SLIDE: Reference-free Evaluation for Machine Translation using a Sliding Document Window

Anonymous ACL submission

Abstract

Reference-based metrics that operate at the sentence-level typically outperform quality estimation metrics, which have access only to the source and system output. This is unsurprising, since references resolve ambiguities that may be present in the source. In this paper, we investigate whether additional source context can effectively substitute for a reference. We present a metric named **SLIDE** (**SL**iding **D**ocument **E**valuator), which operates on blocks of sentences. SLIDE leverages a moving window that slides over each document in the test set, feeding each chunk of sentences into an unmodified, off-the-shelf quality estimation model. We find that SLIDE obtains significantly higher pairwise system accuracy than its sentence-level baseline, in some cases even eliminating the gap with reference-base metrics. This suggests that source context may provide the same information as a human reference in disambiguating source ambiguities. This finding is especially pertinent for reference-free document-level evaluation, wherein SLIDE could provide higher-quality pairwise system assessments while only requiring document boundary annotations.

1 Introduction

The prevailing approach for neural machine translation metrics is to work at the sentence-level, constructing sequences of contextualized encoder states from the source sentence, a reference translation, and a system output. The specific mechanics vary by metric, but a general approach, employed by COMET (Rei et al., 2020), is to pool these encodings into separate sentence-level embeddings, concatenate them, and fed them into a regressor, which is trained against human annotations. Quality Estimation (QE) approaches work similarly, but do not have access to a reference translation.

QE metrics typically trail their reference-based counterparts (Freitag et al., 2022), for obvious reasons. The default evaluation setting for QE is at the sentence-level. But just as there exist many linguistic phenomena that cannot be translated without context, these same phenomena also cannot be properly evaluated in *isolation*. As an example, consider the following English sentences *with context* and their translations into German.

(1) a. *I need my hat*. Where is **it**?
b. *Ich brauche meinen Hut*. Wo ist **er**?

Reference-based evaluation is aided by the fact that the human translation, presumably produced in context, resolves the **ambiguity**. QE approaches (operating at the sentence-level), on the other hand, cannot correctly score this translation. There are many other document-level phenomena that are also typically captured by references, many of which are subtle and hard to measure (Maruf et al., 2019).

It therefore stands to reason that providing disambiguating context could be a useful extension to (reference-free) QE metrics. In this work, we are motivated by two related ideas to address this gap: (i) neural metrics often make use of underlying language models trained on wider contexts, which means there is no real impediment to feeding them multiple sentences, and (ii) a sentence's evaluation will differ based on its order in a block of sentences, so it may be helpful to evaluate each sentence in multiple different contexts. We therefore experiment with a **strided window** approach applied to COMET, whose underlying encoder is InfoXLM (Lample and Conneau, 2019; Chi et al., 2021), trained on wide contexts. We apply a fixed-width sentence window and slide it across the documents within a test set, accumulating scores of each chunk in normal COMET fashion. We experiment with various windows and strides, and find that COMET-QE employed in this fashion outperforms its sentence-level reference-based counterparts in many settings. We conclude that this simple extension to QE might be profitably engaged wherever document boundary annotations are available.

#	docid	sentence
1	7759	There is no city in the high-risk…
2	7759	Iran's First Vice President…
3	7759	"Today there is neither concern…
4	doc0	I see, may I have your eReader…
5	doc0	To find your eReader's software…
6	doc0	1)Go to your Home screen.
7	doc0	2)Tap the More icon at the bottom…
8	doc0	3)Tap Settings.
9	doc0	4)Tap Device information.
10	doc0	5)Beside 'Software version', you'll…

Figure 1: SLIDE extraction for $(w = 4, s = 2)$. The solid green boxes denote extracted chunks, which are then joined with a space and sent to COMET as a single unit. The dashed red boxes denote partial documents: a document that is too short (top), and a document remainder (bottom).

2 The SLIDE Approach

Our main focus is the system ranking question as posed in Kocmi et al. (2021): given two systems, which one is the better one?

The task is to score system output over a test set, comprising multiple documents. We define a window size, w, specifying how many sentences to include as input, and a stride, $s \leq w$, defining how many sentences to advance the window. For a given (w, s) setting, the window is placed at the beginning of the document covering w sentences, and the chunk is sent to COMET as a single input. The window is then incremented by s sentences and a new value computed. This proceeds over all documents in a test set. This is depicted in Figure 1.

