Training and Meta-Evaluating Machine Translation Evaluation Metrics at the Paragraph Level

Daniel Deutsch, Juraj Juraska, Mara Finkelstein, and Markus Freitag
Google
{dandeutsch,jjuraska,marafin,freitag}@google.com

Abstract

As research on machine translation moves to translating text beyond the sentence level, it remains unclear how effective automatic evaluation metrics are at scoring longer translations. In this work, we first propose a method for creating paragraph-level data for training and meta-evaluating metrics from existing sentence-level data. Then, we use these new datasets to benchmark existing sentence-level metrics as well as train learned metrics at the paragraph level. Interestingly, our experimental results demonstrate that using sentence-level metrics to score entire paragraphs is equally as effective as using a metric designed to work at the paragraph level. We speculate this result can be attributed to properties of the task of reference-based evaluation as well as limitations of our datasets with respect to capturing all types of phenomena that occur in paragraph-level translations.

1 Introduction

Automatic evaluation metrics have always been a critical component to the progress of research on machine translation (MT). As the field of MT moves beyond translating individual sentences to translating full paragraphs, book chapters, or documents (Tu et al., 2018; Sun et al., 2022; Thai et al., 2022; Jiang et al., 2023; Post and Junczys-Dowmunt, 2023), automatic metrics need to be designed to work on these longer texts.

Currently, how well automatic metrics agree with human judgments of paragraph translation quality is an open question.[1] Few studies have meta-evaluated metrics on longer texts, and those that have are focused on the literary domain and are limited in the size of the evaluation dataset (Jiang et al., 2022; Thai et al., 2022; Karpinska and Iyyer, 2023). In this work, we investigate training and meta-evaluating metrics for scoring paragraph translations using the benchmark Workshop on Machine Translation (WMT) datasets that are widely used for metric development (Freitag et al., 2022).

Due to the scarcity of human ratings of paragraph translations, we propose a method to create paragraph-level training and meta-evaluation datasets from the existing WMT sentence-level datasets (§3). Although these ratings are typically only used at the sentence level, they were collected on contiguous paragraphs and performed with document context, so they can be used as paragraph-level datasets. We repurpose these datasets to benchmark existing sentence-level metrics as well as train new paragraph-level metrics for scoring paragraph translations (§4).

Our experimental results are somewhat surprising. We find that there appears to be little evidence that training on paragraph-level data is beneficial—at least given the limitations of our experimental setup. Using metrics trained on sentence-level data only to directly score full paragraphs achieves comparable agreement to human ratings as metrics trained on paragraph-level data (§6.1). Sentence-level metrics appear to generalize well to inputs much longer than they were trained on (§6.2).

We hypothesize these observations can be explained by the nature of evaluating translations and characteristics of our paragraph-level dataset (§7). We speculate that long range dependencies—which paragraph-level metrics can model but sentence-level likely do not—may not be too important for achieving high agreement with human ratings. Further, due to the fact that our training and evaluation datasets assume a sentence alignment between the reference and hypothesis paragraphs, certain translation phenomena that sentence-level metrics may struggle to handle, like sentence or information reordering, are not well represented in the dataset,

[1] Translation beyond the sentence level is often referred to as document-level MT. However, there is no clear definition for the term “document.” We use “paragraph” in this work because we feel it most accurately describes the length of text in our datasets. See §2 for more details on this.

Proceedings of the Eighth Conference on Machine Translation (WMT), pages 996–1013
December 6–7, 2023.

limiting our ability to show the benefits of training on paragraph-level ratings.

The contributions of our work include (1) a method for constructing paragraph-level training and meta-evaluation datasets from sentence-level ratings, (2) an experimental study that demonstrates the comparable performance of sentence- and paragraph-level metrics, and (3) an analysis that aims to provide an explanation for our experimental observations.

2 Terminology

Throughout this paper, we use terms like segment, sentence, paragraph, and document to refer to different lengths of text. To the best of our knowledge, there are no agreed upon definitions for these terms in the MT literature, so here we define how they are used for the rest of the paper.

We refer to the input text to an MT system or evaluation metric as a *segment*, irrespective of its length. Traditionally, segments in MT have been roughly equivalent to one sentence, although sometimes they can be short phrases or even longer than a single sentence. Regardless, we use *sentence* to refer to this unit of text since it accurately describes the most common text length that is widely used in MT.

Our work investigates evaluating *paragraphs* of text, which we define to be multi-sentence segments. We do not require that the paragraphs used in this work obey the traditional definition of a paragraph (i.e., a unit of text separated by a newline character). We refrain from calling this unit of text a *document*—which we consider to be all of the possible input text—since each document can be broken down into multiple paragraphs and the term paragraph more accurately describes the length of text we use.

3 Paragraph-Level Datasets

The two main sources for training and meta-evaluating MT metrics are the direct assessment (DA) and Multi-dimensional Quality Metrics (MQM; Lommel et al., 2014; Freitag et al., 2021a) datasets that the Workshop on Machine Translation (WMT) has collected as part of the yearly metrics shared task (Freitag et al., 2022). The DA ratings were done by a mixture of expert and non-expert raters (depending on whether the translation direction is into or out of English) who assigned a quality score in the range 0-100 to translated sentences.

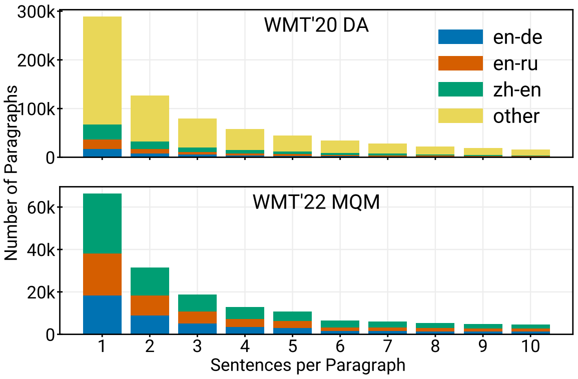

Figure 1: The number of contiguous paragraphs for the given number of sentences per paragraph where each sentence is rated by the same rater. Actual values are included in Appendix A.

Because of differences in rater behavior, the DA scores are z-normalized per rater.[2] In MQM, expert raters identify error spans in translated sentences and assign each error a category and severity level, which are used to calculate a score for that error. A sentence's MQM score is defined as the sum of the errors' scores.

Training and meta-evaluating metrics at the paragraph level requires a collection of translated paragraphs and paragraph-level quality scores. Luckily, the DA data since 2019 and the MQM data can be considered to be paragraph-level ratings. The ratings were performed on contiguous blocks of sentences that were translated by the same system (e.g., the first k sentences per document are rated for a system). Although the scores were collected at the sentence level, the ratings were done in context, meaning the raters had access to the document context for a sentence, so the scores should reflect paragraph- or document-level phenomena like discourse errors. Therefore, we use the sentence-level DA and MQM data to construct paragraph-level datasets as follows.

