1,000 test sets are provided.

2.2 Baseline and Evaluation

The baseline system is a Predictor-Estimator model (Kim et al., 2017) implemented in OpenKiwi[3] (Kepler et al., 2019). The predictor is trained on a parallel corpus used to train the MT model, and predicts each target token from source and target contexts. And the estimator predicts the QE score from features produced by the predictor.

Participants are evaluated by Pearson's correlation metric (Pearson), mean absolute error (MAE), and root mean squared error (RMSE). A z-standardized DA score is used as a gold label.

3 TMUOU System

Our system is an ensemble model of four regression models based on XLM-RoBERTa (Conneau et al., 2020) with language tags. We first explain each base model in Section 3.1, and then introduce the ensemble model in Section 3.2. Finally, Section 3.3 describes the implementation details.

3.1 Base Models

Recently, the fine-tuning approach for masked language models (Devlin et al., 2019) has achieved the highest performance for many language understanding tasks (Wang et al., 2019). The BERT-based regression model (Shimanaka et al., 2019) also achieves high performance in the WMT metric task that estimates the DA score of translation quality (Bojar et al., 2017). We employ XLM-RoBERTa (Conneau et al., 2020), a multilingual masked language model, for this task to estimate the DA score of translation quality from pairs of source and MT output sentences.

E0 Model In this model, we fine-tune the XLM-RoBERTa in the normal way. We input sentence pairs into the model in the following format and use the special token `<s>` at the beginning of the first sentence to estimate the QE score: `<s> source </s> <s> MT output </s>`.

E0+LangTag Model To make it clear to the XLM-RoBERTa which language each sentence is in, we add a special token (LangTag) for language identification, such as `<en>`, at the beginning of each sentence. We have expanded the tokenizer and vocabulary and added the following eight LangTags: `<en> <et> <de> <ne> <ro> <ru> <si> <zh>`. An example of input to the model is as follows: `<s> <en> source </s> <s> <de> MT output </s>`.

E0+AVG Model Averaged token vector is as fruitful as the `<s>` vector at the beginning of the first sentence (Reimers and Gurevych, 2019). We concatenate the averaged token vector and the `<s>` vector to get richer information from sentence pairs.

E0+AVG+LangTag Model This model is a combination of the above models. As shown in Figure 1, we add LangTag at the beginning of each sentence and concatenate the `<s>` vector with the averaged token vector to estimate the QE score.

3.2 Ensemble Model

We ensemble four models described above to make prediction stable. A Gradient Boosting Tree (Fried-

[3]https://github.com/Unbabel/OpenKiwi

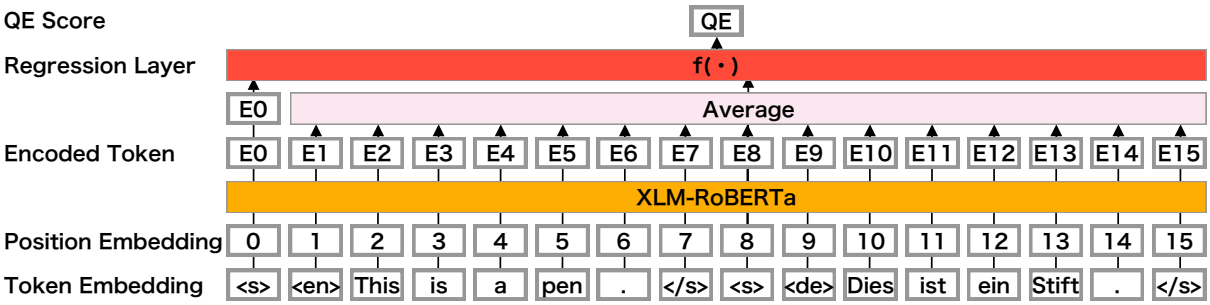

Figure 1: Overview of the TMUOU system.

	En-De	En-Zh	Ro-En	Et-En	Ne-En	Si-En	Ru-En	Multilingual
E0	0.455	**0.490**	0.860	0.747	0.742	0.646	0.693	0.662
E0+LangTag	0.419	0.465	0.874	0.744	0.763	0.648	**0.701**	0.652
E0+AVG	**0.461**	0.440	0.873	0.738	0.751	**0.658**	0.689	0.659
E0+AVG+LangTag	0.410	0.465	**0.885**	**0.764**	**0.769**	0.646	0.699	**0.663**
Ensemble	0.485	0.506	0.897	0.783	0.801	0.691	0.726	0.698

Table 2: Pearson's correlation on the development sets.

man, 2001) is trained using k-fold cross-validation on the development set with the QE scores estimated by each base model as the features. In addition to the QE scores estimated by each base model, the features of the ensemble model also include the sum of MT model scores for each output word and a one-hot vector representing the language pair.

3.3 Implementation Details

We implemented all models based on the Hugging Face (Wolf et al., 2019) XLM-RoBERTa-large model.[4] The hyper parameters are as follows: batch size is 16, weight decay is 0.01, gradient clipping norm is 5.0, dropout for the attention layers and regression layer are 0.1, max epoch is 100. We use early stopping by Pearson metric on the dev sets with patience 5. We use Adam optimizer (Kingma and Ba, 2015) with warm up. The learning rate for the optimizer is $2e^{-5}$, and we gradually decrease the learning rate by a linear scheduler.

