Quality Estimation using Minimum Bayes Risk

Subhajit Naskar, Daniel Deutsch, and Markus Freitag
Google
{snaskar,dandeutsch,freitag}@google.com

Abstract

This report describes the Minimum Bayes Risk Quality Estimation (MBR-QE) submission to the Workshop on Machine Translation's 2023 Metrics Shared Task. MBR decoding with neural utility metrics like BLEURT is known to be effective in generating high quality machine translations. We use the underlying technique of MBR decoding and develop an MBR based reference-free (quality estimation) metric. Our method uses an evaluator machine translation system and a reference-based utility metric (specifically BLEURT and MetricX) to calculate a quality estimation score of a model's output. We report results related to comparing different MBR configurations and utility metrics.

1 Introduction

The task of quality estimation (QE) is to assign a sentence- or word-level quality score to a machine translation (MT) output without the use of a reference translation. In this paper, we describe the methodology used in our sentence-level QE metric submission to the 2023 Workshop on Machine Translation's Metrics Shared Task.

Minimum Bayes Risk (MBR) decoding has been widely used in machine translation to address the limitation of MAP decoding (Kumar and Byrne, 2004; Eikema and Aziz, 2020; Müller and Sennrich, 2021). Freitag et al. (2021b) showed applying MBR decoding using BLEURT (Sellam et al., 2020) as a utility function can out-perform beam search decoding.

MBR decoding can be viewed as a method for reranking candidate outputs from an MT system. It first samples a set of hypothesis translations from the model, scores each hypothesis against a set of pseudo-references (generally, the same set of sample hypotheses) with a utility metric, then selects the hypothesis with the highest average score to be the final translation.

Central to MBR is assigning a quality score to a hypothesis translation without the use of a reference. Because this decoding procedure has been successful in improving the quality of translations from an MT system, in this work, we explore how MBR could be repurposed as a QE metric.

Our proposed metric uses an MT system in conjunction with a utility metric to assign a quality score to a translation without using a reference. The metric assigns a score to a hypothesis translation by using the utility metric to evaluate the hypothesis against a set of pseudo-references that are sampled from the MT model.

In this work, we experiment with creating a metric that uses different MT systems, utility functions, and different pseudo-reference pool sizes. Our experiments demonstrate that (1) a better utility function results in better MBR-QE scores, (2) the choice of MT system can have significant impact on QE metric performance, and (3) the size of the pseudo-reference pool does not have a significant impact on overall metric quality.

Based on our experiments, we chose our primary MBR-QE submission to be an in-house encoder-decoder model with MetricX (Freitag et al., 2022) as the utility function with a pseudo-reference pool size of 256.

2 Metric Descriptions

MBR decoding has two components: an MT system and a utility function. The MT model $P_{model}(y|x)$ estimates the probability of target segments y given a source segment x. The utility function estimates the quality of a translation h given a reference translation r. The best hypothesis is selected using the expected utility with respect to a finite sample generated by the model. The underlying assumption is that the model provides good approximation for the true distribution of human translations.

We adopt that assumption to develop an MBR-

Proceedings of the Eighth Conference on Machine Translation (WMT), pages 806–811
December 6–7, 2023.

based quality estimation metric. The MBR-QE metric uses an MT system P to generate the set of pseudo-references, denoted $\hat{R}$. Then, the utility function defines the quality of a translation h given $\hat{R}$ as the average score over all of $\hat{R}$:

$$\text{MBR-QE}(h) = \frac{1}{|\hat{R}|} \sum_{\hat{r} \epsilon \hat{R}} u(h, \hat{r}) \quad (1)$$

This methodology has multiple potential pitfalls. First, because the distribution P is used to substitute for the distribution of human translations, any significant divergence between these two distributions will lead to the QE score becoming inconsistent because the pool of pseudo-references will not resemble human references. This can be mitigated by using a high quality MT system. Arguably, the MT system should have better performance compared to the MT models that are being evaluated.

Second, our QE metric is dependent on the quality of the utility function. If it has limitations or biases, they will affect the predicted quality scores and introduce inconsistencies between the QE score and ground-truth human quality scores.

We next discuss the experimental setup for analyzing our proposed QE metric.