For each document translated by a system, we run sliding window of size k sentences from the start to the end. If all k sentences in the window have been rated, those k sentences are concatenated together to become a paragraph instance and the window shifts by k. Otherwise, the sliding window shifts by 1 and the process repeats. To maintain consistency between the sentence scores within a paragraph, we additionally require that every sen-

[2] The methodology for collecting DA ratings has changed throughout the years. See Barrault et al. (2020) for the description in 2020, the most recent year used in this work.

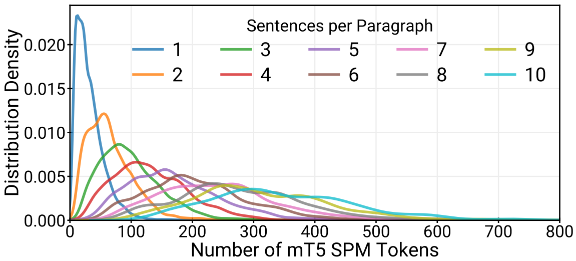

Figure 2: The distribution of paragraph lengths in SPM tokens (i.e., sub-word tokens; Kudo and Richardson, 2018) on the WMT'22 MQM dataset for different numbers of sentences per paragraph. Additional datasets' distributions are included in Appendix A.

tence is scored by the same rater. Then, we define the paragraph-level scores to be the average DA z-score or sum of MQM scores for each sentence in the paragraph.[3] The result is a dataset of rated paragraph translations of k sentences each.

We apply this dataset construction approach to the DA and MQM data for $k = 1, 2, \ldots, 10$ sentences per paragraph. The number of paragraphs is shown in Figure 1 and the distribution of the lengths of the new translated paragraphs is shown in Figure 2. As k increases, the number of paragraphs decreases because there are fewer candidate paragraphs, while the length of the paragraphs increases, roughly by an expected factor of k.

These paragraph-level DA and MQM datasets are used to train and meta-evaluate paragraph-level metrics for the rest of this paper.

4 Paragraph-Level Metrics

We explore two different methods for creating paragraph metrics: directly applying sentence-level metrics to paragraphs (§4.1) and training metrics on paragraph-level data (§4.2).

4.1 Applying Sentence-Level Metrics on Paragraphs

Although automatic metrics that have been used to evaluate sentence-level MT were not explicitly designed to evaluate paragraphs, they can be repurposed to score paragraphs in different ways.

First, the input paragraph can be treated as if it were one long segment and passed to the metric to calculate a score. For metrics that use bag-of-n-grams representations, like BLEU (Papineni et al., 2002), there is no input length limitation. However, some learned metrics, like BLEURT (Sellam et al., 2020), have a maximum possible sequence length due to restrictions related to neural network architectures. Therefore, the length of the input paragraph is restricted in some cases.

Then, if there is assumed to be an alignment between the source, reference, and hypothesis sentences within a paragraph (as is in the case with our datasets), a paragraph score can be calculated by averaging the sentence-level metric's score for each of the k individual sentences. While this sliding window approach more closely aligns how the metrics are being used to how they were designed, we argue this approach is less than ideal because the 1:1 sentence alignment between the source and hypothesis translations will not always exist. However, this approach is useful for understanding and analyzing the behavior of metrics when they are used to score full paragraphs directly.

4.2 Learning Paragraph-Level Metrics

While sentence-level metrics can be repurposed to score paragraphs, the lengths of the input paragraphs are significantly longer than the lengths of individual sentences (compare $k = 1$ to $k > 1$ in Figure 2) and there may be cross-sentence dependencies that are not learned by sentence-level metrics. Therefore, we explore creating a metric specifically for paragraph-level data.

To do so, we train a BLEURT-style regression model on the paragraph-level datasets: The reference and hypothesis paragraphs are tokenized and concatenated together (separated by a special token), then passed as input to a neural network. The network is then trained to predict the hypothesis paragraph's ground-truth quality score. Sections 5.2 and 5.4 contain more information about the model's architecture and implementation details.

It is desirable for the paragraph-level metric to be able to score paragraphs of any length, so we train the metric on paragraphs composed of $k = 1, 2, \ldots, 10$ sentences. Because the number of paragraph instances decreases significantly as k increases (see Figure 1), longer paragraphs will rarely be seen during training. Therefore, we explore two different techniques for weighting training data: one that selects paragraphs uniformly at random

[3] Summing MQM scores was done to generalize an MQM rating for paragraphs since a sentence's MQM score is the total error weight for that sentence. The choice of summing or averaging does not matter for metric meta-evaluation because the correlations are scale invariant.

and one that performs a stratified sample so the training data is composed of an equal number of paragraphs for each value of k.

Next, we describe the experimental setup to evaluate the paragraph-level metrics.

5 Experimental Setup

5.1 Datasets

The paragraph-level datasets used in our experiments are described in Section 3. The WMT'19 (Ma et al., 2019) and '20 (Mathur et al., 2020) paragraph-level DA data is used for training the metrics described in this work, and all metrics are evaluated on the WMT'21 (Freitag et al., 2021b) and WMT'22 (Freitag et al., 2022) paragraph-level MQM data. For both DA and MQM, we use $k = 1, 2, \ldots, 10$ sentences per paragraph. The different paragraph lengths are combined during training but separated for evaluation.

We additionally analyze the behavior of the metrics that we train on judgments collected by Karpinska and Iyyer (2023) on literary translations. Their dataset contains human preference judgments between paragraph translations. The translations come from translation models that translated the input one sentence a time in isolation, one sentence at a time in context, the full paragraph directly, and Google Translate. We evaluate how frequently the metrics agree with the human preference judgments.

5.2 Metrics

Paragraph-Level Metrics. We train two different paragraph-level metrics, one for each of the different weighting techniques, uniform and stratified sampling (see §4.2). We refer to these metrics as PARA-UNIF and PARA-STRAT.

Our metric uses the same architecture as the Metric-X WMT'22 metrics shared task submission (Freitag et al., 2022). The metric builds on the mT5 encoder-decoder language model (Xue et al., 2021), which was originally designed to be a sequence-to-sequence language model. We repurpose the model for our regression task as follows. The inputs to the encoder are the hypothesis and reference translations separated by a special token, and a single dummy token is passed as the first input to the decoder. We arbitrarily selected a reserved vocabulary token, then trained the model so that token's output logit in the first decoding step becomes the score for the input hypothesis translation. This modification of the sequence-to-sequence architecture for regression allows us to utilize all of the pre-trained weights from mT5.

The maximum input sequence length to our metric is 1024 SPM tokens (Kudo and Richardson, 2018). The inputs are truncated during training or inference if the input is larger than 1024.[4] In the worst case, this happens up to 27% of the time on the MQM data for 10 sentences per paragraph (see Appendix A for specific statistics.)

Sentence-Level Baseline. In addition to the paragraph-level metrics, we train a sentence-level version that is trained on the same DA data but only $k = 1$ sentences per paragraph. This baseline metric can be used to directly compare to the paragraph-level metrics that we train because the model architecture, training procedure, etc., are identical. The only difference is the training data. This metric is referred to as SENT-BASE.

