Unbabel's Participation in the WMT20 Metrics Shared Task

Ricardo Rei **Craig Stewart** **Ana C Farinha** **Alon Lavie**
Unbabel AI
{ricardo.rei, craig.stewart, catarina.farinha, alon.lavie}@unbabel.com

Abstract

We present the contribution of the Unbabel team to the WMT 2020 Shared Task on Metrics. We intend to participate on the segment-level, document-level and system-level tracks on all language pairs, as well as the "QE as a Metric" track. Accordingly, we illustrate results of our models in these tracks with reference to test sets from the previous year. Our submissions build upon the recently proposed COMET framework: we train several estimator models to regress on different human-generated quality scores and a novel ranking model trained on relative ranks obtained from Direct Assessments. We also propose a simple technique for converting segment-level predictions into a document-level score. Overall, our systems achieve strong results for all language pairs on previous test sets and in many cases set a new state-of-the-art.

1 Introduction

In this paper we describe our submission to the WMT20 Metrics shared task. Our work is based on the COMET[1] framework, as presented in Rei et al. (2020), and extended here to evaluation of MT output at segment, document and system-level, forming the basis of our submissions to the corresponding task tracks. Recently, automatic evaluation of MT has followed most other sub-fields in NLP with a notable interest in leveraging the power of large, pre-trained language models. Metrics such as BERT REGRESSOR (Shimanaka et al., 2019), BERTSCORE (Zhang et al., 2020), BLEURT (Sellam et al., 2020) and our more recent COMET (Rei et al., 2020), all build upon developments in language modelling to generate automatic metrics with high correlation with human judgement. Our MT evaluation models follow a similar strategy, specifically utilizing the most recent iterations of the XLM-RoBERTa model presented in Conneau et al. (2020).

The uniqueness of our approach comes from our inclusion of the source text as input which was demonstrated in Takahashi et al. (2020) and Rei et al. (2020) to be beneficial to the model. In our contribution to the shared task, we demonstrate methods of further exploiting information in the source text as well as a technique to fully harness the power of pre-trained language models to further improve the prediction accuracy of our evaluation framework when more than one reference translation is available.

For the shared task, we utilize two primary types of models built using the COMET framework, namely; the Estimator models, which regress directly on human scores of MT quality such as Direct Assessment; and the COMET-RANK (base) model used to rank MT outputs and systems.

In addition to the models themselves, we also make the following research contributions:

1. We introduce a method for handling multiple references at inference time and for optimizing the utility of information from all available text inputs

2. We propose a simple technique for calculating a document-level score from a weighted average of segment-level scores

We demonstrate that our COMET framework trained models achieve state-of-the-art results or are competitive on all settings introduced in the WMT19 Metrics shared task, outperforming, in some cases, more recently proposed metrics such as BERTSCORE (Zhang et al., 2020), BLEURT (Sellam et al., 2020) and PRISM (Thompson and Post, 2020).

[1] **C**rosslingual **O**ptimized **M**etric for **E**valuation of **T**ranslation hosted at: https://github.com/Unbabel/COMET

Proceedings of the 5th Conference on Machine Translation (WMT), pages 911–920
Online, November 19–20, 2020.

2 The COMET Framework

As outlined in Rei et al. (2020), the COMET framework allows for training of specialized evaluation metrics that correlate well with different types of human-generated quality scores. The general structure of the framework consists of a cross-lingual encoder that produces a series of token-level vector embeddings for source, hypothesis and reference inputs, a pooling layer which converts the various token-level representations into segment-level vectors for each input, and a predictive neural network that generates a quality score. The latter model can either be trained to regress directly on a score to produce predictions of segment-level quality, or can be trained as a ranker to differentiate MT systems. In our contribution to the shared task, we introduce two varieties of models built on the COMET framework that are extensions of the models evaluated in Rei et al. (2020).