There are a few edge cases that must be considered, both pertaining to situations where we cannot fill a full window:

- *Documents smaller than* w. Windows are constrained by document boundaries. If a document is smaller than the window, we must therefore decide whether to include it.
- *Document remainder*. If a document length d is not evenly divisible by $w - s$, there will be a document remainder at the end of the document of size $(d \bmod w - s)$.

In our initial experiments, we ignore both types of partial documents, i.e., throwing them away, as if they didn't exist in order to evaluate only on chunks with appropriate context available. We will return to this question in Section A.1 to consider alternatives to throwing away partial documents.

Judgement Style	Lang-Pairs	System Pairs
MQM	3	274
DA + SQM	13	564

Table 1: Data used in the evaluation (Freitag et al., 2022)

Finally, we accumulate the scores from all chunks in a test set, and return their average as the system-level score. We call our metric SLIDE, for SLiding Document Evaluator.

3 Experiments

3.1 Evaluation Settings

Our experiments will explore the performance of various SLIDE approaches against the WMT22 metric evaluation task (Freitag et al., 2022). These tasks gathered human rankings of system outputs using different methods of collecting those human judgments: MQM and DA+SQM (Kocmi et al., 2022). The human scores induce a ranking of all system outputs, for each language pair. The N scored system outputs are collected into a set of $\binom{N}{2}$ pairwise system comparisons across all language pairs, considering pairs from each set of systems per language pair in isolation. Metrics are then scored based on their ability to correctly predict the better system in each pair. Table 1 contains information about the data used in the evaluation.

Our key evaluation metric is the pairwise system-level accuracy (Kocmi et al., 2021), a percentage of pairwise system pairs that a metric correctly distinguishes. We experiment with the following SLIDE settings:

- Window values $1 \leq w \leq 10$ and strides of $1 \leq s \leq w$.
- Multiple state-of-the-art QE models (COMET-QE-20 and COMET-QE-22). Further, in addition to QE models, we also look at a reference-based metric, COMET22.[1]
- Incorporating partial documents (§ A.1).

3.2 Results

Figure 2 contains a heatmap depicting pairwise system-level accuracies for each (w, s) value for three models on the MQM task (en-de, en-ru, and zh-en). The first two are QE20 and QE22, as defined above. We also include the reference-based

[1]COMET model `wmt22-comet-da`.

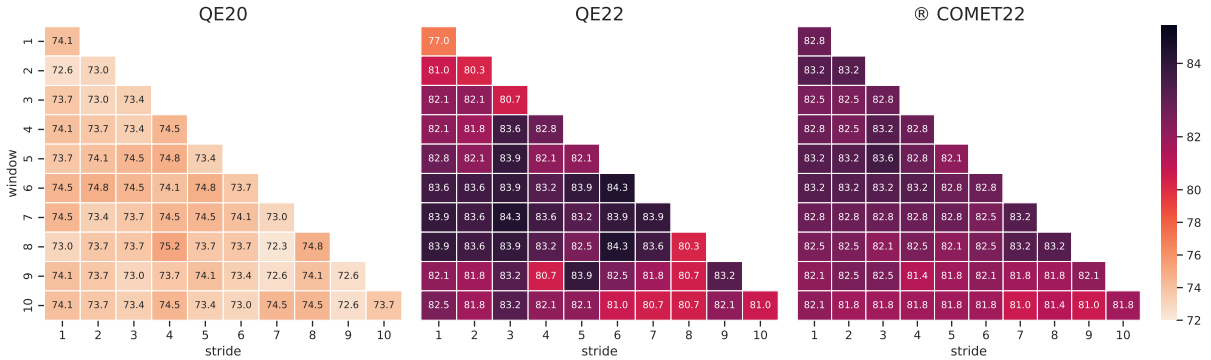


Figure 2: Plot of window vs. stride accuracies for QE20, QE22, and ®COMET22 models on the WMT22 MQM task (en-de, en-ru, and zh-en). Neither COMET20-QE nor the reference-based ®COMET22 sees much improvement from adding context, but the QE22 model does.

®COMET22[2] model.[3] A number of trends emerge. The QE20 does not fair well at all with context, with very few points in the grid even improving over the sentence-level baseline (point $(1, 1)$) at all, and not in any consistent fashion. We see the same thing for model ®COMET22, corroborating the finding of Deutsch et al. (2023), as well as our hypothesis that source context may be redundant *with* the reference.

With the QE22 model, however, the results show a clear trend. Adding context helps significantly; the *worst* points in the grid (8-8 and 2-2) improve by 3.3 points over the no-context baseline. Performance seems to rise as more context is added, although once too much context is used, the points begin to decrease again. Within a particular window (row), it does not seem to matter too much which stride is used, particularly with window sizes 6–8. Further, we hypothesize that the purported large gains based on the utilization of source context is not visible in QE20 owing to the lack of cross-attention in its architecture, which strongly suggests that it is not able to leverage the fine-grained source context, which is demonstrably exploited through better modeling of longer-range dependencies in the cross-attention based QE22 model.

