Long-context Reference-based MT Quality Estimation

Sami Ul Haq[1], **Chinonso Cynthia Osuji**[1], **Thiago Castro Ferreira**[2], **Brian Davis**[1], **Sheila Castilho**[1]
[1]ADAPT Centre, Dublin City University, Dublin, Ireland
[2]Fluminense Federal University, Brazil
{firstname.lastname}@adaptcentre.ie
thiago.castro.ferreira@gmail.com

Abstract

In this paper, we present our submission to the Tenth Conference on Machine Translation (WMT25) Shared Task on Automated Translation Quality Evaluation. Our systems are built upon the COMET framework and trained to predict segment-level Error Span Annotation (ESA) scores using augmented long-context data. To construct long-context training data, we concatenate in-domain, human-annotated sentences and compute a weighted average of their scores. We integrate multiple human judgment datasets (MQM, SQM, and DA) by normalising their scales and train multilingual regression models to predict quality scores from the source, hypothesis, and reference translations. Experimental results show that incorporating long-context information improves correlations with human judgments compared to models trained only on short segments.

1 Introduction

The automatic evaluation of machine translation (MT) is a crucial component of MT research and development. While expert-based human evaluation remains the gold standard, automatic evaluation offers fast and scalable judgments, enabling rapid feedback for optimizing model parameters. Traditionally, automatic MT evaluation metrics have relied on basic lexical-level features, such as counting matching n-grams between the MT hypothesis and the reference translation. Metrics such as BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), and ChrF (Popović, 2015) remain popular due to their lightweight design and computational efficiency (Marie et al., 2021). More recently, neural metrics (either trained on human annotations or based on pre-trained language models) have demonstrated superior capability in comparing and assessing MT quality, often outperforming traditional lexical-based metrics (Freitag et al., 2022). These neural approaches leverage large-scale multilingual data during training and achieve strong performance even when translations diverge lexically from the reference.

This paper presents DCU_ADAPT's submission to the WMT25[1] MT Evaluation Shared Task. The primary focus of this year's task is on systems capable of evaluating translation quality in context, where the context spans entire documents or multiple consecutive segments. We participated in the segment-level quality score prediction track for English–Czech, English–Russian, English–Japanese, and English–Chinese, employing models based on the COMET framework (Rei et al., 2020a).

In our contribution to the shared task, we explore methods for leveraging synthetic data alongside the capabilities of pre-trained and cross-lingual models to predict MT quality estimates for long-sequence or multi-sentence units. Human judgments of MT quality are typically available as short segment-level scores, such as DA (Graham et al., 2017), MQM (Lommel et al., 2014), and SQM (Barrault et al., 2019). Recent pre-trained models support larger context windows and can handle long-sequence inputs, improving discourse-level resolution (Dai et al., 2019)—albeit at the cost of increased memory and computational requirements. However, most existing automatic evaluation metrics (AEMs) predict scores at the sentence level, and those designed for document-level evaluation often perform only shallow context integration during inference (Vernikos et al., 2022). We propose a data augmentation strategy to train multilingual models on long-context annotated data, enabling them to better exploit broader context and reduce inconsistencies caused by sentence-level ambiguity.

The exploration of context in MT is a well-established topic and, in recent years, has become a focal point, driven by the need to incorporate context into both MT systems and their evalua-

[1]https://www2.statmt.org/wmt25/mteval-subtask.html

Proceedings of the Tenth Conference on Machine Translation (WMT), Volume 2: Shared Task Papers, pages 905–912
November 8–9, 2025.

tion methodologies (Bawden et al., 2017; Castilho et al., 2020; Maruf et al., 2021; Castilho et al., 2023; Castilho and Knowles, 2024). There is now broad consensus on the value of document-level evaluation. Since 2019, WMT has conducted human evaluations at the document level, providing evaluators with access to context even when collecting segment-level ratings (Akhbardeh et al., 2021; Kocmi et al., 2022a, 2023, 2024a). Research indicates that the appropriate context span is critical for reliable MT evaluation, with Castilho et al. (2020) showing that incorporating relevant context spans can yield more accurate assessments of translation quality, thereby improving the evaluation process. Several techniques have been proposed to extend evaluation to the document level or to incorporate multi-sentence context into automatic evaluation metrics (Jiang et al., 2021; Vernikos et al., 2022; Rei et al., 2022; Kocmi et al., 2022b; Raunak et al., 2024).

