COMETOID: Distilling Strong Reference-based Machine Translation Metrics Into Even Stronger Quality Estimation Metrics

Thamme Gowda and **Tom Kocmi** and **Marcin Junczys-Dowmunt**
Microsoft Translator
Redmond, WA, USA
{thammegowda,tomkocmi,marcinjd}@microsoft.com

Abstract

This paper describes our submissions to the 2023 Conference on Machine Translation (WMT23) Metrics shared task. Knowledge distillation is commonly used to create smaller student models that mimic a larger teacher model while reducing the model size and hence inference cost in production. In this work, we apply knowledge distillation to machine translation evaluation metrics and distill existing reference-based teacher metrics into reference-free (quality estimation; QE) student metrics. We mainly focus on students of Unbabel's `COMET22` reference-based metric. When evaluating on the official WMT22 Metrics evaluation task, our distilled `Cometoid` QE metrics outperform all other QE metrics on that set while matching or out-performing the reference-based teacher metric. Our metrics never see the human ground-truth scores directly – only the teacher metric was trained on human scores by its original creators. We also distill `ChrF` sentence-level scores into a neural QE metric and find that our reference-free (and fully human-score-free) student metric `ChrFoid` outperforms its teacher metric by over 7% pairwise accuracy on the same WMT22 task, rivaling other existing QE metrics.[1]

1 Introduction

The Conference on Machine Translation (WMT) organizes an annual shared task for meta-evaluation of machine translation (MT) evaluation metrics (Freitag et al., 2022), where numerous MT evaluation metrics are proposed and revised each year. The MT metrics are broadly categorized as: (i) reference-based metrics, which score MT hypothesis against one or more reference translations from humans, and (ii) reference-free metrics, which do not require references and instead score hypothesis directly against the source sentence. Reference-free metrics, also known as quality estimation (QE) metrics, are an attractive choice in scenarios where reference translations are either unavailable or unreliable. However, currently, QE metrics lag behind the reference-based metrics by a considerable margin according to metrics meta-evaluation results (Freitag et al., 2022).

Knowledge distillation (KD) (Liang et al., 2008; Hinton et al., 2015) is commonly used to create smaller student models that mimic larger teacher models (Kim and Rush, 2016) which reduces computational cost when deploying models in production (Kim et al., 2019). Other use cases of KD in MT include distillation from auto-regressive teacher translation models to non-autoregressive students (Zhou et al., 2020) where the students "suffer" from an information bottleneck (here: no access to their own previous output in a time sequence) which impedes their performance when trained on original data. The simplified and probably smoothed output distribution of the teacher is easier to "digest" and often results in improved performance for the student.

In this work, we treat existing reference-based metrics as teachers and by applying knowledge distillation, we create reference-free student metrics that completely eliminate the need for references in evaluation. This is achieved by introducing a hard information bottleneck: just dropping the reference during training while keeping the original reference-based teacher score.

2 Experiments

2.1 Data Preparation

Our training set combines public and internal data sets. The public data is composed of all the MT systems submitted to WMT News (or General) Translation task between years 2009 and 2023. Our internal data set is prepared by translating parallel data using four MT systems: Moses SMT (Koehn et al., 2007), readily available bilingual NMT (Tiede-

[1] Metrics and usage instructions are available at: https://github.com/marian-nmt/wmt23-metrics

Proceedings of the Eighth Conference on Machine Translation (WMT), pages 751–755
December 6–7, 2023.

mann and Thottingal, 2020), multilingual transformer NMT (Gowda et al., 2021), and Microsoft Translator service. The number of examples in our training data is reported in Table 1.

For each training example i, let s_i, r_i and h_i, be source, reference and MT hypothesis segments, respectively. Each example is initially scored using teacher metrics that use reference translations and later references are dropped while training the student metrics. In this work, we use `COMET22` (Rei et al., 2022a) and `ChrF` (Popović, 2015) as teacher metrics. Teacher metrics that need source, reference and hypotheses as inputs – e.g. `COMET22` – produce training data in the form of $(s_i, r_i, h_i) \rightarrow \mathbb{R}$. The reference-only teachers such as `ChrF` produce $(r_i, h_i) \rightarrow \mathbb{R}$. All teacher sentence-level scores are normalized to the $[0, 1]$ range. For `COMET22` this required no change; for `ChrF`, computed by SacreBLEU (Post, 2018), we divide scores by 100.

Distilled students are trained on source-hypothesis pairs $(s_i, h_i) \rightarrow \mathbb{R}$ where the score is from the respective original reference-based teacher. Neither the references nor the human scores are directly seen by the student. However, indirectly, human scores may have been used by the teacher metric, which is the case for `COMET22`, but not for `ChrF`.

