Embarrassingly Easy Document-Level MT Metrics: How to Convert Any Pretrained Metric Into a Document-Level Metric

Giorgos Vernikos[*] EPFL + HEIG-VD
Brian Thompson AWS AI Labs
Prashant Mathur AWS AI Labs
Marcello Federico AWS AI Labs
georgios.vernikos@epfl.ch, {brianjt, pramathu, marcfede}@amazon.com

Abstract

We present a very simple method for extending pretrained machine translation metrics to incorporate document-level context. We apply our method to four popular metrics: BERTScore, Prism, COMET, and the reference-free metric COMET-QE. We evaluate our document-level metrics on the MQM annotations from the WMT 2021 metrics shared task and find that the document-level metrics outperform their sentence-level counterparts in about 85% of the tested conditions, when excluding results on low-quality human references. Additionally, we show that our document-level extension of COMET-QE dramatically improves accuracy on discourse phenomena tasks, supporting our hypothesis that our document-level metrics are resolving ambiguities in the reference sentence by using additional context.

1 Introduction

Automatic evaluation is crucial to the machine translation (MT) community for tracking progress, evaluating new ideas and making modeling choices. While human evaluation is the gold standard for MT evaluation, it is very expensive, and thus most research groups must rely on automatic metrics. Current State-of-the-art (SOTA) metrics are *pretrained* (Kocmi et al., 2021; Freitag et al., 2021b), leveraging existing language models (LMs) or sequence-to-sequence models to judge how well a hypothesis (i.e. MT system output) conveys the same meaning as a human reference translation.

Sentences are often ambiguous, and many recent works have demonstrated that incorporating inter-sentential (i.e. document-level) context is beneficial in both MT (Lopes et al., 2020; Fernandes et al., 2021) and human evaluation of MT (Läubli et al., 2018; Toral, 2020; Freitag et al., 2021a).

A human reference translation is (at least ideally) created taking the entire source document into account. However, just as source sentences are often ambiguous, we hypothesize that human reference sentences also contain ambiguities. Thus, when a system output deviates from the human reference, we may need to look at additional context to determine if those deviations are acceptable, in the context of the full document translation.

In this study, we present a simple procedure for extending pretrained MT metrics to the document level. Prior work has used pretrained models models like BERT (Devlin et al., 2019) to embed a single human reference sentence and hypothesis (e.g. an MT output) sentence. We instead argue that a *better* representation of the reference or hypothesis sentence can be obtained by providing several sentences of context to the pretrained model, allowing the pretrained model to *use surrounding context when embedding each sentence of interest*. Once the embeddings of the reference or hypothesis sentence have been computed (taking into account surrounding sentence context), the metric is computed in the same manner as the sentence-level metric.[1,2]

We apply this method to extend four popular pretrained metrics to the document level:[3]

- BERTScore (Zhang et al., 2020), a text generation metric that uses the alignments from token embeddings of a pretrained BERT model to score the similarity of a hypothesis and reference.
- Prism (Thompson and Post, 2020a), a text generation metric which utilizes a sequence-to-sequence paraphrase model to score how well a hypothesis paraphrases the reference.
- COMET (Rei et al., 2020), an MT metric which fine-tunes a multilingual LM, namely

[*] Work conducted during an internship at Amazon.

[1] In the case of Prism (Thompson and Post, 2020a), we modify this logic slightly to retain only the probabilities of the sentence of interest (see § 3.2).

[2] In the case of COMET/COMET-QE (Rei et al., 2020), which incorporates the source sentence, we provide additional source context in the same manner (see § 3.3 and § 3.4).

[3] We release our code at `https://github.com/amazon-research/doc-mt-metrics`.

XLM-R (Conneau et al., 2020), to predict translation quality given a hypothesis, source, and reference.

- COMET-QE (Rei et al., 2020), the reference-free (i.e. “quality estimation as a metric”) version of COMET.

To test the effectiveness of our document-level metrics, we measure system-level correlation with human judgments. We select the so-called "platinum" Multidimensional Quality Metrics (MQM) judgments collected for the WMT 2021 metrics task (Freitag et al., 2021b). We believe MQM judgments are the best available to test document-level MT metrics as these judgments are made by expert translators that have access to—and are strongly advised to consider—source-side document-level context when judging each target sentence. We perform evaluation on all the WMT 2021 language pairs (En→De, Zh→En, En→Ru) and domains (TED talks and news) for which MQM judgments are available.

We find that our document-level extensions of these four metrics outperform their sentence-level counterparts in 75% of cases considered. Excluding Zh→En news, where the human reference is of low quality (see § 4.1), we see improvements in 85% of cases. This provides strong evidence that document-level context is useful in the automatic evaluation of MT.