For this shared task submission, we use the Estimator model from the COMET framework (Rei et al., 2022), which learns MT quality from human evaluation data such as MQM and DA. To create long-context training data, we combine multiple annotations using a weighted average alongside the original annotations. Our experiments show promising progress toward improved correlation in multi-sentence-level MT quality estimation. Fine-tuning multilingual embedding models demonstrates that it is possible to achieve high correlations with human judgments when evaluating long segments, rather than relying solely on sentence-level score predictions.

We release the data and code produced during this research.

2 Corpora

We used the human annotations from previous WMT shared tasks (Kocmi et al., 2022a, 2023, 2024a) for training our models, which includes human annotations from MQM, DA, and SQM. We train and evaluate our models for English (en) to Czech (cs), Japanese (ja), Chinese (zh), and Russian (ru) language pairs. MQM scores are derived from error annotations and can range from $-\infty$ to 100. Since our goal is to predict ESA scores (Kocmi et al., 2024b), which range between 0 and 100, we normalise (using Equation 1, where x is original and x' is normalised score) and rescale the scores to the [0, 1] interval for training models on the combined dataset.

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)} \tag{1}$$

We create augmented training data for long-span MT quality estimation by taking a weighted average of segment-level scores. During augmentation, we concatenate multiple samples (i.e., 2, 3, 4, and 5 segments) to form long-span texts. To construct the long-span MT evaluation dataset, adjacent short segments are concatenated, and a document-level quality score is computed as a length-weighted average of their original scores. The weighting is based on the total number of characters in the source and machine-translated texts, as formalized in Equation 2.

Let s_1, s_2 are short segments (e.g., source-translation pairs), and raw_1, raw_2 human evaluation scores for these segments. C_1 and C_2 are the total character count of each segment e.g., $C_i = len(s_i)$.

Then the document-level score (raw_{doc}) is calculated as:

$$\text{raw}_{\text{doc}} = \frac{C_1 \cdot \text{raw}_1 + C_2 \cdot \text{raw}_2}{C_1 + C_2} \tag{2}$$

This equation[2] computes a weighted average of two segment-level scores, where the weight is determined by the combined character count of the source and MT for each segment. Longer segments contribute more to the final score, reflecting their higher informational content. The augmentation process increased the average segment length of dataset from 16.84 words per segment to 52.99 words per segment. Since the augmented segments were added to the original dataset, the overall size of the dataset increased by a factor of two.

We then create training, test, and validation sets by randomly sampling segments from the training data. The statistics of the final training, validation, and test sets are shown in Table 1.

3 Experimental Setup

Our system is built on top of the COMET package, utilizing the `comet-train` and `comet-score` commands to train and evaluate our models. We fine-tuned the pre-trained model `Unbabel/wmt22-comet-da`, originally trained on

[2]We adapted the implementation of data augmentation from the Huggingface repository: https://huggingface.co/datasets/ymoslem/wmt-da-human-evaluation-long-context

Type	Split	No. of segments			
		en→cs	en→ja	en→ru	en→zh
DA	train	345K	119K	350K	533K
	dev	43K	15K	43K	66K
	test	38K	13K	39K	59K
MQM	train	–	–	172K	–
	dev	–	–	21K	–
	test	–	–	19K	–
SQM	train	64K	75K	64K	75K
	dev	8K	9K	8K	9K
	test	7K	8K	7K	8K
Total		**505K**	**238K**	**723K**	**750K**

Table 1: Dataset statistics by evaluation type, split, and language pair (K represents values in thousands).

DA data, as well as the multilingual pre-trained model `FacebookAI/xlm-roberta-base` (Liu et al., 2019), using A100 GPU. The `xlm-roberta` model was pre-trained on 2.5TB of filtered CommonCrawl data containing 100 languages and has approximately 279 million parameters.

We trained the models for up to 5 epochs and employed early stopping when the Spearman correlation on the development set did not improve for two consecutive evaluations. For each language pair, the augmented dataset contained five times the original data; however, due to limited GPU memory, we faced out-of-memory issues and restricted training to augmentations with up to two segments. The training process with the augmented dataset took approximately 10 hours for each model.