Other Metrics. In addition to the metrics described in this paper, we evaluate BLEU (Papineni et al., 2002), COMET-22 (Rei et al., 2020, 2022), and PaLM-2 from Fernandes et al. (2023) as sentence-level metrics applied to paragraphs (i.e., §4.1) and document-level metric BlonDE (Jiang et al., 2022). BLEU scores translations using lexical n-gram overlap, and COMET-22 is a learned regression metric that first embeds the input hypothesis, reference, and source, combines them to a joint representation, then finally predicts a score.

The metric from Fernandes et al. (2023) is based on the PaLM-2 large language model (Anil et al., 2023). We evaluate both the zero shot version, in which PaLM-2 is prompted to score a translation on a scale from 0 to 100, and the regression version that finetunes PaLM-2 on MQM ratings to predict a floating point quality score, similar to COMET. Our analysis includes the Bison variant of PaLM-2.

BlonDE evaluates discourse phenomena in document translations via a set of automatically extracted features. It was designed to evaluate texts longer than paragraphs, like book chapters, but we compare against it in this work. BlonDE is available in English only.

We use the SacreBLEU (Post, 2018) implementation of BLEU and the `Unbabel/wmt22-comet-da` COMET-22 model that was trained on sentence-level WMT DA data from 2017-2020.[5]

[4] We experimentally saw no benefit from removing sequences longer than 1024 tokens during training.

[5] Note that the COMET-22 scores we report come from

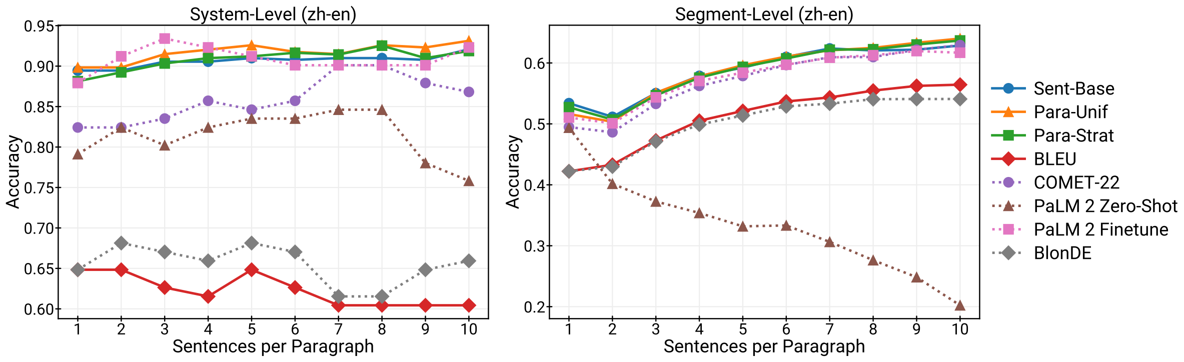

Figure 3: As the number of sentences per paragraph increases, the pairwise accuracy scores (y-axis) of the metrics appears to either not decrease (system-level, left) or increase (segment-level, right). This suggests that accurately scoring a paragraph is an easier task than an individual sentence, even for metrics that are not trained on paragraph-level examples. The results of metrics trained in this work presented here are an average of 5 different runs. Results for other language pairs follow the same trend and are included in Appendix B.

5.3 Meta-Evaluation Metrics

The quality of an evaluation metric is quantified by measuring the correlation of its scores to human ratings of translation quality, a process known as meta-evaluation. In this work, we meta-evaluate metrics using pairwise accuracy at both the system and segment levels.[6] A brief overview of how these accuracy statistics are calculated follows.

At the system-level, an automatic metric and human score is calculated per system by averaging scores over paragraphs. The system-level pairwise accuracy is then computed by enumerating all possible pairs of systems and then calculating the proportion of those pairs for which the automatic metric and human ground-truth ratings agree on their ranking (Kocmi et al., 2021). Thus, the accuracy score can be interpreted as the proportion of pairs of systems that the metric ranked correctly.

At the segment-level, we report segment-level pairwise accuracy using the group-by-item variant of the segment-level correlation in combination with tie calibration (Deutsch et al., 2023). In contrast to system-level accuracy, the group-by-item segment-level correlation calculates the proportion of pairs of *translations* of the same source segment that the metric ranks correctly, then averages that accuracy score over all source segments. The segments used in this evaluation are paragraphs, thus the interpretation of this accuracy score is the proportion of pairs of translations of the same source paragraph that are ranked correctly by the metric.

Because humans frequently assign the same score to translations and regression-based evaluation metrics almost never predict two translations are tied, we follow Deutsch et al. (2023) and run tie calibration before calculating the segment-level accuracy. This procedure automatically introduces ties in the metrics' scores by searching for an ϵ difference in metric score that, when two translations are considered to be a tie if they differ by less than ϵ, achieves the highest accuracy score. We report the accuracy score that corresponds to the best ϵ.

Results using Pearson's correlation follow similar trends to the accuracy results and are available in Appendix B.

5.4 Implementation Details

Our learned metrics are implemented with TensorFlow (Abadi et al., 2015) in the T5X library (Roberts et al., 2022). They are initialized with the XXL version of mT5, which contains 13B parameters. It is trained for a maximum of 20k steps and a batch size of 128 using Adafactor (Shazeer and Stern, 2018) on 64 v3 TPUs. Checkpoint selection was done by selecting the step that has the highest average segment-level pairwise accuracy across language pairs and all values of k sentences per paragraph after applying tie calibration. In general, we observed the specific checkpoint selection strategy was not too important.

only the reference-based regression model, not the ensemble that was submitted to the WMT'22 metrics shared task.

[6] The segment-level correlation could be referred to as a paragraph-level correlation in this work because the segments we evaluate on are paragraphs. However, to be consistent with the evaluation literature, we still use the term segment-level correlation.

6 Results

First, we directly evaluate how well metrics perform when used to directly score paragraphs (§6.1), then we further examine the behavior of different paragraph-level metrics by analyzing their performances with the context of their sentence-level counterparts (§6.2).

6.1 Paragraph-Level Evaluation

Figure 3 plots the system- and segment-level correlation results for different numbers of k sentences per paragraph. Each metric is used to directly score a full paragraph even if the metric was not designed to do so (e.g., Sent-Base or COMET-22). There are several interesting observations.

Paragraph-Level Performance. First, as the length of the paragraphs increases, the system-level correlations remain relatively steady or increase and the segment-level correlations clearly improve for all metrics, except for PaLM-2 zero-shot. This is evidence that scoring paragraphs is an easier task than scoring individual sentences, a result that is counterintuitive; scoring more text should seemingly be a harder task. We hypothesize this result is explained by the fact that some noise in the human and metric scores is averaged away, leaving more reliable signals as the paragraphs get longer. If the metric scores are unbiased estimators, their agreement with human rating should then increase.