3 COMET Models

3.1 Estimator Models

Our Estimators generally follow the architecture proposed in Rei et al. (2020), that is to say we encode segment-level representations using XLM-RoBERTa and pass these outputs through a feed-forward regressor. As in Rei et al. (2020), we train three versions of this basic estimator model against different types of human judgement; *Human-mediated Translation Edit Rate* (HTER) (Snover et al., 2006), a proprietary implementation of *Multidimensional Quality Metric* (MQM) (Lommel et al., 2014) and (in-line with the present task) *Direct Assessments* (DA) (Graham et al., 2013). The hyper-parameters used for these models are exactly as described in Rei et al. (2020), excluding the following alterations: we use XLM-RoBERTa large instead of base and we increase the feed-forward hidden sizes (from 2304 in the first layer and 1152 in the second to 3072 and 1536 hidden units, respectively). We also keep the embedding layer frozen and apply a layer-wise learning rate decay (as proposed in Howard and Ruder (2018)) by which each transformer layer has a learning rate scaled at 0.95 times the rate of the layer above. By doing this, we hope that our metric generalizes better to new language pairs introduced this year.

3.2 Translation Ranking Model

While for the Estimators using a larger pretrained encoder seems to improve performance we found that for the Translation Ranking Model, larger pretrained encoders lead to training instability and an overall worse performance. For that reason we decided to keep the model proposed in (Rei et al., 2020) without any alteration.

4 Corpora

Below we provide an outline of the various datasets used to train our models:

4.1 HTER Corpora

Our HTER corpus is a concatenation of two publicly available corpora: the QT21 corpus and the APE-QUEST corpus. While the QT21 corpus contains segments from the information technology and life sciences domains (Specia et al., 2017), the APE-QUEST contains segments from the legal domain (Ive et al., 2020). Concatenation of these two corpora gives a total of 211K tuples with source sentence, corresponding human-generated reference, MT hypothesis, and post-edited MT (PE). With regard to the language pairs in each corpus, QT21 covers: English to German (en-de), Latvian (en-lt) and Czech (en-cs), and German to English (de-en); while APE-QUEST covers: English-Dutch (en-nl), English-French (en-fr), English-Portuguese (en-pt). Finally, the HTER score is obtained by calculating the translation edit rate (TER) (Snover et al., 2006) between the MT hypothesis and the corresponding PE. By doing this, we were able to create a large HTER corpus covering several language pairs and different domains.

4.2 MQM Corpus

Our MQM corpus is an extension of the proprietary corpus presented in Rei et al. (2020). This internal data consists of customer support chat messages translated using a domain adapted MT model and their corresponding references (consisting of post-edited translations from earlier iterations of the MT systems). The data was then MQM-annotated according to the guidelines set out in Burchardt and Lommel (2014). Our final corpus contains 27K tuples from English into 15 different languages and/or dialects: German (en–de), Spanish (en–es), Latin-American Spanish (en–es-latam), French (en–fr), Italian (en–it), Japanese (en–ja), Dutch (en–nl), Portuguese (en–pt), Brazilian Portuguese (en–pt-br), Russian (en–ru), Swedish (en–sv), Turkish (en–tr), Polish (en–pl), simplified Chinese (en–zh-CN), and Taiwanese Chinese (en–zh-TW).

4.3 DA Corpora

Every year, since 2008, the WMT News Translation shared task organizers collect human judgements in the form of DAs. Since 2017, due to a lack of annotators, these scores are mapped to relative rankings (DARR). We take advantage of this data in two ways: 1) we use the scores directly in order to train an estimator model, 2) as in Rei et al. (2020), we use the DARR to train a translation ranking system. The collective corpora of 2017, 2018 and 2019 contain a total of 24 language pairs, including low-resource languages such as English-Gujarati (en-gu) and English-Kazakh (en-kk). For the purposes of this paper we use the data from 2017 and 2018 to train and the data from 2019 to validate. Later, for participation in the 2020 shared task, we intend to include the data from 2019 in our training corpus.

5 Segment-Level Task

At segment-level, we take each of our Estimator models trained to predict MQM, HTER and DA and predict segment-level scores on the DARR data from WMT19. We then generate pairwise rankings based on these predicted scores. For each language pair we apply the formulation of Kendall's Tau (τ) from the shared task (Ma et al., 2019) as follows:

$$\tau = \frac{\textit{Concordant} - \textit{Discordant}}{\textit{Concordant} + \textit{Discordant}} \quad (1)$$

Concordant here being the number of times a metric assigns a higher score to the "better" hypothesis h^+ and *Discordant*, the number of times a metric assigns a higher score to the "worse" hypothesis h^-, or that the evaluation was otherwise equal.