Dataset	Number of Examples
WMT09-21 systems	4.0M
WMT22 systems	0.5M
WMT23 systems	0.5M
Internal dataset	6.8M

Table 1: Training dataset size.

2.2 Model

Our distilled models have a similar architecture to `COMET-QE` models (Rei et al., 2020a),[2] and are implemented in MarianNMT (Junczys-Dowmunt et al., 2018), a fast NMT toolkit.[3] We slightly simplify the architecture by removing the encoder layer mixing and the batch-normalization present in the original implementation (neither seemed to contribute to any improvements), but we keep the general architecture of the added FFN regressor and the way how the encoder embeddings of source and hypothesis are combined into a single vector. Final output scores are squashed to the $[0, 1]$ range via a sigmoid function.

Similar to `COMET22`, we initialize our student models with the pretrained weights from InfoXLM (Chi et al., 2021),[4] specifically `infoxlm-large` that has 24 transformer layers (Vaswani et al., 2017).

We create the following four student models:

- `Cometoid22-wmt21`: student model distilled from `COMET22` and trained on scored data from the WMT News Translation task from 2009 - 2021 and similarly sized private data.
- `Cometoid22-wmt22`: Same as above, except we include system outputs submitted to WMT22. This is our *primary* submission to WMT23 Metrics shared task.
- `Cometoid22-wmt23`: Same as the above, except we include the system outputs submitted to WMT23.
- `ChrFoid-wmt23`: Same as the above, but we use segment-level `ChrF` as the teacher. This is an experimental model trained after the WMT23 Metrics shared-task deadline and has not been submitted to the shared task.

We evaluate our models on the WMT22 shared task while including WMT22 shared-task system outputs (MT systems and their reference-based scores) in the training data. This may seem suspicious at first, but note that our models do not use any human scores (the actual ground-truth of the task) in the training process, neither did the reference-based teachers which were trained before the WMT22 shared task. For the part of the evaluation where system submissions are available, this can be seen as part of an involved scoring process where the teacher remains blind to WMT22/WMT23 outputs, but the student does see them during distillation.

However, we are aware that this view may be disputable, hence we have submitted our `Cometoid22-wmt22` (blind to WMT23 outputs) as the primary submission to the WMT23 shared task instead of `Cometoid22-wmt23` that has seen scored WMT23 outputs (but not the actual ground-truth). We also provide results for `Cometoid22-wmt21` which is fully blind in regard to both – WMT22 and WMT23 outputs.

[2] https://huggingface.co/Unbabel/wmt20-comet-qe-da

[3] https://marian-nmt.github.io

[4] https://huggingface.co/microsoft/infoxlm-large

Metric	DA+SQM	MQM
Metricx_xxl_MQM_2020	**0.861**	0.850
Metricx_xl_MQM_2020	0.859	0.843
Cometoid22-wmt23 QE	0.859	0.803
Metricx_xxl_DA_2019	0.857	**0.865**
Cometoid22-wmt22 QE	0.857	0.807
Metricx_xl_DA_2019	0.850	**0.865**
Cometoid22-wmt21 QE	0.848	0.788
UniTE	0.847	0.828
COMET22	0.839	0.839
UniTE-ref	0.838	0.818
COMETKiwi(WMT22) QE	0.832	0.788
Cross-QE QE	0.832	0.781
ChrFoid-wmt23 QE	0.832	0.777
COMETKiwi (public) QE	0.816	0.770
ChrF	0.758	0.734

Table 2: WMT22 Evaluation system-level pairwise accuracy with DA+SQM (13 language pairs) and MQM (3 language pairs only). Rows are ordered by DA+SQM accuracy. Cometoid22 metrics are the best reference-free (QE) metrics.

2.3 Training

We ensure that scores from teacher metrics are in $[0, 1]$ range and optimize student metrics using cross-entropy loss.[5] Rei et al. (2020b) found that freezing InfoXLM layers for a number of epochs and training only the added parameters is beneficial, however, we were unable to confirm this with our metrics; we have fine-tuned all parameters till convergence according to perplexity on a small heldout subset of the data. For the final primary submission, we added the heldout data back to the training data and trained for the same number of iterations. We see minor improvements from Mixup regularization (Pinto et al., 2022) which we use for all student trainings.

3 Results and Analysis

We report system level pairwise accuracy obtained using `mt-metrics-eval`,[6] the official meta-evaluation pipeline used in WMT22 Metrics task. Table 2 shows that our COMETOID metrics are the top-performing QE metrics on the WMT22 Metrics data set. Interestingly, COMETOID student models also outperform the `COMET22` reference-based teacher model on DA+SQM data (we do fare worse on the smaller MQM data set only). Last but not least, `ChrFoid` – our student metric distilled from the `ChrF` (Popović, 2015) string-based metric – does surprisingly well and out-performs the teacher metric by a considerable margin despite now being reference-free.