PaLM-2 zero-shot is an outlier in this case because it predicts a large number of ties between translations. Prompting large language models for MT evaluation is known to result in the model predicting a small number of unique scores, resulting in many ties (Kocmi and Federmann, 2023; Fernandes et al., 2023). As the length of the paragraph increases, the number of MQM ties decreases. Since pairwise accuracy penalizes incorrect tie predictions, the zero shot model has worse performance on longer texts. See Figure 4 for a visualization of the number of ties in the PaLM-2 output and MQM scores.

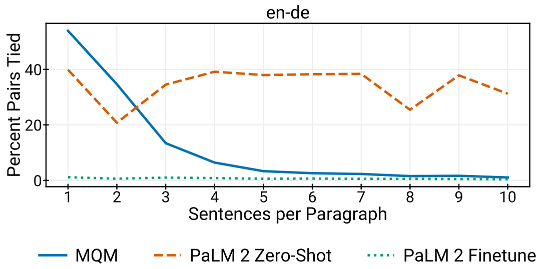

Figure 4: There are fewer MQM ties as the number of sentences per paragraph increases. The finetuned PaLM-2 model outputs a very small number of ties, whereas the zero-shot model consistently predicts a large number of ties. Since the pairwise accuracy meta-evaluation metric penalizes metrics for incorrect tie predictions, the zero-shot model will have worse performance as the inputs get longer.

Dataset	1 Sent. per Para.			10 Sent. per Para.		
	25th	50th	75th	25th	50th	75th
WMT'19 DA	20	31	47	300	362	431
WMT'20 DA	24	38	58	318	410	524
WMT'21 MQM	28	41	57	370	433	516
WMT'22 MQM	15	27	43	265	333	426

Table 1: The SPM token lengths for the given percentiles are in general around 10 times larger with 10 sentences per paragraph compared to a single sentence. Visualizations of the distributions for every paragraph length can be found in Appendix A.

Sentence vs. Paragraph Level. Then, there appears to be little evidence that training on paragraph-level examples results in better correlations to human ratings on paragraph-level test data. For instance, increasing the weight of the paragraph-level data during training does not help compared to uniformly sampling data (compare Para-Strat to Para-Unif). Further, the baseline metric Sent-Base that shares the same architecture as our paragraph-level metrics but is only trained on sentence-level data ($k = 1$) performs just as well as the paragraph-level metrics. This observation is additionally supported by COMET-22's results. The difference between the metrics we train versus COMET is relatively constant for all values of k, demonstrating that COMET is not systematically worse on longer inputs.

The generalization of sentence-level metrics on paragraph-level data is rather surprising. The length of the inputs for scoring paragraphs is up to 10x longer than those for scoring sentences (see Table 1). Even though the length of the test data is out-of-distribution with respect to the training data, the sentence-level metrics predict reliable scores on the paragraph-level data. Next, we further analyze the sentence-level metrics to better understand their scores.

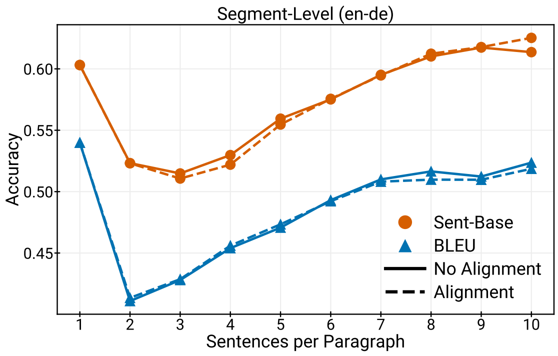

Figure 5: Metrics that score a paragraph directly (solid line) versus those that assume an alignment between the reference and hypothesis and calculate a score by averaging across the k sentence-level values (dashed line) perform very similarly. The drop from 1 sentence to 2 sentences per paragraph is likely due to the fact that a large number of ties in the ground-truth get broken, so introducing more ties via tie calibration is less helpful since doing so is right less often. This phenomenon does not happen with Pearson correlations (see Appendix B).

6.2 Understanding Sentence-Level Metrics

To further analyze the performance of the sentence-level metrics on paragraph-level data, we compare the two versions of applying a sentence-level metric to paragraphs discussed in §4.1. One version directly scores a full paragraph (thus, making no assumption about an alignment between the hypothesis and reference), whereas the other averages the scores of evaluating the individual k hypothesis sentences against the corresponding reference sentence (thus, assuming a sentence-level alignment exists).

Figure 5 shows that for two sentence-level metrics, the baseline trained in this work and BLEU, the performance of the two paragraph scoring variants is very similar. Then, Figure 6 shows that the Pearson correlation between the scores for those two variants is very high (≥ 0.85).

Together, these results point to the fact that there is little difference between these two methods. Directly scoring a paragraph or scoring individual sentences yield both similar scores and similar agreement to human ratings. The sentence-level metrics appear to be scoring full paragraphs in a desirable way—by calculating some average score across sentences.

This result is not obvious. As the length of the input increases, the bag-of-n-grams representation used by lexical matching metrics like BLEU have an increased potential for erroneous matches between the hypothesis and reference sentences, which could result in misleading scores. Learned metrics, like the ones trained in this work, have not been trained on a significant amount of very long data, so it is not clear that the scoring functions they learn would generalize well to longer inputs. Despite this, the sentence-level metrics appear to predict high-quality scores for paragraphs.

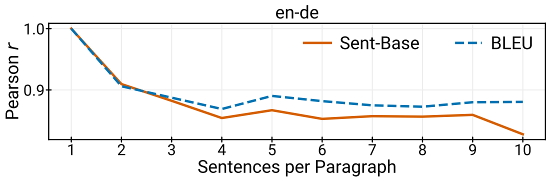

Figure 6: The plot shows the Pearson correlation on en-de between directly predicting a score for a paragraph of k sentences and calculating a paragraph score by averaging over k sentence-level scores. The correlations are quite high, demonstrating that the both methods result in very similar scores.

In Section 7, we propose a hypothesis for why this is the case and why training on paragraph-level data does not appear to result in a better metric.

6.3 Literary Translation Evaluation

We compared how frequently SENT-BASE and PARA-STRAT agree with the 540 pairwise human preference judgments between paragraph literary translations from Karpinska and Iyyer (2023). We found that the two models agreed 285 and 305 times, respectively. While it is a positive signal that the paragraph-level model appears to be better aligned with human preferences of longer texts, the difference was not quite statistically significant under a pairwise permutation test with $\alpha = 0.05$ ($p = 0.09$). Future work should perform a more in-depth analysis of this data and collect a larger number of paragraph translations and judgments.

7 Discussion

In theory, training on paragraph-level data should have advantages compared to training on sentence-level data. The metric (1) should be able to handle longer input sequences, (2) it should be able to capture long range dependencies, and (3) it should be able to model different paragraph-level phenomena like information or sentence reordering. However, we were not able to demonstrate these advantages in practice, and we theorize why as follows.

Source Context: Maria said no.
Source: She did not slap the green witch.

Reference Context: Maria dijo no.
Reference: No le dió una bofetada a la bruja verde.

Hypothesis: Ella(✓)/Él(✗) no le dió una bofetada a la bruja verde.