We trained one baseline model (the fine-tuned version of `wmt22-comet-da`) and three main models (one primary and two secondary submissions) using the data augmentation approach, retaining only the last two checkpoints for each. Spearman correlation on the test split was used to select the best checkpoint per language pair, and this best model was used for the final submission.

For official WMT `test25`[3] evaluation, the full segment was used since it fell within the maximum sequence length (512) supported by both the baseline and fine-tuned models. COMET scores typically range between 0 and 1, but can sometimes exceed 1, indicating exceptionally high-quality segments. To align with the ESA metric's scoring strategy (Kocmi et al., 2024b), we upscaled and rounded the scores to a range between 0 and 100.

[3]https://github.com/wmt-conference/wmt25-mteval/blob/main/data/testset/mteval-task1-test25.tsv.gz

4 Results

As described in Section 2, our experiments use the normalized versions of multiple human annotations collected from previous WMT shared tasks. To evaluate and compare our approach, we applied a similar augmentation method to construct long-sequence test data, incorporating annotations from DA, MQM, and SQM. Our baseline sentence-level quality estimation models are `wmt22-comet-da` (referred to as COMET-22) and BERTSCORE. COMET-22-LS, a fine-tuned version of `wmt22-comet-da` on long-span data, and ROBERTA-LS, fine-tuned from `FacebookAI/xlm-roberta-base`, both serve as long-sequence quality score prediction models. Following Rei et al. (2020a), the models were trained on triplets of (source, hypothesis, reference) and output a score between 0 and 1 reflecting the translation quality relative to both source and reference.

We used the Pearson correlation coefficient to evaluate the models' performance. Segment-level Pearson correlations on the self-test set are presented in Tables 2 and 3. Our results indicate that metrics from models trained on long-context inputs generally outperform sentence-level metrics, in some cases by a significant margin.

The annotation-wise segment-level correlation results in Table 2 demonstrate that the unsupervised baseline metric, BERTSCORE, exhibits relatively weak correlations across all language pairs. The sentence-level COMET-22 model shows improved correlations, especially for DA annotations, reflecting its training on DA data. Moreover, it outperforms BERTSCORE on SQM and MQM annotations, indicating its ability to mimic human annotations by learning from data. Our fine-tuned long-sequence baseline model, COMET-22-LS, surpasses the sentence-level baselines, achieving performance close to, and in some cases better than, our primary submission models based on ROBERTA-LS. Notably, COMET-22-LS achieves results on MQM annotations that are very close to those of ROBERTA-LS, outperforming sentence-level baselines by a substantial margin. This suggests that training on longer context sequences provides considerable benefits, especially for complex annotation types like MQM, which capture fine-grained translation errors.

Table 3 summarizes correlation results across all language pairs using joint annotations. Across nearly all language pairs, our models outperform

	DA				SQM				MQM
	en→ru	**en→cs**	**en→ja**	**en→zh**	**en→ru**	**en→cs**	**en→ja**	**en→zh**	**en→ru**
BERTSCORE	0.399	0.480	0.418	0.328	0.290	0.207	0.290	0.107	-0.04
COMET-22	0.571	0.637	0.511	0.443	0.450	0.400	0.352	0.220	0.075
COMET-22-LS	0.848	**0.894**	0.777	**0.772**	**0.572**	0.701	**0.668**	0.585	0.866
ROBERTA-LS	**0.874**	0.890	**0.780**	0.770	0.557	**0.707**	0.666	**0.600**	**0.874**

Table 2: System-level Pearson correlation results for MQM, SQM, and DA annotations. Bold values indicate systems that achieved higher correlations with human judgments. LS denotes models trained on long-span input data.

	en→ru	**en→cs**	**en→ja**	**en→zh**	avg.
BERTSCORE	0.216	0.344	0.354	0.217	0.283
COMET-22	0.365	0.519	0.432	0.331	0.412
COMET-22-LS	0.762	0.798	0.722	0.679	0.740
ROBERTA-LS	**0.768**	**0.799**	**0.723**	**0.685**	**0.744**

Table 3: Segment-level Pearson correlation scores for the language pairs `en-ru`, `en-cs`, `en-ja`, and `en-zh`. Bold values indicate stronger correlations with human judgments. LS denotes models trained on long-span input data.

baseline metrics in correlation with human judgments. Ideally, COMET-22-LS, fine-tuned from `wmt22-comet-da` (which is already trained for evaluation tasks), should have outperformed ROBERTA-LS. However, the performance difference between the two models is not substantial. This may be because `wmt22-comet-da` was trained only on DA data, while the current task also includes SQM and MQM annotations, which typically follow different scoring strategies. Furthermore, as mentioned earlier, due to resource constraints, we trained our models for only five epochs and on a limited number of augmented segments. With larger augmentation and more robust training, the performance gap between the two models may become more pronounced.