3 Experimental Setup

3.1 Pseudo-Reference Generation

Our MBR-QE metric relies on the assumption that if the MT system that generates the pseudo-references can be used as an approximation for the distribution of human translations, then the aggregated utility metric score can be used a quality estimate for hypothesis. Therefore, the MT model and method for generating pseudo-references is critical for the effectiveness of this metric.

MT Systems. The MT system used for our shared task submission is an in-house encoder-decoder translation model that is similar to the Google Translate production model. In this report, we also experiment with generating pseudo-references from the PaLM 2 (Bison) large language model (Anil et al., 2023) using 5-shot prompting.

Sampling Method. We generate pseudo-references from the MT system using epsilon sampling (Hewitt et al., 2022; Freitag et al., 2023) with $p = 0.02$ and sampling temperature 1.0. We experiment with using a different number of pseudo-references.

3.2 Utility Functions

Freitag et al. (2021b) showed that MBR decoding works well with neural evaluation metrics. We experiment with 2 neural metrics as the utility function in MBR-QE.

BLEURT v0.2 (Sellam et al., 2020; Pu et al., 2021): BLEURT v0.2 is a learned regression-based metric that is trained to predict the quality of a translation given a reference. It is pre-initialized with RemBERT (Chung et al., 2020) and finetuned using a combination of WMT human evaluation data from 2015-2019 and synthetic data.

MetricX (Freitag et al., 2022): MetricX is a learned regression-based metric that is based on mT5 (Xue et al., 2021). It is trained on a combination of direct assessment and MQM (Lommel et al., 2014; Freitag et al., 2021a) data that was collected by WMT. We use the reference-based version that uses mT5-XXL.

3.3 Meta-Evaluation

We use four different correlations to calculate the metrics' agreements with human judgments. At the system-level, we use pairwise accuracy (Kocmi et al., 2021) and Pearson's r. System-level Pearson's r captures how strong the linear relationship is between the metric and human scores for MT systems. Pairwise accuracy evaluates a metric's ranking of MT systems by calculating the proportion of all possible pairs of MT systems that are ranked the same by the metric and human scores.

At the segment-level, we use group-by-item pairwise accuracy with tie calibration (Deutsch et al., 2023) and no-grouping Pearson's r. The no-grouping Pearson's r calculates the linear relationship between the metric and human scores across translations from every system and document. The group-by-item pairwise accuracy calculates the proportion of all possible pairs of translations for the same input segment that are ranked the same or tied by the metric and human. Then the accuracy scores are averaged over all possible input segments. We use tie calibration (Deutsch et al., 2023) that automatically introduces ties into metric scores based on a threshold. This tie calibration is required as regression-based metrics rarely predict ties.

Our experiments are performed using the WMT'22 English-to-German (en-de) and Chinese-to-English (zh-en) MQM ratings (Freitag et al., 2022). These datasets are commonly used for meta-

evaluation and are the latest available from the Metrics Shared Task. We did not evaluate using en-ru since it is not included as a language pair in the WMT'23 evaluation.

4 Experimental Results

The main experimental results are shown in Tables 1 and 2. Table 1 compares the two utility functions with various pseudo-reference pool sizes when using the in-house MT system, and Table 2 does the same but for the PaLM 2-based system.

Comparing Utility Functions. For both MT systems and all pseudo-reference pool sizes, the MBR-QE metric that uses MetricX as a utility function in general has higher correlations than when BLEURT is used. This result is expected since MetricX was the best performing metric in the WMT'22 evaluation. This is evidence that the quality of the utility function is important for the quality of the MBR-QE score.

Comparing MT Systems. When comparing whether the encoder-decoder MT system or PaLM 2 is used to generate the pool of pseudo-references, there is no clear winner between the two. The MBR-QE score has a higher correlation at the segment-level with the encoder-decoder model, but the correlations are higher at the system-level with PaLM 2. It is not clear why this is the case.

Pseudo-Reference Pool Size. Overall, the correlations are surprisingly stable for each of the different numbers of pseudo-references. Most of the differences comes between pairwise accuracy at the system-level, but this correlation can be sensitive; there are not many system pairs, so if one or two system rankings change, it can have a large impact on the overall accuracy. In the future, we could explore decreasing the pseudo-reference pool size even further to understand its impact on the overall MBR-QE metric quality.