Figure 7: An English-to-Spanish translation example where the reference translation does not have enough information to correctly evaluate the hypothesis. Gender in Spanish is marked on pronouns, and Spanish is a pro-drop language, which means the pronoun can be omitted if the context is clear. In this example, the pronoun is dropped from the reference, so determining whether the pronoun used in the hypothesis requires taking into account the previous reference sentence. We suspect such examples are not frequent, and if they do exist, the information required to resolve the ambiguity is relatively local to the reference sentence.

First, the analysis in §6.2 shows that sentence-level metrics generalize well to significantly longer input, so advantage (1) may not be so relevant. We hypothesize that the scoring function learned by sentence-level metrics like SENT-BASE or COMET could score a token in the hypothesis based on some alignment to the reference using its relative position in the translation. This function would be agnostic with respect to the global positioning, and thus the scoring function would generalize well to longer inputs. If this were true, training on paragraph-level data would not be necessary to obtain good performance on long sequences.

Second, evaluating translation quality seems to be a very "local" problem in the sense that modeling long range dependencies is not frequently necessary for evaluation. Often, the reference phrase that aligns to a hypothesis phrase has enough information to accurately evaluate the hypothesis. If it does not, the information is likely nearby, not several sentences away (see Figure 7). Although the sentence-level metrics were not trained on multiple sentences, we suspect they are able to capture nearby dependencies across sentences when evaluating paragraphs. In theory, a paragraph-level metric would have the ability to model long range dependencies since it could observe them during training. However, if they are infrequent, advantage (2) over sentence-level metrics may be small.

Finally, the ability for our learned paragraph metrics to capture phenomena like sentence reordering is limited by our dataset construction method. Since the paragraphs in our training and test sets come from MT systems that translated one sentence at a time, there are no phenomena like sentence reordering present in the datasets. Therefore, the paragraph-level metric cannot learn to model such cases, and the metrics are never evaluated on them either. Thus, the limitations of the dataset mean that we cannot demonstrate advantage (3).

We believe that paragraph-level metrics are necessary for evaluating true paragraph translations, where MT systems can be more creative with how a full paragraph is translated, rather than paragraph translations that are created by translating individual sentences. We hypothesize that sentence-level metrics will not generalize well when there is no sentence alignment or there is significant information reordering. To accurately evaluate actual paragraph translations, metrics need to be trained on similar data. Future work should invest in collecting human ratings for paragraph-level translations so that new metrics can be trained and evaluated.

8 Related Work

The vast majority of research on MT evaluation has worked at the sentence level (Papineni et al., 2002; Banerjee and Lavie, 2005; Snover et al., 2006; Popović, 2015, 2017; Lo, 2019; Sellam et al., 2020; Rei et al., 2020, 2022; Thompson and Post, 2020; Wan et al., 2022), although there has been recent interest in moving beyond sentence-level evaluation. Vernikos et al. (2022) propose a method to incorporate document-level context into a sentence-level metric by using the additional context when computing the representations for the hypothesis and reference sentences. Although they use document context in their metric, it is still scores single sentences at a time, in contrast to the paragraph-level metrics in our work that predict a score for entire paragraphs at once. Then Jiang et al. (2022) propose a document-level metric called BlonDE that targets evaluating discourse phenomena as opposed to overall translation quality (i.e., they do not model translation accuracy errors). To the best of our knowledge, ours is the first study aimed at training a learned metric that directly scores entire paragraphs.

Other studies that have evaluated sentence-level metrics beyond the sentence-level have done so in the literary domain. Thai et al. (2022) show that automatic metrics prefer MT output over human translations, and Karpinska and Iyyer (2023) show that metrics prefer actual translations of paragraphs over sentence-by-sentence translations. Our work is complementary to theirs as we focus on the news domain, train metrics on paragraph-level data, and evaluate on a much larger set of human ratings. It is not clear whether conclusions reached about metrics in the news domain will apply to the literary domain or vice versa.

Some researchers have developed challenge sets that can be used to probe how well metrics capture discourse phenomena that appear when translating more than one sentence at a time (Bawden et al., 2018; Müller et al., 2018; Lopes et al., 2020). However, these challenge sets can be trivial for reference-based metrics because the reference often resolves the ambiguity in the translation. To the best of our knowledge, a challenge set that forces reference-based metrics to use context outside of a single reference sentence during evaluation (see Figure 7) does not exist.

Research on generating translations of text longer than single sentences directly use sentence-level metrics to score translations (Tiedemann and Scherrer, 2017; Miculicich et al., 2018; Ma et al., 2020; Wu et al., 2023; Post and Junczys-Dowmunt, 2023). Our work can be viewed as a justification for doing so.

9 Conclusion

In this work, we proposed a method for constructing paragraph-level datasets for training and meta-evaluating MT evaluation metrics from sentence-level data. Our experimental results showed that metrics trained on paragraph-level data do not necessarily out-perform those trained on sentence-level data, potentially due to the fact that sentence-level metrics seem to generalize well to longer inputs and limitations of our paragraph-level datasets. Future work should invest in collecting human judgments for paragraph translations generated by MT systems that directly translate full paragraphs instead of translating one sentence at a time. Such a dataset would be more likely to contain phenomena that do not exist at the sentence level, which we hypothesize would be more likely to require metrics designed to work at the paragraph level.

Limitations

There are a couple of limitations related to our dataset construction approach that are worth enumerating.

As discussed in Section 7, our ability to evaluate metrics' performances on all types of paragraph-level translations is limited by our dataset construction method. Our translated paragraphs are generated by MT systems that translate one sentence at time, which results in sentence aligned data. Therefore, we are unable to evaluate metrics on true paragraph-level translations that might have sentence or information reordering.

Then, the WMT data no longer contains information about the white space between the original source sentences. Therefore, the DA and MQM paragraph-level datasets do not contain the paragraph breaks that were in the original document. Each of the k sentences is concatenated together and separated by a space in our work, so it is very likely that the artificially constructed paragraphs do not perfectly resemble real paragraphs.

References

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2015. TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org.

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.

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, Ann Arbor, Michigan. Association for Computational Linguistics.

Loïc Barrault, Magdalena Biesialska, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Matthias Huck, Eric Joanis, Tom Kocmi, Philipp Koehn, Chi-kiu Lo, Nikola Ljubešić, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Santanu Pal, Matt Post, and Marcos Zampieri. 2020. Findings of the 2020 Conference on Machine Translation (WMT20). In *Proceedings of the Fifth Conference on Machine Translation*, pages 1–55, Online. Association for Computational Linguistics.

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.

Daniel Deutsch, George Foster, and Markus Freitag. 2023. Ties Matter: Meta-Evaluating Modern Metrics with Pairwise Accuracy and Tie Calibration.

Patrick Fernandes, Daniel Deutsch, Mara Finkelstein, Parker Riley, André F. T. Martins, Graham Neubig, Ankush Garg, Jonathan H. Clark, Markus Freitag, and Orhan Firat. 2023. The Devil is in the Errors: Leveraging Large Language Models for Fine-grained Machine Translation Evaluation.