We also conduct analysis to better understand the performance improvement that we observe. We demonstrate that our document-level extension of COMET-QE significantly improves over its sentence-level counterpart on targeted tasks evaluating discourse phenomena, namely pronoun resolution and Word Sense Disambiguation (WSD).[4] This finding provides further evidence that our document-level metrics are using context to resolve ambiguities in the reference sentence. We also show that using reference context is better than using context from the MT output, likely because the MT output contains more errors than the reference.

In summary, our contributions are:

1. We present a simple but effective method to extend pretrained sentence-level metris to the document level, and apply it to four popular metrics.
2. We show that the proposed document-level metrics tend to have better correlation with human judgments than their sentence-level counterparts.
3. We improve over both COMET and COMET-QE, which appear to be the previous SOTA automatic metric and reference-free metric, respectively (Freitag et al., 2021b; Kocmi et al., 2021).
4. We conduct analysis to show that the improvements observed using our approach can be attributed to better context utilization, and also show that using reference context is better than using context from the hypothesis.

2 Related Work

Our work has parallels in human MT evaluation, where document-level judgments are required to distinguish human translation quality from MT system quality (Läubli et al., 2018; Toral, 2020). Castilho et al. (2020) showed that many source sentences are ambiguous, but that ambiguities are often resolved using only a few additional sentences of context. This suggests that we do not need to incorporate very many additional sentences of context into a document-level metric in order to see an improvement in quality.

Pretrained metrics are metrics which leverage large existing pretrained LMs or sequence-to-sequence models, and include YiSi (Lo, 2019), COMET (Rei et al., 2020), BERTscore (Zhang et al., 2020), Prism (Thompson and Post, 2020a), BLEURT (Sellam et al., 2020), and others. Pretrained metrics have been shown to consistently outperform surface-level metrics such as BLEU (Papineni et al., 2002), TER (Snover et al., 2006), and chrF (Popović, 2015) – see Mathur et al. (2020); Kocmi et al. (2021); Freitag et al. (2021b).

Prior to the rise of pretrained metrics, several works targeted discourse-level phenomena in MT metrics such as pronominal anaphora (Hardmeier and Federico, 2010; Miculicich Werlen and Popescu-Belis, 2017; Jwalapuram et al., 2019) and lexical cohesion (Wong and Kit, 2012; Gong et al., 2015). For a detailed overview of evaluation of discourse-level phenomena, we direct the reader to Maruf et al. (2021). Recently, Jiang et al. (2022) proposed BlonDe, a document-level metric that focuses on discourse phenomena in order to score a translated document. However, we find that BlonDe substantially under-performs modern pretrained metrics, despite taking advantage of document-level context (see § 5.1).

[4] The use of a reference would make these tasks trivial, so we limit our analysis to the reference-free COMET-QE.

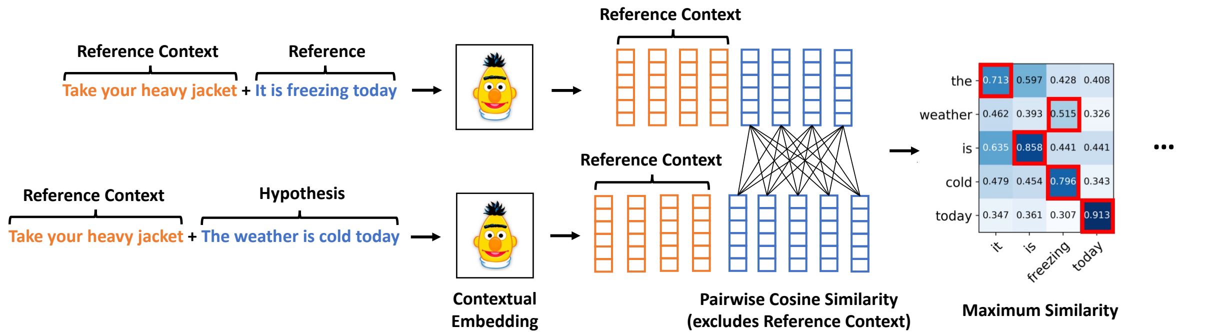
Reference Context
Reference
Take your heavy jacket + It is freezing today
Reference Context
Hypothesis
Take your heavy jacket + The weather is cold today
Contextual Embedding
Reference Context
Reference Context
Pairwise Cosine Similarity (excludes Reference Context)
the 0.713 0.597 0.428 0.408
weather 0.462 0.393 0.515 0.326
is 0.635 0.858 0.441 0.441
cold 0.479 0.454 0.796 0.343
today 0.347 0.361 0.307 0.913
it
is
freezing
today
...
Maximum Similarity