Comparing to Other Metrics. Table 3 shows the comparion between our submission, denoted MBR-QE, to other QE metrics that were the top-performing QE metrics in the WMT'22 Metrics Shared Task, CometKiwi (Kepler et al., 2019; Rei et al., 2022b) and UniTE-src (Wan et al., 2022). The table additionally contains results for the best reference-based metrics MetricX and COMET-22 (Rei et al., 2022a).

Compared to the QE metrics, MBR-QE in general has the best-performance across most evaluation settings, demonstrating that it is a state-of-the-art QE metric. In some cases, it even out-performs the reference-based metrics, namely in the system-level Pearson correlation.

MBR-QE leverages MetricX as the utility function. MBR-QE still under-performs with respect to MetricX, demonstrating that the human references are still valuable and that the pseudo-references do not perfectly match the distribution of human translations, which is expected given that the MT system is not perfect. However, the gap in performance between the two metrics is relatively small in some settings.

4.1 Submission Summary

Both of our submissions to the Metrics Shared task use the in-house MT system to generate 256 pseudo-references with epsilon sampling ($p = 0.02$ and temperature 1.0). Our primary submission uses MetricX as the utility function, and the contrastive submission uses BLEURT.

5 Related Work

Incorporating evaluation metrics into reranking the outputs from MT systems has been very successful. For example, the Freitag et al. (2021b) showed that reranking translations with BLEURT as part of MBR produced higher-quality translations. This work served as the inspiration for our QE metric submission.

Research on quality estimation focuses on predicting word- and sentence-level quality scores (Zerva et al., 2022). The most successful approaches to predicting sentence-level scores are learned regression-based metrics that are trained to predict ground-truth quality scores, like CometKiwi (Kepler et al., 2019; Rei et al., 2022b) or UniTE-src (Wan et al., 2022). Our metric is quite different from these approaches in that it is not directly trained to predict quality scores, but rather it leverages a reference-based metric combined with an MT system to score a translation. To the best of our knowledge, ours is the first metric that uses MBR to build a QE metric.

6 Conclusion

In this report, we proposed a new QE metric called MBR-QE that repurposes an MT system in combination with MBR to score a translation without ac-

Utility Metric	Pseudo-Ref Pool Size	SEG pairwise acc. en-de	SEG pairwise acc. zh-en	SEG Pearson en-de	SEG Pearson zh-en	SYS pairwise acc. en-de	SYS pairwise acc. zh-en	SYS Pearson en-de	SYS Pearson zh-en
BLEURT	64	0.5772	0.5142	0.4750	0.4628	0.7692	0.7802	0.6773	0.8907
	128	0.5777	0.5145	0.4752	**0.4631**	0.7692	0.7802	0.6749	0.8899
	256	0.5782	0.5151	0.4747	0.4626	0.7692	0.7692	0.6751	0.8901
MetricX	64	**0.5986**	0.5292	0.4891	0.4513	0.7564	**0.8132**	**0.8654**	0.8654
	128	0.5944	0.5300	0.4873	0.4519	**0.7821**	**0.8132**	0.8391	0.9579
	256	0.5979	**0.5306**	**0.4897**	0.4524	0.7692	**0.8132**	0.8647	**0.9586**

Table 1: MBR-QE correlations on the WMT'22 MQM data comparing when BLEURT and MetricX are used as utility functions with different pseudo-reference pool sizes are sampled from the in-house encoder-decoder model.

Utility Metric	Pseudo-Ref Pool Size	SEG pairwise acc. en-de	SEG pairwise acc. zh-en	SEG Pearson en-de	SEG Pearson zh-en	SYS pairwise acc. en-de	SYS pairwise acc. zh-en	SYS Pearson en-de	SYS Pearson zh-en
BLEURT	64	0.5614	0.4824	0.4261	0.4153	0.8077	0.7802	0.7636	0.9253
	128	0.5612	0.4827	0.4246	0.4143	0.8077	0.7802	0.7631	0.9250
	256	0.5616	0.4831	0.4249	0.4152	0.8077	0.7692	0.7635	0.9249
MetricX	64	0.5764	**0.5022**	0.4574	0.4259	0.7949	**0.8571**	**0.9154**	**0.9846**
	128	0.5737	0.5000	0.4621	0.4339	**0.8333**	0.8242	0.9145	0.9844
	256	**0.5767**	0.5021	**0.4626**	**0.4265**	0.8077	**0.8571**	0.9212	0.9845

Table 2: MBR-QE correlations on the WMT'22 MQM data comparing when BLEURT and MetricX are used as utility functions with different pseudo-reference pool sizes are sampled from using PaLM 2 as a translation system with 5-shot propmting.