6 Document-Level Task

In the WMT2019 News Translation the organizers introduced a document-level translation task (Barrault et al., 2019) for en-de and en-cs. This means that for those language pairs we are able to obtain document-level direct assessments. We can compute a score taking into account an entire document and correlate it with the human evaluation also carried out at document-level.

For our document-level submission we propose the generation of a document-level score as a weighted average of the predicted scores for each segment composing that document (hereinafter called micro-average score), where the same is weighted by segment length.

To calculate this score at inference time we pass the entire document (divided into segments) through the model as a single batch. This has the added effect of reducing inference time.

7 System-level Task

Following previous years, the metric used in the System-level Task will be Pearson's r correlation score. The correlation is calculated between the average of all DA human z-scores for a given system and language pair, and the average of the corresponding scores predicted by a given metric. Because the goal of some metrics is to maximize the correlation with human judgements (i.e. BLEU), while for others is to minimize that correlation (i.e. HTER), the value reported is its absolute value.

7.1 Robustness to high-permorming systems

One important finding from WMT19 is the general deterioration of metrics' performance when considering only the top n MT systems (Ma et al., 2019). Previously, we showed robustness of our metrics in this scenario in terms of Kendall's Tau at segment-level (Rei et al., 2020). Mathur et al. (2020) show that at system-level, Pearson correlation is highly influenced by outliers and that performances for metrics such as BLEU drop significantly when considering only the top systems. To address this, we propose a system-level pairwise comparison measured with the same Kendall's Tau formulation used for segment-level analysis outlined in section 5 above. By doing this, we are not only better handling possible outliers, but emulating a real world application of these metrics: In most cases (both in academia and industry), we want a metric that can successfully differentiate between two systems, even if those systems are very close in terms of performance, which is often the case.

8 Quality Estimation as a Metric

Given the clear parallels between the COMET framework and modern approaches to Quality Estimation such as Kepler et al. (2019), we used our framework to participate in the "QE as a Metric" track of the shared task by removing the reference at input and proportionately reducing the dimensions of the feed-forward network to accommodate the reduced input.

Table 1: Segment-level Kendall's Tau (τ) correlations for language pairs from English-to-other for the WMT19 Metrics DARR corpus.

	en-cs	**en-de**	**en-fi**	**en-gu**	**en-kk**	**en-lt**	**en-ru**	**en-zh**	
Nº Tuples	27178	99840	31820	11355	18172	17401	24334	18658	**avg.**
BLEU	0.364	0.248	0.395	0.463	0.363	0.333	0.4691	0.235	0.410
CHRF	0.444	0.321	0.518	0.548	0.510	0.438	0.548	0.241	0.510
BERTSCORE (F1)	0.486	0.350	0.526	0.559	0.534	0.464	0.581	0.350	0.550
PRISM	0.580	0.416	0.590	-	0.529	0.555	0.581	0.373	0.518
COMET-MQM (large)	0.595	0.405	0.594	0.580	0.546	0.607	0.693	0.400	0.553
COMET-HTER (large)	0.610	0.427	0.610	0.587	0.569	0.615	0.707	0.405	0.566
COMET-DA (large)	**0.618**	**0.435**	0.620	**0.617**	0.585	0.619	**0.711**	0.427	0.579
COMET-RANK (base)	0.603	0.427	**0.664**	0.611	**0.693**	**0.665**	0.580	**0.449**	**0.587**

Table 2: Segment-level Kendall's Tau (τ) correlations on language pairs with English as a target for the WMT19 Metrics DARR corpus.