Figure 1: To extend BERTScore to the document level, we add reference context (e.g. "Take your heavy jacket") to both the reference sentence (e.g. "It is freezing today") and hypothesis sentence (e.g. "The weather is cold today"). This context is used to improve the embeddings of the reference and hypothesis sentences (e.g. helping the model understand that "it" is likely referring to weather). However, the additional context is not used when performing alignment and scoring, which follows standard sentence-level BERTScore. The same methodology is applied to Prism and COMET/COMET-QE (not shown). Image adapted from Zhang et al. (2020).

3 Method

At a high level, we propose a very simple procedure for extending pretrained MT metrics to the document level: As in standard sentence-level metrics, we produce a score for a single hypothesis sentence compared to a single human reference translation sentence. However, we use additional context[5,6] from the reference translation when computing the contextual embeddings for both the hypothesis sentence and reference sentence. Once the hypothesis and reference sentence have been embedded, we discard the extra context sentences before computing metric scores following the same process as the corresponding sentence-level metric. Additional details are provided for each metric below.

For the following discussion, let s refer to the source sentence, h refer to the hypothesis (i.e. MT system output) sentence, r refer to the human reference translation sentence, and let c_s, c_h and c_r refer to the source, hypothesis, and reference context, respectively.

3.1 Document-level BERTScore

BERTScore (Zhang et al., 2020) is an unsupervised text generation metric that leverages the power of a pretrained large LM to score generated text. BERTScore encodes tokens of both the reference and the hypothesis with a pretrained LM and computes soft alignments based on token similarities. The alignment matrix is then used to calculate the precision, recall and F1 scores of the hypothesis compared to the reference.

To extend BERTScore to the document level, we use the reference context $\langle c_r \rangle$ while encoding the hypothesis or the reference with the LM. However, we align only the tokens of the reference/hypothesis sentence being scored (see Figure 1 for an illustration).

For BERTScore we use the default LM option for each language pair, which is the multilingual BERT-base (Devlin et al., 2019) for all En→* pairs and RoBERTa-large (Liu et al., 2019) for *→En pairs. BERT and RoBERTa are naively document-level; specifically, the LMs are trained on up to 512 tokens at a time, which is significantly longer than the average sentence length. Thus no changes to the underlying model were required to extend BERTscore to the document level.

3.2 Document-level Prism

Prism (Thompson and Post, 2020a,b) is an unsupervised text generation metric that uses a sequence-to-sequence paraphraser to evaluate how well a hypothesis paraphrases a human reference translation. Specifically, to score a translation the reference is fed to the encoder and the hypothesis is force-decoded in the decoder via teacher forcing. The token-level probabilities of the reference are aggregated to produce a score and the process is repeated with the hypothesis in the encoder side and the reference in the decoder. The final score is the average of the two scores.

In order to generalize Prism for document-level

[5] We use two preceding sentences from the reference as context, but our method could be applied to additional previous and/or subsequent sentences.

[6] We only use valid context. For example, when using a nominal value of two prior sentences as context, the first sentence in a document gets no context sentences and the second sentence gets one context sentence.

evaluation we concatenate the reference context c_r to both the reference and hypothesis $\langle c_r; r, c_r; h\rangle$. The context is used as a prompt; that is, we only aggregate token-level probabilities for the sentence being evaluated. The authors of Prism release the sentence-level multilingual MT model that they zero-shot paraphrase model. However, we require a document-level model to extend Prism to the document level. One option for extending Prism to the document level is to train a document-level, multilingual MT model. While document-level data collection methods and datasets do exist (Guo et al., 2019; Thompson and Koehn, 2020; Cettolo et al., 2012; Lison et al., 2018), document-level data is not currently available in nearly as many language pairs as sentence-level data. To extend Prism to the document level, we instead use mBART-50 (Tang et al., 2020), a multilingual encoder-decoder LM. mBART-50 is trained on document fragments of up to 512 tokens, in 50 languages, resulting in a multilingual document-level paraphraser. Note that while an mBART model fine-tuned on (sentence-level) translations is available, we do not use it because we require a document-level model. As a result, although the mBART model we use is multilingual, it is not a translation model so we cannot use it for the reference-free version of Prism.