These results suggest that combining multiple segments within the same document or domain is more effective than independently scoring segmented sentences and averaging their scores (Raunak et al., 2024). The improvement may be attributed to the model's ability to capture contextual information across long-sequence segments, thereby enabling more context-aware quality estimation.

However, handling longer text sequences poses challenges due to the input size limitations of the underlying models, as highlighted by Gong et al. (2020). This necessitates careful segmentation and score-averaging strategies to compute scores at the paragraph or document level. We also conduct a preliminary analysis of the score distributions (Figure 1) and find that sentence-level baseline scores (COMET-22) are mostly concentrated between 60 and 100, with a pronounced peak around 90. In contrast, the ROBERTA-LS model, trained on multi-sentence inputs, produces a more widely spread score distribution that better reflects the variability typically observed in human judgments (Toral et al., 2018). This wider spread may be due to the long-span training data being based on weighted average scores that encompass a broader range of score scales. By contrast, the narrower distribution of COMET-22 scores could stem from the characteristics of the human-annotated data on which it was trained, where non-expert evaluators have been shown to assign disproportionately higher fluency and adequacy ratings, resulting in smaller score gaps and reduced variance compared to expert assessments (Toral et al., 2018). This indicates that models trained with longer context are more sensitive to subtle quality differences, reflecting a more nuanced understanding of translation quality.

5 Related Work

In recent years, metrics based on large pre-trained models have emerged as strong alternatives to traditional n-gram-based approaches, enabling better capture of semantic similarity between words beyond mere lexical matching. These metrics broadly fall into two categories: embedding-based metrics and fine-tuned metrics.

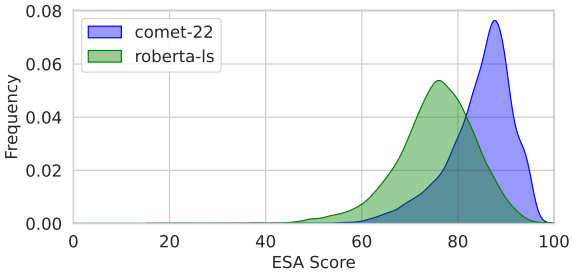

Figure 1: Distribution of segment-level scores assigned by sentence-level COMET-22 and long-sequence ROBERTA-LS model.

Embedding-based metrics typically represent an advancement over n-gram matching by using dense word representations in an embedding space to compute scores that reflect semantic similarity between reference and hypothesis segments. Notable examples include YISI-1 (Lo, 2019), MOVERSCORE (Chow et al., 2019), and BERTSCORE (Zhang et al., 2019), which leverage embedding models for soft alignment between two segments to capture semantic similarity effectively.

Fine-tuned metrics, on the other hand, involve learnable models such as RUSE (Shimanaka et al., 2018), BLEURT (Sellam et al., 2020), and COMET (Rei et al., 2020a, 2022) that directly optimize underlying embedding models to maximize correlation with human judgments. These models have demonstrated promising results in producing reliable quality scores for test sets such as DA or MQM. While most of these metrics perform reference-based evaluation, recent advancements leveraging highly multilingual pre-trained encoders like multilingual BERT (Devlin, 2018) and RoBERTa (Liu et al., 2019; Conneau et al., 2019) have enabled reference-less systems to show encouraging correlations with human judgments (Freitag et al., 2023).

Most automatic evaluation approaches rely on decontextualized assessments, where translations are judged at the sentence level. However, sentences are often inherently ambiguous, and incorporating document-level context has been shown to be beneficial for both MT and its evaluation (Läubli et al., 2018; Castilho et al., 2020; Castilho and Knowles, 2024; Vernikos et al., 2022). Consequently, a few automatic metrics have been developed to extend evaluation beyond the word or sentence level (Vernikos et al., 2022; Jiang et al., 2021). These metrics aim to address discourse-level phenomena such as lexical consistency, coherence, ellipsis, and pronoun resolution (Voita et al., 2018; Bawden et al., 2017).