Comparing chunked ($w = s$) vs. overlapped ($w > s$) values, we also see no particular pattern. Increasing the context size (up to *some* point) matters, but overlapping the document chunks neither helps nor hurts, on average.

Next, we situate some of these results against leading metrics and other metrics of interest from the WMT22 task in Table 2.[4] We include the best variant of SLIDE, the worst variant, and, in a nod towards model selection, the best variant with a stride of 1. Here, we can see that SLIDE propels quality estimation up the chart, where it even competes with reference-based versions of the underlying evaluators. In particular, SLIDE $(6, 6)$ outperforms ®COMET22, and a number of variants are at the same level as ®COMET20. While these results don't answer the difficult question of model selection, the trends empirically validate our key hypothesis (that source-based context could substitute for the reference in evaluation). We note again that even the worst SLIDE model is already quite an improvement over the baseline.

3.3 Analysis & Further Experiments

We note that it is not obvious that adding context to sentences might improve metric performance when one considers that the machine translation systems being compared are all sentence-based translations. Any information that is newly available for the evaluator would therefore *not* have been available to the underlying translation engine. As a result, the improvements the metric is picking up on must come from the new sources included in the evaluation. We further experiment with the QE22 model with different variants of SLIDE in Appendix A.

4 Related Work

The vast majority of work in machine translation metrics is focused on sentence-level evaluation. At

[2] As a service to the reader, we annotate reference-using models with ®.

[3] COMET model `wmt22-comet-da`

[4] We observe that the public model release of COMETKiwi, `wmt22-cometkiwi-da`, performs notably worse than the one submitted to the WMT22 task. From private correspondence with the authors, we learned that that the model entered into the shared task was an ensemble.

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
SLIDE$(7, 1)$	0.839	0.814
® COMET-20	0.836	0.823
® Doc-COMET(2)	0.836	
® UniTE	0.828	0.847
® MS-COMET-22	0.828	0.830
® UniTE-ref	0.818	0.838
® MATESE	0.810	
SLIDE$(2, 1)$	0.807	0.825
® YiSi-1	0.792	0.782
COMETKiwi (WMT-22)	0.788	0.832
Doc-COMET(2)	0.737	0.810
COMETKiwi (Public)	0.770	0.816
® chrF	0.734	0.758
® BLEU	0.708	0.704

Table 2: Pairwise system accuracy against the WMT22-MQM annotations. Metrics that use a reference are marked with ®. Our entries are of the form SLIDE (w, s). We have retained many other WMT22 scores for comparison purposes.

a high level, it is useful to distinguish two uses of context for machine translation evaluation.

In the first setting, contextualized metrics are designed to test document-translation capabilities of a model. Chief among these are contrastive test sets, in which a model is tested in its ability to rank good translations from bad ones in a contextualized manner (Müller et al., 2018; Bawden et al., 2018; Voita et al., 2019; Lopes et al., 2020). Other approaches have employed test-time NLP tool chains to target and reward correct prediction of discourse phenomena (Jiang et al., 2022; Fernandes et al., 2023).

In the second setting, metrics make use of context to provide or refine their model of what makes a good translation. This is useful even for ranking sentence-based translation systems. The first such system may be DocCOMET (Vernikos et al., 2022), which used context to modify the encodings of sentences, but then removed that context before invoking COMET's classifier. They looked at both reference-based and QE metrics, and evaluated on Pearson's correlation against human scores. Context-COMET (Hendy et al., 2023) took an approach similar to that described here, but was not evaluated at all. More recently, Deutsch et al. (2023) showed that paragraph-level evaluation works just as well for reference-based metrics, even with underlying metrics trained in a sentence-level fashion. However, they did not experiment with reference-less metrics (the quality estimation task).

Our approach here fits within this second setting. In contrast to this prior work, our approach requires no changes to the underlying codebase, is evaluated on pairwise system-level accuracy, and focuses on the quality estimation task, where context is the most promising.