4 Related Work

Reference-free (QE) metrics: Comet20-QE (Rei et al., 2020b) and CometKiwi22 (Rei et al., 2022c) are popular QE metrics. UniTE (Wan et al., 2022) supports inference in reference-free mode, in addition to reference-based mode. These metrics rely on scores from human evaluators during training and are limited by availability of high quality human ratings. Our metrics are trained with scores from teacher models and are trained on larger training data than what has been rated by human evaluators.

Distillation: Pu et al. (2021) and Rei et al. (2022b) apply knowledge distillation to the reference-based metrics, however, their distillation is aimed at reducing the model size for the sake of reducing computational cost during inference. Our work differs from theirs, as we distill with the aim of removing the need for human references at inference time.

5 Conclusion

We believe this work describes a perhaps simpler avenue towards more powerful QE metrics than proposed so far: *build strong reference-based first, next distill into even stronger QE metrics*. It further seems that performance improves with adding fully synthetic data (via adding larger amounts of inputs and automatically scored outputs). This effect seems also applicable to "dumb" metrics like `ChrF`: we have arrived at CHRFOID, a QE metric that has seen no human scores at all, and yet rivals the performance of the best previously available QE metrics. Knowledge distillation combined with a strong information bottleneck (reference-based to reference-free) seems to be the key in this new approach.

Limitations

Using available system outputs of the *same* shared task for training the metric may be a disputable approach even if the ground-truth was not used. Training time and model size of our distilled metrics are similar to the other popular metrics, and may be a limitation.

[5] Our preliminary experiments with mean absolute error loss performed inferior to cross-entropy.

[6] `https://github.com/google-research/mt-metrics-eval`

Ethics Statement

Knowledge distillation of existing models is always close to "model-stealing". The information provided here should be used responsibly and with publicly available models or according to terms of service.

Acknowledgements

Authors like to thank Roman Grundkiewicz for sharing some of the private data sets used in this work; Hieu Hoang for help with training Moses SMT, one of many MT systems used to create training data for distillation. Authors also like to thank the developers of `mt-metrics-eval` for creating and open-sourcing the meta-evaluation tool.

References

Zewen Chi, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-Ling Mao, Heyan Huang, and Ming Zhou. 2021. InfoXLM: An information-theoretic framework for cross-lingual language model pre-training. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3576–3588, Online. Association for Computational Linguistics.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu 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.

Thamme Gowda, Zhao Zhang, Chris Mattmann, and Jonathan May. 2021. Many-to-English machine translation tools, data, and pretrained models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 306–316, Online. Association for Computational Linguistics.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia. Association for Computational Linguistics.

Yoon Kim and Alexander M. Rush. 2016. Sequence-level knowledge distillation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1317–1327, Austin, Texas. Association for Computational Linguistics.

Young Jin Kim, Marcin Junczys-Dowmunt, Hany Hassan, Alham Fikri Aji, Kenneth Heafield, Roman Grundkiewicz, and Nikolay Bogoychev. 2019. From research to production and back: Ludicrously fast neural machine translation. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pages 280–288, Hong Kong. Association for Computational Linguistics.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

Percy Liang, Hal Daumé, and Dan Klein. 2008. Structure compilation: Trading structure for features. ICML '08, page 592–599, New York, NY, USA. Association for Computing Machinery.

Francesco Pinto, Harry Yang, Ser Nam Lim, Philip Torr, and Puneet Dokania. 2022. Using mixup as a regularizer can surprisingly improve accuracy & out-of-distribution robustness. In *Advances in Neural Information Processing Systems*, volume 35, pages 14608–14622. Curran Associates, Inc.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Amy Pu, Hyung Won Chung, Ankur Parikh, Sebastian Gehrmann, and Thibault Sellam. 2021. Learning compact metrics for MT. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 751–762, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022a. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Ricardo Rei, Ana C Farinha, José G.C. de Souza, Pedro G. Ramos, André F.T. Martins, Luisa Coheur, and Alon Lavie. 2022b. Searching for COMETINHO: The little metric that could. In *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, pages 61–70, Ghent, Belgium. European Association for Machine Translation.

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

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020b. Unbabel's participation in the WMT20 metrics shared task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 911–920, Online. 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. 2022c. 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.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT – building open translation services for the world. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 479–480, Lisboa, Portugal. European Association for Machine Translation.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Yu Wan, Dayiheng Liu, Baosong Yang, Haibo Zhang, Boxing Chen, Derek Wong, and Lidia Chao. 2022. UniTE: Unified translation evaluation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8117–8127, Dublin, Ireland. Association for Computational Linguistics.

Chunting Zhou, Jiatao Gu, and Graham Neubig. 2020. Understanding knowledge distillation in non-autoregressive machine translation. In *International Conference on Learning Representations*.