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, 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.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, George Foster, Alon Lavie, and Ondřej Bojar. 2021b. Results of the WMT21 Metrics Shared Task: Evaluating Metrics with Expert-based Human Evaluations on TED and News Domain. In *Proceedings of the Sixth Conference on Machine Translation*, pages 733–774, Online. Association for Computational Linguistics.

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.

Yuchen Eleanor Jiang, Tianyu Liu, Shuming Ma, Dongdong Zhang, Mrinmaya Sachan, and Ryan Cotterell. 2023. Discourse-Centric Evaluation of Document-level Machine Translation with a New Densely Annotated Parallel Corpus of Novels. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7853–7872, Toronto, Canada. Association for Computational Linguistics.

Marzena Karpinska and Mohit Iyyer. 2023. Large language models effectively leverage document-level context for literary translation, but critical errors persist.

Tom Kocmi and Christian Federmann. 2023. Large Language Models Are State-of-the-Art Evaluators of Translation Quality.

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.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing.

In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

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.

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.

Qingsong Ma, Johnny Wei, Ondřej Bojar, and Yvette Graham. 2019. Results of the WMT19 Metrics Shared Task: Segment-Level and Strong MT Systems Pose Big Challenges. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 62–90, Florence, Italy. Association for Computational Linguistics.

Shuming Ma, Dongdong Zhang, and Ming Zhou. 2020. A Simple and Effective Unified Encoder for Document-Level Machine Translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3505–3511, Online. Association for Computational Linguistics.

Nitika Mathur, Johnny Wei, Markus Freitag, Qingsong Ma, and Ondřej Bojar. 2020. Results of the WMT20 Metrics Shared Task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 688–725, Online. Association for Computational Linguistics.

Lesly Miculicich, Dhananjay Ram, Nikolaos Pappas, and James Henderson. 2018. Document-Level Neural Machine Translation with Hierarchical Attention Networks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2947–2954, Brussels, Belgium. Association for Computational Linguistics.

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.

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, Lisbon, Portugal. Association for Computational Linguistics.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. 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.

Matt Post and Marcin Junczys-Dowmunt. 2023. Escaping the sentence-level paradigm in machine translation.

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. 2020. 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.

Adam Roberts, Hyung Won Chung, Anselm Levskaya, Gaurav Mishra, James Bradbury, Daniel Andor, Sharan Narang, Brian Lester, Colin Gaffney, Afroz Mohiuddin, Curtis Hawthorne, Aitor Lewkowycz, Alex Salcianu, Marc van Zee, Jacob Austin, Sebastian Goodman, Livio Baldini Soares, Haitang Hu, Sasha Tsvyashchenko, Aakanksha Chowdhery, Jasmijn Bastings, Jannis Bulian, Xavier Garcia, Jianmo Ni, Andrew Chen, Kathleen Kenealy, Jonathan H. Clark, Stephan Lee, Dan Garrette, James Lee-Thorp, Colin Raffel, Noam Shazeer, Marvin Ritter, Maarten Bosma, Alexandre Passos, Jeremy Maitin-Shepard, Noah Fiedel, Mark Omernick, Brennan Saeta, Ryan Sepassi, Alexander Spiridonov, Joshua Newlan, and Andrea Gesmundo. 2022. Scaling Up Models and Data with `t5x` and `seqio`. *arXiv preprint arXiv:2203.17189*.

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.

Noam Shazeer and Mitchell Stern. 2018. Adafactor: Adaptive learning rates with sublinear memory cost. In *International Conference on Machine Learning*, pages 4596–4604. PMLR.

Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. A Study of Translation Edit Rate with Targeted Human Annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas.

Zewei Sun, Mingxuan Wang, Hao Zhou, Chengqi Zhao, Shujian Huang, Jiajun Chen, and Lei Li. 2022. Rethinking Document-level Neural Machine Translation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3537–3548, Dublin, Ireland. Association for Computational Linguistics.

Katherine Thai, Marzena Karpinska, Kalpesh Krishna, Bill Ray, Moira Inghilleri, John Wieting, and Mohit Iyyer. 2022. Exploring Document-Level Literary Machine Translation with Parallel Paragraphs from World Literature. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9882–9902, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Brian Thompson and Matt Post. 2020. Automatic Machine Translation Evaluation in Many Languages via Zero-Shot Paraphrasing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 90–121, Online. Association for Computational Linguistics.

Jörg Tiedemann and Yves Scherrer. 2017. Neural Machine Translation with Extended Context. In *Proceedings of the Third Workshop on Discourse in Machine Translation*, pages 82–92, Copenhagen, Denmark. Association for Computational Linguistics.

Zhaopeng Tu, Yang Liu, Shuming Shi, and Tong Zhang. 2018. Learning to Remember Translation History with a Continuous Cache. *Transactions of the Association for Computational Linguistics*, 6:407–420.

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.

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.

Minghao Wu, George Foster, Lizhen Qu, and Gholamreza Haffari. 2023. Document Flattening: Beyond Concatenating Context for Document-Level Neural Machine Translation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 448–462, Dubrovnik, Croatia. 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.

A Dataset Statistics

The exact number of paragraph-level instances by WMT year and language pair that we generaetd from our dataset construction procedure (see §3) can be found in Table 2 for DA and Table 3 for MQM. Figure 8 visualizes the distribution of the lengths of the hypotheses in the paragraph-level datasets based on mT5 SPM tokens. Then, Table 4 contains the number of paragraph examples that are too long to fit into the 1024 SPM maximum context length that is used by the metrics trained in this work.

B Additional Results

Figure 9 contains the system- and segment-level accuracy correlations on the en-de and en-ru language pairs from WMT'22 MQM that were not presented in the main body of the paper. Figure 10 contains the correlations for all 3 language pairs but uses Pearson correlation instead of pairwise accuracy.

Figure 11 shows the correlation between the two ways to apply a segment-level metric to paragraph-level data, directly scoring the paragraph or averaging the k segment scores, on the en-ru and zh-en WMT'22 MQM dataset.

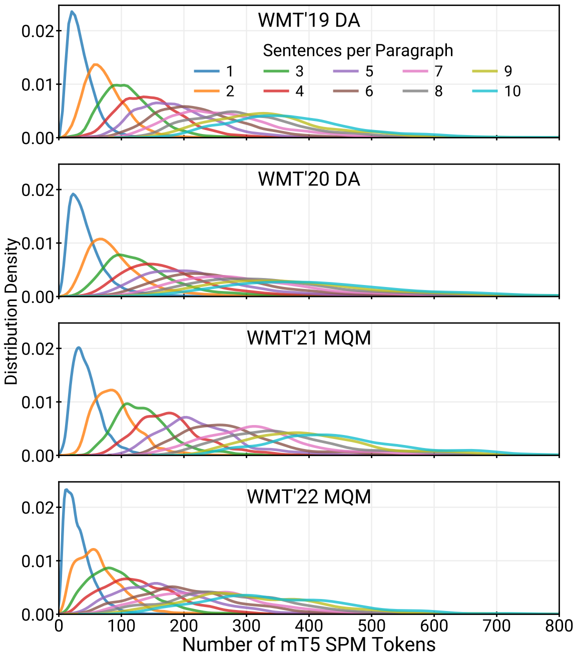

Figure 8: The distribution of the length of the hypothesis translations for the direct assessment (DA) and MQM datasets for a given number of sentences per paragraph.