3.3 Document-level COMET

COMET (Rei et al., 2020) is a supervised metric that is trained on human judgments. COMET encodes the source, hypothesis and reference via a multilingual pretrained LM and the representation of each sentence is the average of its output token embeddings. The encoded representations are further combined via subtraction and multiplication and fed to a regressor that predicts a score for each translated sentence. We use COMET-MQM_2021 (Rei et al., 2021), which is built on top of XLM-RoBERTa-large (Conneau et al., 2020). The COMET models are pretrained on direct assessment judgements from WMT 2015 to WMT 2020 and fine-tuned on MQM z-scores from Freitag et al. (2021a).

To extend COMET to the document level, we integrate source context c_s and reference context c_r by concatenating them with the source and hypothesis/reference in the encoder. We obtain sentence representations by averaging the output embeddings of the tokens of the current sentence only before passing them to the regressor.

As with BERTscore, the model underlying COMET is inherently document-level. However, the underlying LM is fine-tuned for a few epochs on human judgments from previous WMT campaigns that consist of a single (source, reference, and hypothesis) sentence and the corresponding score. As the amount of fine-tuning is quite limited, we hypothesize that the model has still retained its ability to handle text beyond sentence level, and this assumption appears to be confirmed by experimental results (see § 5.1).

3.4 Document-level COMET-QE

COMET-QE (Rei et al., 2021) is the reference-free version of COMET. We use the latest COMET-MQM-QE_2021, trained similarly to the COMET-MQM_2021 discussed above. Although COMET-QE does not does not have access to the reference it has been shown to perform reasonably well compared to strong reference-based metrics (Kocmi et al., 2021).

Similar to reference-based COMET, to extend COMET-QE to the document level, for each source s and hypothesis h, we concatenate the previous source and hypothesis sentences as context $\langle c_s; s, c_h; h\rangle$ and score the hypothesis h in question.

The pretrained model for COMET-QE is the same as the one used in COMET, therefore no further modifications are required to extend COMET to the document level.

4 Experiments

Motivated by the finding of Scherrer et al. (2019); Kim et al. (2019); Castilho et al. (2020) that two previous sentences are sufficient context to correctly resolve ambiguities in the majority of sentences, we use two previous reference sentences as context unless otherwise noted. Sentences are separated using the separator token of each model: [SEP] for RoBERTa and <\s> for XLM-R and mBART-50. We use reference context c_r as reference for the hypothesis, as opposed to hypothesis context c_h. This is done in order to avoid propagation of translation errors (see § 6.1 for an ablation using hypothesis context instead of reference context).

4.1 Human Judgment Experiments

We compare our document-level metrics judgments of MT outputs with those of the human-generated

Model	Input	TED talks			News		
		En→De	En→Ru	Zh→En	En→De	En→Ru	Zh→En
BlonDe	$\langle c_h, h, c_r, r \rangle$	-	-	-0.232	-	-	0.212
Prism (m39v1)	$\langle h, r \rangle$	0.656	0.867	0.272	0.841	0.799	0.558
Prism (mBART-50)	$\langle h, r \rangle$	0.486	0.845	0.240	0.661	0.710	0.363
Doc-Prism (mBART-50)	$\langle c_r; h, c_r; r \rangle$	**0.692**	**0.852**	**0.372**	**0.825***	**0.777**	**0.374**
BERTScore	$\langle h, r \rangle$	0.506	0.831	0.293	0.930	**0.629**	**0.575***
Doc-BERTScore	$\langle c_r; h, c_r; r \rangle$	**0.613***	**0.836**	**0.344***	**0.948***	0.622	0.535
COMET	$\langle s, h, r \rangle$	**0.818**	0.841	0.266	0.772	0.659	**0.628**
Doc-COMET	$\langle c_s; s, c_r; h, c_r; r \rangle$	0.816	**0.849**	**0.297**	**0.802***	**0.676**	0.513
COMET-QE	$\langle s, h \rangle$	0.694	0.818	**-0.209**	0.711	0.688	**0.529**
Doc-COMET-QE	$\langle c_s; s, c_h; h \rangle$	**0.724**	**0.830**	-0.255	**0.733**	**0.733***	0.462

Table 1: System-level correlation with WMT 2021 MQM annotations for Prism, BERTScore, COMET and COMET-QE and their generalization for document-level evaluation (Doc-*, this work). Within each document/sentence-level pair, **bold** denotes the best correlation and "*" denotes a statistically significant ($p < 0.05$) difference. Excluding Zh→En news data, which has a very low-quality human reference (see § 4.1), our document-level metrics outperform their sentence-level counterparts in 17 of 20 (85%) of cases, and 6 of 6 (100%) of statistically significantly different cases.