However, existing methods typically use a limited number of surrounding sentences as context, allowing models to incorporate neighboring information when embedding each sentence and computing scores at the sentence level (Rei et al., 2020b; Vernikos et al., 2022; Hu et al., 2023). In contrast, long-sequence or document-level evaluation processes the entire segment as a single input, enabling deeper discourse-level resolution and offering a promising yet still underexplored avenue for improving alignment with human judgments.

6 Conclusion

In this paper, we present DCU_ADAPT's contribution to the WMT25 MT Evaluation Shared Task. We leverage the COMET framework and train regression models to predict ESA quality scores. In line with the Shared Task goals, we augment the provided training data and optimize our models to evaluate long, multi-sentence units of text. By fine-tuning multilingual models for cross-lingual transfer, we utilize source, reference, and hypothesis as inputs. Our primary submission — a fine-tuned pre-trained model trained on augmented data — demonstrates higher or otherwise competitive correlation levels with human judgments across multiple languages. Further investigation comparing long-text evaluation after segmentation with sentence-level evaluation is a promising direction for future work.

The data and code produced during this Shared Task participation are available at: https://github.com/sami-haq99/CAEMT/tree/main/wmt-2025-submission.

Limitations

We trained our models using augmented long-segment level scores from the MQM, DA, and SQM datasets. However, we only evaluated the models on self-test data carefully extracted from the training set; evaluating on benchmark datasets will better clarify the true benefits of our approach. Additionally, we normalized the data using min-max normalization to combine different datasets and upscaled the predicted scores to match the ESA metric's score range. Furthermore, our training and testing were conducted on GPUs with at least 40GB of memory; due to time constraints, we were unable to evaluate performance on CPUs.

Ethics Statement

Our research focuses on evaluating long-sequence outputs of MT systems using quality estimation models trained on augmented data. We are committed to conducting and reporting our evaluations with the highest levels of transparency and fairness. By upholding these principles, we aim to contribute to reliable and objective assessment practices in MT evaluation.

Acknowledgements

This work was conducted with the financial support of the Research Ireland Centre for Research Training in Digitally-Enhanced Reality (d-real) under Grant No. 18/CRT/6224, and Research Ireland Centre for Research Training in Artificial Intelligence under Grant No. 18/CRT/6223. For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

The Authors also benefit from being members of the ADAPT SFI Research Centre at Dublin City University, funded by the Science Foundation Ireland under Grant Agreement No. 13/RC/2106_P2.

References

Farhad Akhbardeh, Arkady Arkhangorodsky, Magdalena Biesialska, Ondřej Bojar, Rajen Chatterjee, Vishrav Chaudhary, Marta R. Costa-jussa, Cristina España-Bonet, Angela Fan, Christian Federmann, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Leonie Harter, Kenneth Heafield, Christopher Homan, Matthias Huck, Kwabena Amponsah-Kaakyire, Jungo Kasai, Daniel Khashabi, Kevin Knight, Tom Kocmi, Philipp Koehn, Nicholas Lourie, Christof Monz, Makoto Morishita, Masaaki Nagata, Ajay Nagesh, Toshiaki Nakazawa, Matteo Negri, Santanu Pal, Allahsera Auguste Tapo, Marco Turchi, Valentin Vydrin, and Marcos Zampieri. 2021. Findings of the 2021 Conference on Machine Translation (WMT21). In *Proceedings of the Sixth Conference on Machine Translation*, pages 1–88, Online. Association for Computational Linguistics.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Loïc Barrault, Ondřej Bojar, Marta R Costa-Jussa, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, et al. 2019. Findings of the 2019 conference on machine translation (wmt19). ACL.

Rachel Bawden, Rico Sennrich, Alexandra Birch, and Barry Haddow. 2017. Evaluating discourse phenomena in neural machine translation. *arXiv preprint arXiv:1711.00513*.

Sheila Castilho and Rebecca Knowles. 2024. A survey of context in neural machine translation and its evaluation. *Natural Language Processing*, pages 1–31.

Sheila Castilho, Clodagh Mallon, Rahel Meister, and Shengya Yue. 2023. Do online machine translation systems care for context? what about a gpt model?