Metric	SEG pairwise acc. en-de	SEG pairwise acc. zh-en	SEG Pearson en-de	SEG Pearson zh-en	SYS pairwise acc. en-de	SYS pairwise acc. zh-en	SYS Pearson en-de	SYS Pearson zh-en
Quality Estimation (Reference-Free) Metrics								
MBR-QE	**0.598**	**0.531**	**0.490**	0.452	**0.769**	**0.813**	**0.865**	**0.959**
COMETKIWI	0.572	0.509	0.432	**0.509**	0.692	0.758	0.674	0.866
UNITE-SRC	0.582	0.508	0.397	0.404	0.742	0.708	0.509	0.874
Reference-Based Metrics								
MetricX	0.605	0.544	0.549	0.581	0.829	0.867	0.847	0.920
COMET-22	0.594	0.536	0.512	0.585	0.790	0.886	0.771	0.942

Table 3: A comparison of our submission, denoted MBR-QE (scoring translations with MetricX against translations generated by our in-house MT system) to other QE metrics (top) and reference-based metrics (bottom). MBR-QE is overall the best-performing metric amongst the QE metrics, and it even improves over the reference-based metrics in system-level Pearson.

cess to a reference. Our experiments demonstrated that the choice of MBR utility function is important, the choice of MT system can impact downstream metric correlations, and the pseudo-reference pool size does not have a significant impact on results.

References

Rohan Anil, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, Eric Chu, Jonathan H. Clark, Laurent El Shafey, Yanping Huang, Kathy Meier-Hellstern, Gaurav Mishra, Erica Moreira, Mark Omernick, Kevin Robinson, Sebastian Ruder, Yi Tay, Kefan Xiao, Yuanzhong Xu, Yujing Zhang, Gustavo Hernandez Abrego, Junwhan Ahn, Jacob Austin, Paul Barham, Jan Botha, James Bradbury, Siddhartha Brahma, Kevin Brooks, Michele Catasta, Yong Cheng, Colin Cherry, Christopher A. Choquette-Choo, Aakanksha Chowdhery, Clément Crepy, Shachi Dave, Mostafa Dehghani, Sunipa Dev, Jacob Devlin, Mark Díaz, Nan Du, Ethan Dyer, Vlad Feinberg, Fangxiaoyu Feng, Vlad Fienber, Markus Freitag, Xavier Garcia, Sebastian Gehrmann, Lucas Gonzalez, Guy Gur-Ari, Steven Hand, Hadi Hashemi, Le Hou, Joshua Howland, Andrea Hu, Jeffrey Hui, Jeremy Hurwitz, Michael Isard, Abe Ittycheriah, Matthew Jagielski, Wenhao Jia, Kathleen Kenealy, Maxim Krikun, Sneha Kudugunta, Chang Lan, Katherine Lee, Benjamin Lee, Eric Li, Music Li, Wei Li, YaGuang Li, Jian Li, Hyeontaek Lim, Hanzhao Lin, Zhongtao Liu, Frederick Liu, Marcello Maggioni, Aroma Mahendru, Joshua Maynez, Vedant Misra, Maysam Moussalem, Zachary Nado, John Nham, Eric Ni, Andrew Nystrom, Alicia Parrish, Marie Pellat, Martin Polacek, Alex Polozov, Reiner Pope, Siyuan Qiao, Emily Reif, Bryan Richter, Parker Riley, Alex Castro Ros, Aurko Roy, Brennan Saeta, Rajkumar Samuel, Renee Shelby, Ambrose Slone, Daniel Smilkov, David R. So, Daniel Sohn, Simon Tokumine, Dasha Valter, Vijay Vasudevan, Kiran Vodrahalli, Xuezhi Wang, Pidong Wang, Zirui Wang, Tao Wang, John Wieting, Yuhuai Wu, Kelvin Xu, Yunhan Xu, Linting Xue, Pengcheng Yin, Jiahui Yu, Qiao Zhang, Steven Zheng, Ce Zheng, Weikang Zhou, Denny Zhou, Slav Petrov, and Yonghui Wu. 2023. PaLM 2 Technical Report.