MQM annotations from the 2021 WMT metrics shared task (Freitag et al., 2021a). We select MQM for several reasons: They are produced by professional translators (compared to crowd workers or translation researchers) and require explicit error annotations that are believed to lead to higher quality annotations. Also, MQM annotators are specifically instructed to "*identify all errors within each segment in a document, paying particular attention to document context*." In 2021, in addition to the news domain, annotations were also produced for translations of TED talks in three language pairs: En→De, Zh→En and En→Ru.

One potential problem with the metrics dataset is the quality of the Zh→En news human reference. The WMT metrics shared task organizers acquired MQM scores for the human references, in addition to MT system outputs. The Zh→En reference received an MQM score of just 4.27, only slightly better than the best MT system at 4.42 (Freitag et al., 2021b). For reference, 0.0 is a perfect score and a score of 5.0 corresponds to one major error (or many minor errors) per sentence. In contrast, for the same language pair, the TED reference has an MQM score of 0.42 vs the best MT system at 1.65.

4.2 Discourse Phenomena Experiments

In order to confirm that any gains we see from document-level metrics are in fact due to their ability to correctly handle ambiguities in the reference which can be resolved using document-level context, we also perform targeted evaluation of discourse phenomena using contrastive sets. These testsets are common in the evaluation of document-level MT systems where a context-aware model should ideally assign the highest probability to the correct translation; all translations are plausible and only the use of context can reveal the correct translations. For our case, since we are evaluating MT metrics, we treat each sentence as a different hypothesis and calculate how often our metric ranks the correct translation the highest. Since the use of a reference would make this task trivial for reference-based metrics, we only evaluate on COMET-QE. We use ContraPro (Müller et al., 2018), a selection of sentences from OpenSubtitles2018 (Lison et al., 2018) that contain the English anaphoric pronoun *it* in the source side. Starting from the correct translation in German, contrastive translations are automatically created to contain the German pronouns *er*, *sie* and *es*. In order to identify the correct translation the model must look into previous context. We also evaluate on a similar dataset for En→Fr created by Lopes et al. (2020) for the translation of *it* and *they* into *il*, *elle*, *ils*, *elles* in French. Finally, we evaluate on DiscEvalMT (Bawden et al., 2018), a contrastive test which consists of 200 examples of anaphoric pronoun translation for En→Fr and 200 examples of WSD.

4.3 Baseline Methods

For correlation with human MT quality judgments, in addition to the sentence-level version of each metric we extend, we also compare to

Model	En→De			En→Fr				
	Intra	Inter	Total	Intra	Inter	Total	Anaphora	WSD
Lopes et al. (2020)	-	-	70.8	-	-	83.2	82.5	55.0
COMET-QE	78.2	40.9	48.4	76.3	76.6	76.5	50.0	50.0
Doc-COMET-QE (this work)	**80.5**	**72.6**	**74.2**	**88.7**	**88.0**	**88.3**	**83.5**	**68.0**

Table 2: Accuracy (percentage correct) for targeted evaluation of contextual phenomena. Our document-level version of COMET-QE substantially outperforms the sentence-level COMET-QE, and also outperforms the best methods proposed by Lopes et al. (2020), demonstrating that it is successfully incorporating contextual information.

BlonDe (Jiang et al., 2022), an overlap-based document-level metric that focuses on discourse phenomena.[7] We also compare to Prism using the m39v1 model released by the authors of Prism.

For discourse phenomena, we compare our document-level COMET-QE model to the sentence-level COMET-QE as well as the best reported results of Lopes et al. (2020).

5 Results

5.1 Correlation with Human Judgments

We present the system-level Pearson correlation with the human annotations of the 2021 WMT metrics task for all metrics (sentence- and document-level) in Table 1. Statistical significance ($p < 0.05$) is computed for each sentence- vs document-level metric pair following Freitag et al. (2021b) using the PERM-BOTH hypothesis test (Deutsch et al., 2021). We also provide the results of BlonDe (only for *→En since this metric relies on entity taggers and discourse markers that are only trained in English) and Prism with the original model (m39v1) for comparison.

Overall, adding document-level context leads to improved correlation with human judgments for all metrics. Our document-level metrics outperform their sentence-level counterparts in 18 of 24 (75%) of cases considered. Excluding Zh→En news data, which has a very low-quality human reference (see § 4.1), our document-level metrics outperform their sentence-level counterparts in 17 of 20 (85%) of cases. Looking at only pairs with statistically significant differences, our document-level metrics outperform their sentence-level counterparts in 6 of 7 cases (86%), and 6 of 6 (100%) of cases excluding Zh→En news.