For the ensemble model, we trained gradient boosting regressor with least square loss implemented in scikit-learn (Pedregosa et al., 2011) with 10 folds cross-validation. The hyper parameters are as follows: the initial learning rate is 0.1, the number of estimators are 100, the subsample ratio is 1.0, the criterion is mean squared error with improvement score by Friedman, the minimum amount of sample split is 2, max depth of the tree is 3.

[4]https://huggingface.co/xlm-roberta-large

	MAE	RMSE	Pearson
Bergamot-LATTE	0.408	0.527	0.718
TMUOU	**0.418**	**0.543**	**0.686**
IST and Unbabel	0.433	0.569	0.673
TransQuest	0.480	0.596	0.722
NiuTrans	0.529	0.653	0.732
WL Research	0.538	0.683	0.546
IST and Unbabel	0.547	0.719	0.583
Baseline	0.788	0.999	0.376
Bergamot-LATTE	0.895	1.062	0.489
nc	0.918	1.141	0.462

Table 3: Official results in ascending order of MAE.

4 Results

Table 2 shows the Pearson's correlation of each model on the development sets. Although there is no significant difference in the performance of the base models, the E0+AVG+LangTag model achieves higher performance in the majority of language pairs. The ensemble model achieves the highest performance in all language pairs. QE performance of to-English language pairs tends to be higher than that of from-English language pairs.

Table 3 presents the official results for a multilingual track. Participants are listed in ascending order of MAE. We submitted the ensemble model and ranked 4th in Pearson and 2nd in MAE and RMSE on a multilingual track.

5 Conclusions

We describe the TMUOU submission for the WMT20 Shared Task on Quality Estimation. Our system is an ensemble model based on XLM-RoBERTa, which takes into account averaged token vectors and language identifiers to improve performance. In the official evaluation, we ranked 4th in Pearson and 2nd in MAE and RMSE on a multilingual track.

Acknowledgments

This work was supported by JST ACT-X Grant Number JPMJAX1907, and JSPS KAKENHI Grant Number JP20K19861.

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 *Proceedings of the 20th International Conference on Computational Linguistics*, pages 315–321.

Ondřej Bojar, Yvette Graham, and Amir Kamran. 2017. Results of the WMT17 Metrics Shared Task. In *Proceedings of the Second Conference on Machine Translation*, pages 489–513.

Daniel Cer, Yinfei Yang, Sheng yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Brian Strope, and Ray Kurzweil. 2018. Universal Sentence Encoder for English. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 169–174.

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.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised Learning of Universal Sentence Representations from Natural Language Inference Data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186.

Erick Fonseca, Lisa Yankovskaya, André F. T. Martins, Mark Fishel, and Christian Federmann. 2019. Findings of the WMT 2019 Shared Tasks on Quality Estimation. In *Proceedings of the Fourth Conference on Machine Translation*, pages 1–12.

Jerome H. Friedman. 2001. Greedy Function Approximation: A Gradient Boosting Machine. *Annals of Statistics*, 29(5):1189–1232.

Simona Gandrabur and George Foster. 2003. Confidence Estimation for Translation Prediction. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 95–102.

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*, pages 6098–6111.

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*, pages 117–122.

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.

Diederik P. Kingma and Jimmy Lei Ba. 2015. Adam: A Method for Stochastic Optimization. In *Proceedings of the 3rd International Conference on Learning Representations*.

Lajanugen Logeswaran and Honglak Lee. 2018. An Efficient Framework for Learning Sentence Representations. In *Proceedings of the 6th International Conference on Learning Representations*.

Qingsong Ma, Ondřej Bojar, and Yvette Graham. 2018. Results of the WMT18 Metrics Shared Task: Both Characters and Embeddings Achieve Good Performance. In *Proceedings of the Third Conference on Machine Translation*, pages 682–701.

Qingsong Ma, Johnny Wei, Ondřej Bojar, and Yvette Graham. 2019. Results of the WMT19 Metrics Shared Task: Segment-Level and Strong MT Systems Pose Big Challenges. In *Proceedings of the Fourth Conference on Machine Translation*, pages 62–90.

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*

the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 48–53.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12(85):2825–2830.

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*, pages 3982–3992.

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*, pages 764–771.

Hiroki Shimanaka, Tomoyuki Kajiwara, and Mamoru Komachi. 2019. Machine Translation Evaluation with BERT Regressor. *arXiv:1907.12679*.

Matthew Snover, Bonnie Dorr, Richard 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*, pages 223–231.

Lucia Specia, Frédéric Blain, Marina Fomicheva, Erick Fonseca, Vishrav Chaudhary, Francisco Guzmán, and André FT Martins. 2020. Findings of the WMT 2020 Shared Task on Quality Estimation. In *Proceedings of the Fifth Conference on Machine Translation: Shared Task Papers*.

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

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In *Proceedings of the 7th International Conference on Learning Representations*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. HuggingFace's Transformers: State-of-the-art Natural Language Processing. *arXiv:1910.03771*.