The SLIDE metric submission to the WMT 2023 metrics task

Vikas Raunak and **Tom Kocmi** and **Matt Post**
Microsoft
Redmond, Washington, USA
{viraunak,tom.kocmi,mattpost}@microsoft.com

Abstract

We describe our submission of a new metric, SLIDE (Raunak et al., 2023), to the WMT 2023 metrics task. SLIDE is a reference-free quality-estimation metric that works by constructing a fixed sentence-length window over the documents in a test set, concatenating chunks and then sending them for scoring as a single unit by COMET (Rei et al., 2022a). We find that SLIDE improves dramatically over its context-less counterpart on the two WMT22 evaluation campaigns (MQM and DA+SQM).

1 Introduction

Reference-based metrics such as COMET typically perform far above their source-based quality estimation (QE) counterparts. One explanation is that the human reference provides an answer or grounding to many types of translation ambiguities, such as pronoun selection, that may be impossible to predict from just a single input sentence alone. A handful of approaches have looked at extending metrics with source- and target-side context (Vernikos et al., 2022; Deutsch et al., 2023; Raunak et al., 2023) in hopes of providing stronger correlation with human judgments. We base our submission on SLIDE (Raunak et al., 2023), which explicitly postulates and provides evidence for the claim that source-side context may work to provide the same information as human references.

2 Metric settings

SLIDE is parameterized by (w, s), a window and a stride. The window, w, is a fixed-size sentence window that is moved across each document in a test set. The sentences in the window are concatenated on the source and system systems with a space, and then sent directly to the underlying QE model, COMETKiwi (Rei et al., 2022b) in our submission, for evaluation as a single chunk. The window is then incremented by s sentences, and a new value computed. These values are treated independently, summed and averaged over a test set in typical fashion. Documents that are shorter than the window size, and the "remainder" portions of documents that cannot be perfectly tiled by the window and stride, are skipped.

In practice, we used a (w, s) value of $(6, 6)$ for all languages except EN-DE and DE-EN. For those languages, the data was provided at the paragraph level. We therefore simply took the provided segmentations one-by-one, without providing a window or stride. We chose this value because it had some of the best reported results in Raunak et al. (2023, Figure 1). Table 1 repeats Table 2 from

Metric	MQM	DA+SQM
® metricx_xl_DA_2019	0.865	0.850
® metricx_xxl_MQM_2020	0.850	0.861
® BLEURT-20	0.847	0.827
® metricx_xl_MQM_2020	0.843	0.859
SLIDE$(6, 6)$	0.843	0.838
® COMET-22	0.839	0.839
® COMET-20	0.836	0.823
® Doc-COMET	0.836	0.810
® UniTE	0.828	0.847
® MS-COMET-22	0.828	0.830
® UniTE-ref	0.818	0.838
® MATESE	0.810	-
® YiSi-1	0.792	0.782
COMETKiwi (WMT-22)	0.788	0.832
COMETKiwi (public)	0.770	0.816
Doc-COMET	0.752	0.810
® chrF	0.734	0.758
® BLEU	0.708	0.704

Table 1: Pairwise system accuracy against the WMT22-MQM and DA+SQM annotations. Metrics that use a reference are marked with ®. We mark our entries in bold. **COMETKiwi (public)** uses no context. Our entry to the WMT23 task, SLIDE (6,6), improves over it in both settings.

Proceedings of the Eighth Conference on Machine Translation (WMT), pages 812–814
December 6–7, 2023.

their paper, depicting results on the WMT22 tasks with the pairwise accuracy (Kocmi et al., 2021). Our entries are marked in bold. SLIDE improves dramatically over its context-less counterpart. We also call attention to COMETKiwi (WMT-22); this is the number from the official submission to the task, which performs much better than the publicly available model.

3 Results

The WMT23 test set (Freitag et al., 2023; Kocmi et al., 2023) for each language pair comprises a set of documents containing between 1 and 173 lines, with a mean of 9.7 and a median of 7 across 14 language pairs.

At the time of publication, official results were not available, so we cannot comment on how well the strong results from Raunak et al. (2023) generalized to the new settings in WMT23.

We note also that we discovered after the submission that a bug in our code resulted in debugging output appearing in the data to be scored by COMET. This unfortunately affects the scores and means that SLIDE's placement in the official rankings are incorrect.

4 Conclusion

In this system description, we presented our submission to the WMT 2023 metrics task. SLIDE is designed as a reference-free quality-estimation metric which leverages the strength of contextual information by constructing a fixed sentence-length window over documents in a test set. The initial findings from Raunak et al. (2023) showcased the potential of SLIDE to deliver enhanced performance over context-less metrics, particularly in the WMT22 evaluation campaigns.

While we anticipate the official results from the WMT23 metrics task, bug in our code might have affected SLIDE's standing in the rankings.

We believe that SLIDE is a step forward in our collective endeavor to create metrics that align more closely with human judgments. Future works may explore optimizing window and stride configurations or integrating advanced algorithms to further exploit the potential of context in quality estimation tasks.

References

Daniel Deutsch, Juraj Juraska, Mara Finkelstein, and Markus Freitag. 2023. Training and meta-evaluating machine translation evaluation metrics at the paragraph level.

Markus Freitag, Nitika Mathur, Chi kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Tom Kocmi, Frédéric Blain, Daniel Deutsch, Craig Stewart, Chrysoula Zerva, Sheila Castilho, Alon Lavie, and George Foster. 2023. Results of wmt23 metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Philipp Koehn, Benjamin Marie, Christof Monz, Makoto Morishita, Kenton Murray, Masaaki Nagata, Toshiaki Nakazawa, Martin Popel, and Maja Popović. 2023. Findings of the 2023 conference on machine translation (WMT23). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). 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.

Vikas Raunak, Tom Kocmi, and Matt Post. 2023. Slide: Reference-free evaluation for machine translation using a sliding document window.

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, 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.

Giorgos Vernikos, Brian Thompson, Prashant Mathur, and Marcello Federico. 2022. Embarrassingly easy document-level MT metrics: How to convert any

pretrained metric into a document-level metric. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 118–128, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.