We see that document-level metrics outperform sentence-level metrics in only 1 of 4 cases on Zh→En news This suggests that the document-level metrics are sensitive to errors in the reference context. This hypothesis is further supported by analysis in § 6.1.

For Prism, we observe that the sentence-level results with the original m39v1 model are better than the sentence-level results with mBART-50. However, by using document-level context we are able to improve over the sentence-level Prism with mBART-50 in every language pair/domain. This narrows the gap between Prism with mBART and Prism with m39v1, outperforming the stronger m39v1 model in two TED language pairs.

Although the COMET models are fine-tuned on single sentences, experimental results suggest they are able to retain their ability to handle inter-sentential dependencies. We considered retraining COMET excluding older direct assessment judgments which did not take document-level context into account; however this would have severely limited the amount of (already very limited!) training data.

Finally, we observe that BlonDe performs significantly worse than the pretrained metrics as well as our document-level extensions, underperforming everything except document-level COMET-QE in TED Zh→En.

5.2 Discourse Phenomena Improvements

We provide the results of targeted evaluation on contrastive datasets for COMET-QE and Doc-COMET-QE in Table 2. We also provide the scores of the best-performing document-MT model for each dataset from Lopes et al. (2020) for comparison. The reference-based metrics are not considered in this section as the use of a reference would make the task trivial.

We observe that the document-level COMET-QE substantially outperforms the sentence-level COMET-QE, and even outperforms document-

[7] We report BlonDe results in English only, as BlonDe uses a discourse marker script from Sileo et al. (2019) which was trained only in English. BlonDe could likely be extended to other languages but we did not attempt to do so.

	Context	Doc-Prism	Doc-BERTScore	Doc-COMET
hypothesis	$\langle c_s; s, c_r; r, c_h; h\rangle$	0.595	0.624	0.630
reference	$\langle c_s; s, c_r; r, c_r; h\rangle$	**0.649**	**0.650**	**0.659**

Table 3: Average correlation with MQM human judgments of our document-level metrics using previous hypothesis sentences as context vs. previous reference sentence as context. COMET-QE is excluded because it does not depend on the reference. For all three methods, we see better correlation using the reference for hypothesis context. We hypothesize that this is because using previous hypothesis sentences allows for propagation of errors (i.e. an error in a previous sentence can impair the judgment of the current sentence).

level translation models optimized for discourse tasks. Surprisingly, we observe improvements in the evaluation of pronoun translation not only when the necessary information is located in a previous sentence (Inter) but even in the case where the antecedent can be found in the same sentence (Intra), suggesting additional context is helpful in these cases as well. Apart from pronoun translation, our approach also improves over both the sentence-level metric and the document-level MT of Lopes et al. (2020) at WSD. These findings all support our hypothesis that our document-level metrics are resolving ambiguities in the reference sentence by using additional context.

6 Ablations

6.1 Hypothesis vs Reference Context

For our document-level MT metrics described prior to this point, we use the reference context c_r (as opposed to the hypothesis context c_h) as context for the hypothesis. Our reasoning behind this decision is that previous translations could contain errors that might bias the document-level metric into rewarding erroneous translations. To test this, we conduct an ablation experiment in which we concatenate the hypothesis context to the hypothesis while the context of the remaining inputs (i.e. the reference and the source sentence) remains unchanged. Table 3 shows the average correlation across all language pairs and domains using either the hypothesis context or the reference context. We do not provide these scores for COMET-QE as it does not have access to the reference.

We observe that the use of the hypothesis context degrades performance for all metrics, which is in line with the findings of Fernandes et al. (2021) for document-level MT. We suspect that this is because the previous hypothesis sentences contain more errors than previous reference sentences, and thus using previous hypothesis sentences allows for more propagation of errors (i.e. an error in a previous sentence can impair the judgment of the current sentence).

One disadvantage of using reference context for the hypothesis is that we cannot measure document-level fluency, that is, how well a document flows from one sentence to the next. Our analysis suggests that either document level fluency is of less concern than error propagation, and/or that MQM judgments are not adequately capturing document-level fluency.

6.2 Amount of Context

In our experiments so far we have used the previous two sentences as context, motivated by the finding of Scherrer et al. (2019); Kim et al. (2019); Castilho et al. (2020) that two previous sentences are sufficient context to resolve ambiguities in the majority of sentences. Figure 2 shows the results for [0, 1, 2] previous sentences as context for news articles and TED talks. In the news domain we observe that for En→De and En→Ru), adding more context helps. On the other hand, for Zh→En, adding context appears to be harmful. We believe this is likely explained by the relatively low-quality human references in Zh→En (see § 4.1). For TED talks, although the results are somewhat noisy, we also observe that more context tends to improve correlation across all three language pairs.