Hyung Won Chung, Thibault Fevry, Henry Tsai, Melvin Johnson, and Sebastian Ruder. 2020. Rethinking embedding coupling in pre-trained language models. *arXiv preprint arXiv:2010.12821*.

Daniel Deutsch, George Foster, and Markus Freitag. 2023. Ties Matter: Modifying Kendall's Tau for Modern Metric Meta-Evaluation.

Bryan Eikema and Wilker Aziz. 2020. Is MAP decoding all you need? the inadequacy of the mode in neural machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4506–4520, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021a. Experts, Errors, and Context: A Large-Scale Study of Human Evaluation for Machine Translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.

Markus Freitag, Behrooz Ghorbani, and Patrick Fernandes. 2023. Epsilon sampling rocks: Investigating sampling strategies for minimum bayes risk decoding for machine translation.

Markus Freitag, David Grangier, Qijun Tan, and Bowen Liang. 2021b. Minimum bayes risk decoding with neural metrics of translation quality. *CoRR*, abs/2111.09388.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chikiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André F. T. Martins. 2022. Results of WMT22 metrics shared task: Stop using BLEU – neural metrics are better and more robust. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 46–68, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

John Hewitt, Christopher Manning, and Percy Liang. 2022. Truncation sampling as language model desmoothing. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3414–3427, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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.

Tom Kocmi, Christian Federmann, Roman Grundkiewicz, Marcin Junczys-Dowmunt, Hitokazu Matsushita, and Arul Menezes. 2021. To Ship or Not to Ship: An Extensive Evaluation of Automatic Metrics for Machine Translation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 478–494, Online. Association for Computational Linguistics.

Shankar Kumar and William Byrne. 2004. Minimum Bayes-risk decoding for statistical machine translation. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 169–176, Boston, Massachusetts, USA. Association for Computational Linguistics.

Arle Lommel, Hans Uszkoreit, and Aljoscha Burchardt. 2014. Multidimensional Quality Metrics (MQM): A Framework for Declaring and Describing Translation Quality Metrics. *Tradumàtica*, (12):0455–463.

Mathias Müller and Rico Sennrich. 2021. Understanding the properties of minimum Bayes risk decoding in neural machine translation. 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 259–272, Online. Association for Computational Linguistics.

Amy Pu, Hyung Won Chung, Ankur P Parikh, Sebastian Gehrmann, and Thibault Sellam. 2021. Learning compact metrics for mt. In *Proceedings of EMNLP*.

Ricardo Rei, JosÃ© G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and AndrÃ© F. T. Martins. 2022a. Comet-22: Unbabel-ist 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation*, pages 578–585, Abu Dhabi. Association for Computational Linguistics.

Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. 2022b. CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Thibault Sellam, Amy Pu, Hyung Won Chung, Sebastian Gehrmann, Qijun Tan, Markus Freitag, Dipanjan Das, and Ankur Parikh. 2020. Learning to evaluate translation beyond English: BLEURT submissions to the WMT metrics 2020 shared task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 921–927, Online. Association for Computational Linguistics.

Yu Wan, Keqin Bao, Dayiheng Liu, Baosong Yang, Derek F. Wong, Lidia S. Chao, Wenqiang Lei, and Jun Xie. 2022. Alibaba-translate china's submission for wmt2022 metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation*, pages 586–592, Abu Dhabi. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Chrysoula Zerva, Frédéric Blain, Ricardo Rei, Piyawat Lertvittayakumjorn, José G. C. de Souza, Steffen Eger, Diptesh Kanojia, Duarte Alves, Constantin Orăsan, Marina Fomicheva, André F. T. Martins, and Lucia Specia. 2022. Findings of the WMT 2022 shared task on quality estimation. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 69–99, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.