	de-en	**fi-en**	**gu-en**	**kk-en**	**lt-en**	**ru-en**	**zh-en**	
Nº Tuples	85365	32179	20110	9728	21862	39852	31070	**avg.**
BLEU	0.054	0.236	0.194	0.276	0.249	0.115	0.321	0.206
CHRF	0.123	0.292	0.240	0.323	0.304	0.177	0.371	0.261
BERTSCORE (F1)	0.191	0.354	0.292	0.351	0.381	0.221	0.433	0.318
BLEURT (large-512)	0.174	**0.374**	0.313	0.372	0.388	0.220	0.436	0.325
PRISM	0.189	0.366	**0.320**	0.362	0.382	0.220	0.434	0.325
COMET-MQM (large)	0.191	0.360	0.289	0.346	0.373	0.213	0.426	0.314
COMET-HTER (large)	0.193	0.351	0.286	0.340	0.375	0.209	0.429	0.312
COMET-DA (large)	**0.220**	0.368	0.316	**0.378**	**0.405**	**0.231**	**0.462**	**0.340**
COMET-RANK (base)	0.202	**0.399**	**0.341**	0.358	**0.407**	0.180	0.445	0.333

9 Multi-Reference Handling

In this year's shared task we are provided with a second human-generated reference for German-to-English (de-en), Russian-English (ru-en) and Chinese-to-English (zh-en). Given that our base framework currently supports the input of only one single reference, we introduce a method of leveraging information from a second reference at inference time.

During standard training of our models, we input source, hypothesis and reference in that order, resulting in a concatenation of embeddings as detailed further in Rei et al. (2020). During training, with a probability of $p = 0.5$ we switch the positions of source and reference, such that the system receives the reference as the source and vice versa. This has two primary effects on our model. Firstly, during fine-tuning of the underlying language model, the source embeddings are aligned with the target language embedding space resulting in more useful source embeddings. Secondly, it forces the model to treat source and reference as interchangeable inputs, allowing it to handle switching of inputs at inference time without excessively hindering the model's predictive ability. Finally, at inference time we embed source $\boldsymbol{s}$, hypothesis $\boldsymbol{h}$, reference $\boldsymbol{r}$ and the alternative reference $\hat{\boldsymbol{r}}$. These embeddings are then passed to the feed-forward neural network in the following six permutations: $[\boldsymbol{s}; \boldsymbol{h}; \boldsymbol{r}]$, $[\boldsymbol{r}; \boldsymbol{h}; \boldsymbol{s}]$, $[\boldsymbol{s}; \boldsymbol{h}; \hat{\boldsymbol{r}}]$, $[\hat{\boldsymbol{r}}; \boldsymbol{h}; \boldsymbol{s}]$, $[\boldsymbol{r}; \boldsymbol{h}; \hat{\boldsymbol{r}}]$ and $[\hat{\boldsymbol{r}}; \boldsymbol{h}; \boldsymbol{r}]$.

Six passes through the feed-forward gives us six predictions. Final, aggregated scores are achieved by taking the mean of the six predictions and multiplying it by 1 minus the standard deviation (σ). The intuition being that $1 - \sigma$ gives something of an idea of confidence of the model at the segment-level and that scaling the mean prediction to penalize lower confidence might align better with human judgement.

10 Experimental Results

Below we present results of our various COMET models on WMT19 evaluation sets as described

Table 3: Kendall's Tau (τ) correlation and standard deviation (σ) across all language pairs for the top 5 high-performing systems.

Model	Avg. Kendall (all)	Avg. Kendall (en)
BLEU	0.387±0.366	0.257±0.395
CHRF	0.387±0.463	0.343±0.513
BERTSCORE (F1)	0.453±0.267	0.429±0.279
BLEURT	-	0.571±0.355
PRISM	0.52±0.270	0.514±0.279
COMET-MQM (large)	0.587±0.277	0.543±0.276
COMET-HTER (large)	0.547±0.325	0.486±0.363
COMET-DA (large)	**0.653±0.233**	**0.629±0.269**
COMET-RANK (base)	0.547±0.256	0.543±0.276

above. Segment-level and document-level results are outlined in the corresponding tables within the body of the paper, the remaining tables for other task results are contained in the Appendices hereto.