7 Conclusion

We proposed a simple and effective approach to generalize pretrained MT metrics to the document level. We apply our approach to BERTScore, Prism, COMET-QE, and COMET-QE, and we believe that it could easily be extended to other pretrained sentence-level metrics. To the best of our knowledge, our work is the first example of pretrained document-level MT metrics.

We demonstrate that the use of document-level context in pretrained metrics improves correlation with human judgments, and that the improvements

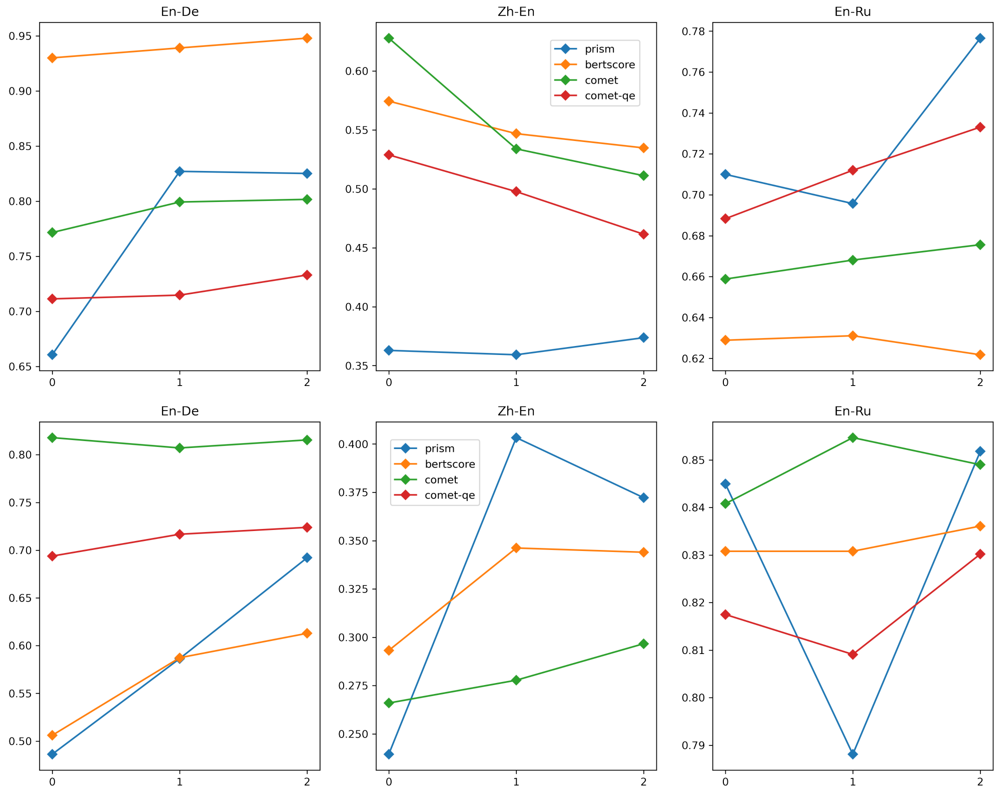

Figure 2: System-level Pearson correlation with human correlation vs. number of sentences of context for News (upper) and TED talks (lower). Although the results are noisy, in general we observe that correlation improves as the amount of context increases. The one exception is Zh→En News, which we attribute to poor human references (see § 4.1).

are likely due to fact that the document-level metrics can resolving ambiguities in the reference sentence by using additional context. We present results on MT evaluation but our approach may also be beneficial in other Natural Language Generation (NLG) tasks where discourse phenomena are present (e.g paraphrasing, data to text generation, chatbots, etc).

In conclusion, we argue that the MT community (and possibly the greater NLG community) should adopt metrics—such as those presented in this work—which take document-level context into account. This would better align automatic metrics with human evaluation, where document-level judgements have been shown to be more discriminative than sentence-level judgements. We also recommend that future research in metrics explore novel ways to incorporate context.

References

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

Sheila Castilho, Maja Popović, and Andy Way. 2020. On context span needed for machine translation evaluation. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3735–3742, Marseille, France. European Language Resources Association.

Mauro Cettolo, Christian Girardi, and Marcello Federico. 2012. WIT3: Web inventory of transcribed and translated talks. In *Proceedings of the 16th Annual conference of the European Association for Machine Translation*, pages 261–268, Trento, Italy. European Association for Machine Translation.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Daniel Deutsch, Rotem Dror, and Dan Roth. 2021. A Statistical Analysis of Summarization Evaluation Metrics Using Resampling Methods. *Transactions of the Association for Computational Linguistics*, 9:1132–1146.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Patrick Fernandes, Kayo Yin, Graham Neubig, and André F. T. Martins. 2021. Measuring and increasing context usage in context-aware machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6467–6478, Online. Association for Computational Linguistics.