Sheila Castilho, Maja Popović, and Andy Way. 2020. On context span needed for machine translation evaluation. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3735–3742, Marseille, France. European Language Resources Association.

Julian Chow, Lucia Specia, and Pranava Madhyastha. 2019. WMDO: Fluency-based word mover's distance for machine translation evaluation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 494–500, Florence, Italy. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V Le, and Ruslan Salakhutdinov. 2019. Transformer-xl: Attentive language models beyond a fixed-length context. *arXiv preprint arXiv:1901.02860*.

Jacob Devlin. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Markus Freitag, Nitika Mathur, Chi-kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Tom Kocmi, Frédéric Blain, Daniel Deutsch, Craig Stewart, et al. 2023. Results of wmt23 metrics shared task: Metrics might be guilty but references are not innocent. In *Proceedings of the Eighth Conference on Machine Translation*, pages 578–628.

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.

Hongyu Gong, Yelong Shen, Dian Yu, Jianshu Chen, and Dong Yu. 2020. Recurrent chunking mechanisms for long-text machine reading comprehension.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2017. Can machine translation systems be evaluated by the crowd alone. *Natural Language Engineering*, 23(1):3–30.

Xinyu Hu, Xunjian Yin, and Xiaojun Wan. 2023. Exploring context-aware evaluation metrics for machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15291–15298.

Yuchen Eleanor Jiang, Tianyu Liu, Shuming Ma, Dongdong Zhang, Jian Yang, Haoyang Huang, Rico Sennrich, Ryan Cotterell, Mrinmaya Sachan, and Ming Zhou. 2021. Blonde: An automatic evaluation metric for document-level machine translation. *arXiv preprint arXiv:2103.11878*.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, et al. 2024a. Findings of the wmt24 general machine translation shared task: the llm era is here but mt is not solved yet. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1–46.

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, Makoto Nagata, Toshiaki Nakazawa, Martin Popel, Maja Popović, and Mariya Shmatova. 2023. Findings of the 2023 Conference on Machine Translation (WMT23): LLMs Are Here but Not Quite There Yet. In *Proceedings of the Eighth Conference on Machine Translation*, pages 1–42, Singapore. 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. 2022a. 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, Hitokazu Matsushita, and Christian Federmann. 2022b. MS-COMET: More and better human judgements improve metric performance. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 541–548, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Tom Kocmi, Vilém Zouhar, Eleftherios Avramidis, Roman Grundkiewicz, Marzena Karpinska, Maja Popović, Mrinmaya Sachan, and Mariya Shmatova. 2024b. Error span annotation: A balanced approach for human evaluation of machine translation. *arXiv preprint arXiv:2406.11580*.

Samuel Läubli, Rico Sennrich, and Martin Volk. 2018. Has machine translation achieved human parity? a case for document-level evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4791–4796, Brussels, Belgium. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Chi-kiu Lo. 2019. YiSi - a unified semantic MT quality evaluation and estimation metric for languages with different levels of available resources. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 507–513, Florence, Italy. 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.

Pierre Marie et al. 2021. Scientific credibility of machine translation research: A meta-evaluation of 769 papers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 566–577. Association for Computational Linguistics.

Sameen Maruf, Fahimeh Saleh, and Gholamreza Haffari. 2021. A survey on document-level neural machine translation: Methods and evaluation. *ACM Computing Surveys (CSUR)*, 54(2):1–36.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

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.

Vikas Raunak, Tom Kocmi, and Matt Post. 2024. 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.

2022. 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, 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, Catarina Farinha, and Alon Lavie. 2020b. Unbabel's participation in the wmt20 metrics shared task. *arXiv preprint arXiv:2010.15535*.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

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: Shared Task Papers*, pages 751–758, Belgium, Brussels. Association for Computational Linguistics.

Antonio Toral, Sheila Castilho, Ke Hu, and Andy Way. 2018. Attaining the Unattainable? Reassessing Claims of Human Parity in Neural Machine Translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 113–123, Brussels, Belgium. 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. *arXiv preprint arXiv:2209.13654*.

Elena Voita, Pavel Serdyukov, Rico Sennrich, and Ivan Titov. 2018. Context-aware neural machine translation learns anaphora resolution. *arXiv preprint arXiv:1805.10163*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.