5 Conclusion and Future Work

Incorporating context into COMET-QE provides critical information that appears to help the metric better adjudicate the difference between systems. Even just a single sentence of context *drastically* improves the ability of the model to discriminate between systems. The method works well even though the scores are accumulated over groups of sentences, which is different from the sentence-level manner in which COMET is trained, and even when the accumulated scores come from overlapping blocks. The results here required no change to the underlying sentence-based evaluators, and in that sense come for free, so long as document boundary annotations are available. We therefore recommend that source context be included with any neural quality estimation metric. An interesting further evaluation would throw some document-context translation engines into the mix; we speculate that contextualized quality estimation (like that provided by SLIDE) should help discriminate those systems too.

6 Limitations

In order to leverage the quality-estimation evaluation benefits from SLIDE, document boundaries must be available on the source sentences. Even though this is a benign requirement in most cases, this is a strict limitation on the application of SLIDE in cases where the document boundary annotations are unavailable.

References

Rachel Bawden, Rico Sennrich, Alexandra Birch, and Barry Haddow. 2018. Evaluating discourse phenomena in neural machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1304–1313, New Orleans, Louisiana. Association for Computational Linguistics.

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.

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

Patrick Fernandes, Kayo Yin, Emmy Liu, André Martins, and Graham Neubig. 2023. When does translation require context? a data-driven, multilingual exploration. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 606–626, Toronto, Canada. 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.

Naman Goyal, Jingfei Du, Myle Ott, Giri Anantharaman, and Alexis Conneau. 2021. Larger-scale transformers for multilingual masked language modeling. In *Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021)*, pages 29–33, Online. Association for Computational Linguistics.

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How good are gpt models at machine translation? a comprehensive evaluation.

Yuchen Jiang, Tianyu Liu, Shuming Ma, Dongdong Zhang, Jian Yang, Haoyang Huang, Rico Sennrich, Ryan Cotterell, Mrinmaya Sachan, and Ming Zhou. 2022. BlonDe: An automatic evaluation metric for document-level machine translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1550–1565, Seattle, United States. Association for Computational Linguistics.

Tom Kocmi, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Thamme Gowda, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Rebecca Knowles, Philipp Koehn, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Michal Novák, Martin Popel, and Maja Popović. 2022. Findings of the 2022 conference on machine translation (WMT22). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 1–45, Abu Dhabi, United Arab Emirates (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.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *CoRR*, abs/1901.07291.

António Lopes, M. Amin Farajian, Rachel Bawden, Michael Zhang, and André F. T. Martins. 2020. Document-level neural MT: A systematic comparison. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 225–234, Lisboa, Portugal. European Association for Machine Translation.

Sameen Maruf, Fahimeh Saleh, and Gholamreza Haffari. 2019. A survey on document-level machine translation: Methods and evaluation. *CoRR*, abs/1912.08494.

Mathias Müller, Annette Rios, Elena Voita, and Rico Sennrich. 2018. A large-scale test set for the evaluation of context-aware pronoun translation in neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 61–72, Brussels, Belgium. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. 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.

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.

Elena Voita, Rico Sennrich, and Ivan Titov. 2019. When a good translation is wrong in context: Context-aware machine translation improves on deixis, ellipsis, and lexical cohesion. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1198–1212, Florence, Italy. Association for Computational Linguistics.

A Further SLIDE Experiments

In order to further test the hypothesis that source-based context could substitute for the reference in evaluation, we experiment with two more variants of the SLIDE approach.

A.1 Partial windows

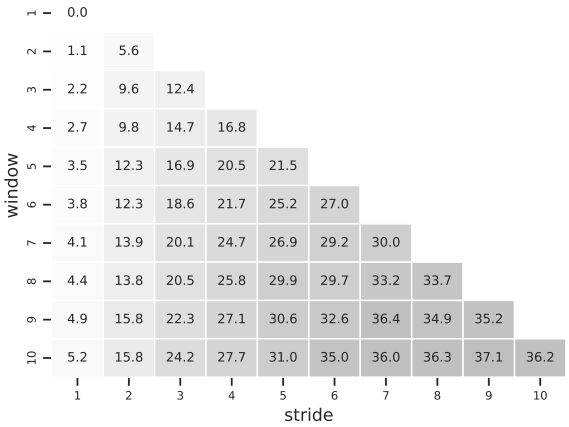


Figure 3: Percentage of original-document *sentences* that are dropped when partial documents are ignored.

The main approach only builds full window sizes. This means that two types of sentences are excluded: (i) those occurring in test set documents that are shorter than the window size, and (ii) those for which $w - s$ does not evenly divide the document length. Examples of each of those cases are depicted in Figure 1. Further, Figure 3 lists the percentage of sentences that are left out under the default chunking strategy that omits these two types.

We construct an experiment that includes these partials. Documents smaller than the window size are included as a complete document, and remainders are added as as partial chunk. The resulting heatmap for QE22 is depicted in Figure 4 (a).