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, 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.

Zhengxian Gong, Min Zhang, and Guodong Zhou. 2015. Document-level machine translation evaluation with gist consistency and text cohesion. In *Proceedings of the Second Workshop on Discourse in Machine Translation*, pages 33–40, Lisbon, Portugal. Association for Computational Linguistics.

Mandy Guo, Yinfei Yang, Keith Stevens, Daniel Cer, Heming Ge, Yun-hsuan Sung, Brian Strope, and Ray Kurzweil. 2019. Hierarchical document encoder for parallel corpus mining. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 64–72, Florence, Italy. Association for Computational Linguistics.

Christian Hardmeier and Marcello Federico. 2010. Modelling pronominal anaphora in statistical machine translation. In *Proceedings of the 7th International Workshop on Spoken Language Translation: Papers*, pages 283–289, Paris, France.

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.

Prathyusha Jwalapuram, Shafiq Joty, Irina Temnikova, and Preslav Nakov. 2019. Evaluating pronominal anaphora in machine translation: An evaluation measure and a test suite. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2964–2975, Hong Kong, China. Association for Computational Linguistics.

Yunsu Kim, Duc Thanh Tran, and Hermann Ney. 2019. When and why is document-level context useful in neural machine translation? In *Proceedings of the Fourth Workshop on Discourse in Machine Translation (DiscoMT 2019)*, pages 24–34, Hong Kong, China. Association for Computational Linguistics.

Tom Kocmi, Christian Federmann, Roman Grundkiewicz, Marcin Junczys-Dowmunt, Hitokazu Matsushita, and Arul Menezes. 2021. To ship or not to ship: An extensive evaluation of automatic metrics for machine translation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 478–494, Online. Association for Computational Linguistics.

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.

Pierre Lison, Jörg Tiedemann, and Milen Kouylekov. 2018. OpenSubtitles2018: Statistical rescoring of sentence alignments in large, noisy parallel corpora. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

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. *CoRR*, abs/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.

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

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

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 Werlen and Andrei Popescu-Belis. 2017. Validation of an automatic metric for the accuracy of pronoun translation (APT). In *Proceedings of the Third Workshop on Discourse in Machine Translation*, pages 17–25, Copenhagen, Denmark. 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.

Ricardo Rei, Ana C Farinha, Chrysoula Zerva, Daan van Stigt, Craig Stewart, Pedro Ramos, Taisiya Glushkova, André F. T. Martins, and Alon Lavie. 2021. Are references really needed? unbabel-IST 2021 submission for the metrics shared task. In *Proceedings of the Sixth Conference on Machine Translation*, pages 1030–1040, Online. 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.

Yves Scherrer, Jörg Tiedemann, and Sharid Loáiciga. 2019. Analysing concatenation approaches to document-level NMT in two different domains. In *Proceedings of the Fourth Workshop on Discourse in Machine Translation (DiscoMT 2019)*, pages 51–61, Hong Kong, China. Association for Computational Linguistics.

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.

Damien Sileo, Tim Van De Cruys, Camille Pradel, and Philippe Muller. 2019. Mining discourse markers for unsupervised sentence representation learning. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3477–3486, Minneapolis, Minnesota. Association for Computational Linguistics.

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.

Y. Tang, C. Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning. *ArXiv*, abs/2008.00401.

Brian Thompson and Philipp Koehn. 2020. Exploiting sentence order in document alignment. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5997–6007, Online. Association for Computational Linguistics.

Brian Thompson and Matt Post. 2020a. 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.

Brian Thompson and Matt Post. 2020b. Paraphrase generation as zero-shot multilingual translation: Disentangling semantic similarity from lexical and syntactic diversity. In *Proceedings of the Fifth Conference on Machine Translation*, pages 561–570, Online. Association for Computational Linguistics.

Antonio Toral. 2020. Reassessing claims of human parity and super-human performance in machine translation at WMT 2019. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 185–194, Lisboa, Portugal. European Association for Machine Translation.

Billy T. M. Wong and Chunyu Kit. 2012. Extending machine translation evaluation metrics with lexical cohesion to document level. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1060–1068, Jeju Island, Korea. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating text generation with BERT. In *The 8th International Conference on Learning Representations (ICLR)*, Addis Ababa, Ethiopia. OpenReview.net.