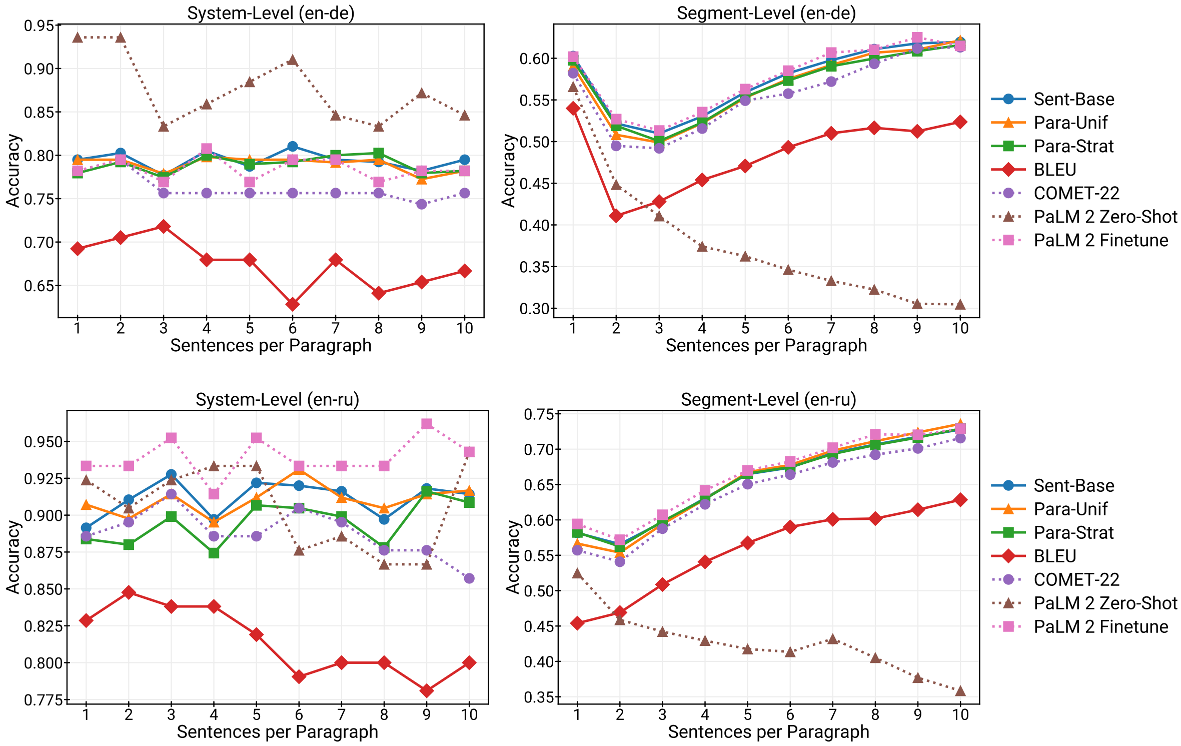

Figure 9: System- and segment-level accuracy results for the en-de and en-ru language pairs on the paragraph-level WMT'22 MQM data for different numbers of k sentences per paragraph. In general, the system-level correlations are relatively flat and the segment-level correlations increase as the number of sentences per paragraph increases. BlonDE is not included because it only supports English.

Year	**LP**	**Sentences per Paragraph**									
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
2019	de-cs	16900	1032	95	12	1	0	0	0	0	0
2019	de-en	34756	16754	10896	7735	5976	4660	3947	3147	2730	2345
2019	de-fr	6700	173	5	0	0	0	0	0	0	0
2019	en-cs	27445	13241	8710	6152	4834	3865	3215	2607	2371	1967
2019	en-de	45131	21777	14311	10124	7932	6363	5274	4285	3906	3232
2019	en-fi	20618	9937	6557	4611	3628	2910	2419	1945	1799	1482
2019	en-gu	10151	4890	3229	2267	1774	1423	1221	964	884	722
2019	en-kk	12922	6221	4115	2888	2253	1813	1562	1223	1112	910
2019	en-lt	13217	6363	4219	2963	2319	1863	1603	1257	1137	944
2019	en-ru	22600	10902	7180	5069	3974	3185	2650	2137	1966	1633
2019	en-zh	26530	12810	8434	5944	4673	3758	3102	2520	2308	1904
2019	fi-en	20286	362	21	2	0	0	0	0	0	0
2019	fr-de	4000	87	3	0	0	0	0	0	0	0
2019	gu-en	14860	550	40	2	0	0	0	0	0	0
2019	kk-en	15763	705	77	10	0	0	0	0	0	0
2019	lt-en	16046	489	32	2	0	0	0	0	0	0
2019	ru-en	24247	785	83	10	1	0	0	0	0	0
2019	zh-en	50722	15164	9347	6774	5030	4087	3312	2714	2226	1797
2020	cs-en	9381	4322	2628	1797	1323	940	685	404	241	138
2020	de-en	12541	5825	3451	2422	1808	1220	927	652	507	378
2020	en-cs	34180	16371	10324	7358	5591	4501	3474	2749	2270	2035
2020	en-de	17393	8337	5253	3723	2859	2283	1729	1362	1138	1033
2020	en-iu	6145	3028	1990	1479	1152	937	801	693	600	538
2020	en-ja	21999	10672	6769	5036	3907	3093	2513	2109	1812	1635
2020	en-pl	18342	8891	5636	4192	3266	2569	2089	1756	1514	1377
2020	en-ru	19543	9494	6058	4433	3477	2750	2279	1847	1602	1468
2020	en-ta	9175	4439	2825	2100	1634	1301	1035	875	746	680
2020	en-zh	41965	20069	12656	9034	6843	5510	4260	3371	2782	2483
2020	iu-en	12172	75	0	0	0	0	0	0	0	0
2020	ja-en	9879	4710	3047	2103	1715	1321	1053	845	759	639
2020	km-en	6951	72	0	0	0	0	0	0	0	0
2020	pl-en	12435	6048	3871	2857	2184	1708	1445	1265	1030	844
2020	ps-en	7138	110	2	0	0	0	0	0	0	0
2020	ru-en	11244	5369	3408	2405	1832	1488	1179	952	785	604
2020	ta-en	7842	3762	2406	1723	1322	1065	847	694	572	473
2020	zh-en	30325	14567	9253	6674	5106	4078	3374	2811	2223	1824

Table 2: The number of paragraphs with the given number of sentences per paragraph from the direct assessment data from WMT'19 and WMT'20. Each paragraph is required to a contiguous block of sentences that are rated by the same rater.