A problem with blindly incorporating these partials is that they are not properly weighted; in producing the document-level score, all chunks are averaged with a uniform weight. It stands to reason that smaller chunks should contribute less to the overall score, proportional to their sentence length. Figure 4 (b) adjusts for this, upweighting each chunk by multiplying its score by the number of sentences it contains, prior to averaging.

w	en-de	en-ru	zh-en
1	0	0	0
2	0	0	0.2
3	0	0	1.1
4	0.1	0.1	4.5
5	1.0	0.6	10.2
6	6.2	5.7	19.5
7	13.9	13.9	30.5
8	24.3	24.2	43.3
9	31.9	31.3	55.3
10	38.3	38.3	64.1

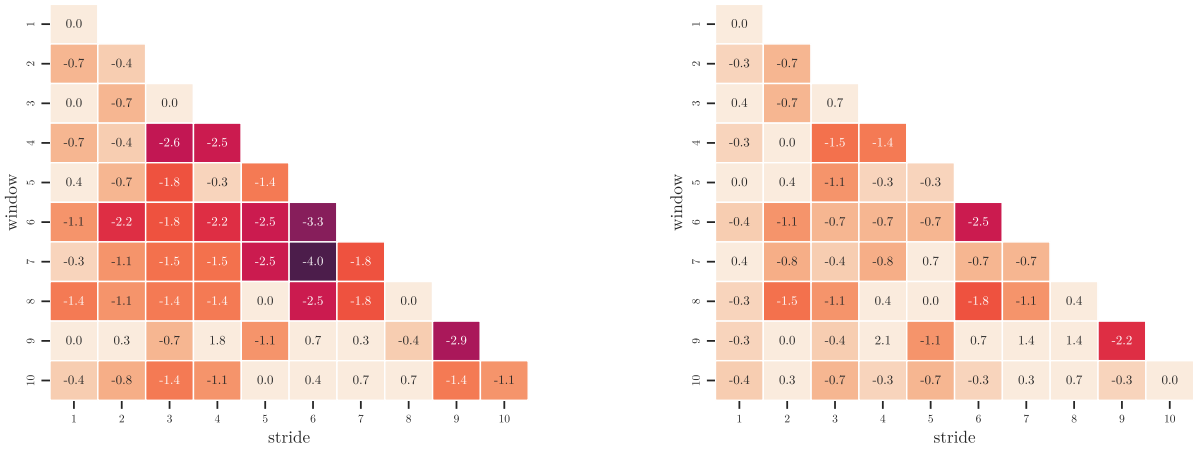
Table 3: Percentage of *chunks* with a tokenized length > 512, which means they will be cropped by the underlying InfoXLM model (Chi et al., 2021; Goyal et al., 2021).

Comparing to the middle figure from Figure 2, we see a notable drop in system-level accuracy for both plots relative to the only-full-window approach of SLIDE. Both systems, however, **continue to improve** over the original, context-less QE22 model. We also see that that weighting the partials helps mitigate the problem.

A.2 Variable chunking

The problem of partial chunks is not the only design choice when instantiating SLIDE. Another choice has to deal with managing the maximum token length of the underlying model used by COMET, InfoXLM. COMET-Kiwi (the QE22 model) encodes the entire input as as single string, instead of separately encoding the source and the system output. This raises the chance that we will hit the maximum token length. In fact, this is the case. Table 3 lists the percentage of input chunks that rise above the maximum encoding length.

We therefore further experiment with a general solution that will address both the issue of partial document chunks and truncation. Our approach is to retain the notion of a window and a stride, with sentences as atomic units. However, the window now specifies a maximum number of tokens. For a given token-based window size w_t, we add (source, system) sentence pairs so long as the number of tokens in the sentence pairs does not exceed w_t. The stride parameter remains purely sentence-based. To produce the system-level score, we use



(a) Partial windows.

(b) Partial windows with weighting.

Figure 4: Heatmaps of MQM accuracy difference relative to Figure 2(b) for SLIDE when all partials are incorporated. In (a), they are treated as equal chunks when producing the document-level score; in (b), the score of each chunk is upweighted based on the number of sentences it contains. Both variants improve over context-less QE22, but are generally worse than SLIDE_f

the sentence-based weighting approach, that multiples each chunk score by the number of sentences in that chunk, prior to averaging.

Unfortunately, this token based chunking approach failed to produce a consistent gain. Setting a maximum window size of 500 and greedily building chunks as large as possible produced an MQM score of 0.818, considerably lower than the SLIDE results in Table 2.