Dataset	LP	Sentences per Paragraph									
		1	2	3	4	5	6	7	8	9	10
WMT’21	en-de	7905	3825	2460	1800	1395	1140	870	765	660	585
WMT’21	en-ru	7905	3825	2460	1800	1395	1140	870	765	660	585
WMT’21	zh-en	9058	4340	2814	1974	1596	1190	994	770	658	644
WMT’22	en-de	18410	8932	5236	3486	3080	1610	1568	1470	1372	1330
WMT’22	en-ru	19725	9570	5610	3735	3300	1725	1680	1575	1470	1425
WMT’22	zh-en	28110	13005	7935	5655	4245	3285	2670	2160	1935	1710

Table 3: The number of paragraphs with the given number of sentences per paragraph from the MQM data from WMT’21 and WMT’22. Each paragraph is required to a contiguous block of sentences that are rated by the same rater.

Dataset	Sentences per Paragraph									
	1	2	3	4	5	6	7	8	9	10
WMT’19 DA	2 (0%)	3 (0%)	4 (0%)	15 (0%)	48 (0%)	196 (1%)	440 (2%)	702 (3%)	1349 (7%)	1944 (11%)
WMT’20 DA	4 (0%)	179 (0%)	667 (1%)	1148 (2%)	1598 (4%)	2222 (6%)	2879 (10%)	3389 (15%)	4041 (22%)	4688 (29%)
WMT’21 MQM	0 (0%)	0 (0%)	0 (0%)	0 (0%)	3 (0%)	23 (1%)	103 (4%)	245 (11%)	295 (15%)	488 (27%)
WMT’22 MQM	0 (0%)	0 (0%)	6 (0%)	11 (0%)	56 (1%)	74 (1%)	110 (2%)	202 (4%)	266 (6%)	450 (10%)

Table 4: The number (and percent) of paragraphs for which the number of SPM tokens in the reference and hypothesis combined is larger than the maximum allowable input length by our metric, 1024. If the input is too long, it is truncated.

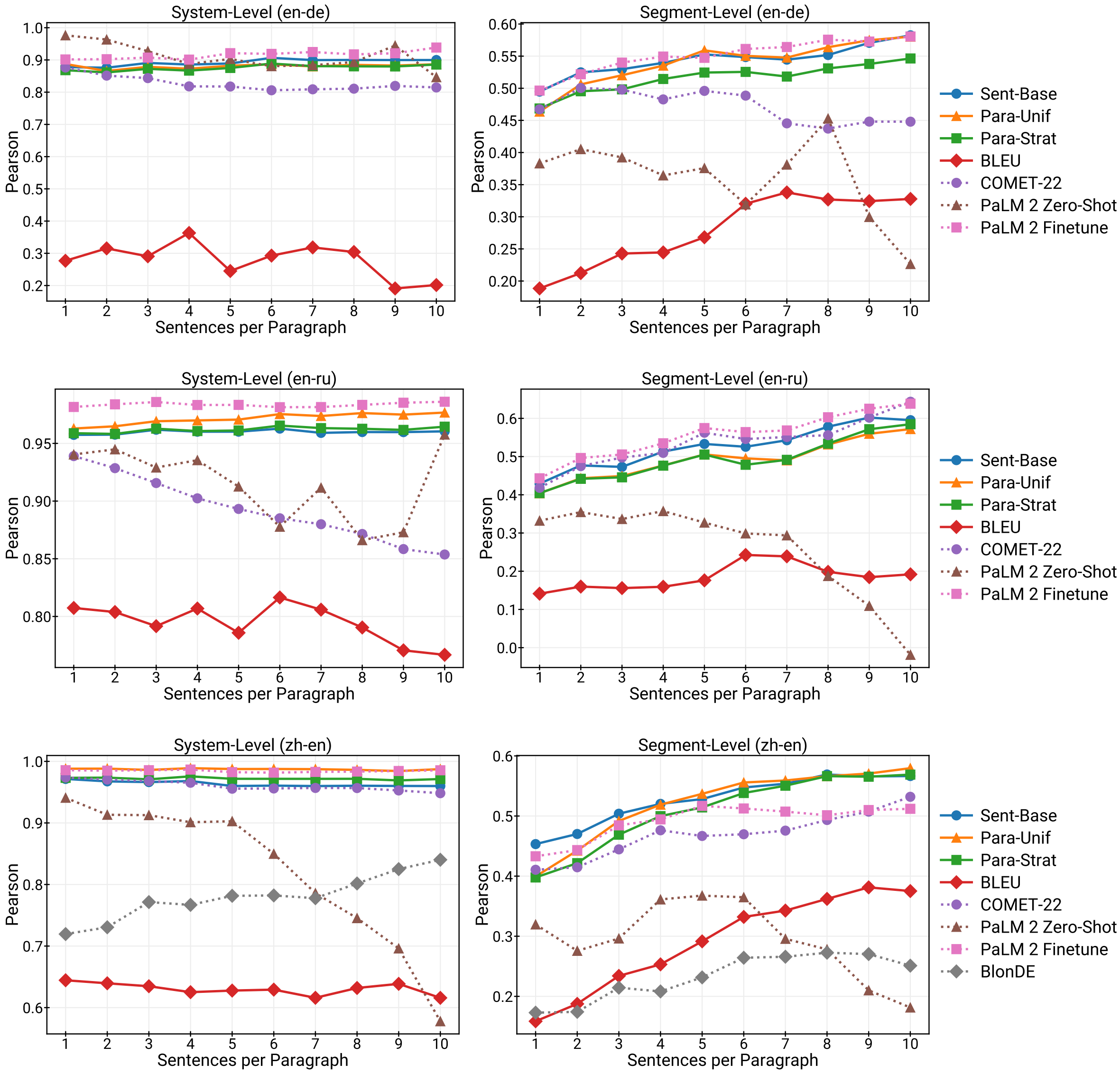

Figure 10: The system- and segment-level correlation results when using Pearson correlation follow very similar trends to those that use pairwise accuracy. The segment-level Pearson uses the “no grouping” variant from Deutsch et al. (2023) to avoid the NaN problem that happens with the “group-by-item” variant, which was used in combination with pairwise accuracy in the main body of the paper.

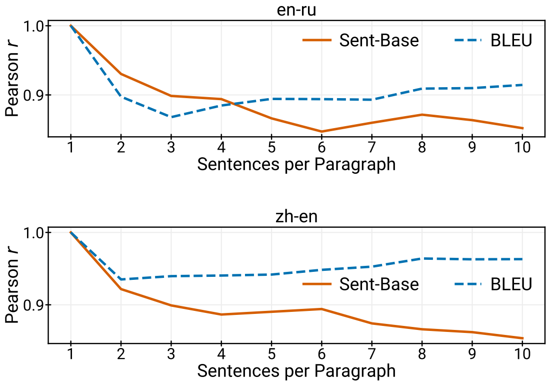

Figure 11: The correlation between metric scores for directly scoring paragraphs and averaging the score of evaluating the k sentences per paragraph independently on the WMT'22 MQM data.