10.1 Segment-level Task

Our segment-level results on the shared task test sets for WMT19 are detailed in tables 1 and 2. We note that for all language pairs out of English (Table 1) both our DA Estimator and our COMET-RANK (base) outperform prior metrics, in some cases by a significant margins. The same can be said in most language pairs into English, where we consistently perform at the level competitive with or exceeding prior metric performance in this task. Table 6 (contained in the appendices) further illustrates performance of our models on non-English language pairs. We note that in all settings our COMET models outperform state-of-the-art for these language pairs.

10.2 System-level Task

System-level results are outlined in tables 7, 8 and 9 in the appendix. In most language pairs we outperform the best metrics in terms of correlation with human judgement. For those language pairs for which our metrics are outperformed by others, we note that ours are at least competitive with other, recent metrics.

An unexpected result is that at system-level our COMET-RANK (base) does not perform as well as our Estimators, regardless of its strong segment-level results. We believe that this is an artifact of training directly on DARR data. Since in WMT shared tasks, the DA rating scale employed is defined at the 0-25-50-75-100 point margins, the minimum required difference between two hypothesis to produce DARR judgement is 25 points (Ma et al., 2019). All other segments are discarded, as within that range the notion of which hypothesis is better becomes ambiguous. As a result we believe that our ranker model learns to successfully discriminate less ambiguous examples and struggles to correctly assign a score otherwise.

10.2.1 Robustness to high-performing systems

As outlined above, we also complement our evaluation at system-level with an analysis of metric performance in terms of the pairwise ranking of the top five performing systems from each language pair. For each setting we output the Kendall's Tau (that is to say the formulation outlined in section 5 above) and report the mean and standard deviation of results across language pairs in table 3.

In both settings we note that our DA Estimator (large) model significantly outperforms other metrics both in terms of mean and standard deviation. This strongly suggests that not only do we perform well in terms of system-level Pearson but that at a practical level, our model can much more successfully differentiate high-performing systems.

10.3 Document-level Task

Table 10 compares the micro-averaging against a simple unweighted average. From table 10 we can observe that micro-averaging outperforms macro-averaging by a small margin. Table 4 summarizes our results for the Document-level Task using our segment-level Estimators with micro-averaging. In this task, the HTER Estimator shows generally superior performance on average surpassing our best performing segment-level model, the DA Estimator. An important conclusion to draw from the strong document-level correlations noted here is that a

model trained to generate segment-level scores, can also perform well as a document-level metric.

Table 4: Pearson correlation (r) between Document-level DAs and micro average segment-level scores for English-to-German and English-to-Czech.

	en-cs	**en-de**	
Nº Documents	1115	2355	**avg.**
COMET-MQM (large)	0.638	0.516	0.577
COMET-HTER (large)	0.655	**0.558**	**0.607**
COMET-DA (large)	**0.667**	0.528	0.598

Table 5: Pearson correlations (r) and adequacy (as reported in Freitag et al. (2020)) for segment-level DA using our DA Estimator (large) model on WMT19 Metrics shared task test data for en-de. We show Pearson's r for the single reference scenario using the corresponding reference ('1-ref') and the multi-reference scenario where the reference is combined with the original in the manner outlined in section 9 above ('2-ref').

Reference	**Adequacy**	r **(1-ref)**	r **(2-ref)**
WMT	85.3	0.523	-
AR	86.7	0.539	0.555
WMTp	81.8	0.470	0.529
ARp	80.8	0.476	0.537

10.4 Multi-Reference Handling

Additional references were obtained for two language pairs: en-de and de-en. For the former, we conducted experiments using 3 additional references from Freitag et al. (2020): AR reference (an additional high quality reference translation), ARp reference (a paraphrased-as-much-as-possible version of AR), and WMTp reference (a paraphrased-as-much-as-possible version of the original WMT reference); for the latter, we use the alternative reference given in the WMT19 News shared task test set. Conveniently, Freitag et al. (2020) also offer a notion of the quality of the extra references for en-de by providing human-generated adequacy assessments for each. In table 5 we show the performance of our DA Estimator (large) model with each reference, either as a single reference or combined in the manner described in section 9 above with the original reference.

While we lack data to draw any statistically significant conclusions, there is a strong suggestion from these results of a positive correlation between reference quality and utility to the predictive model.

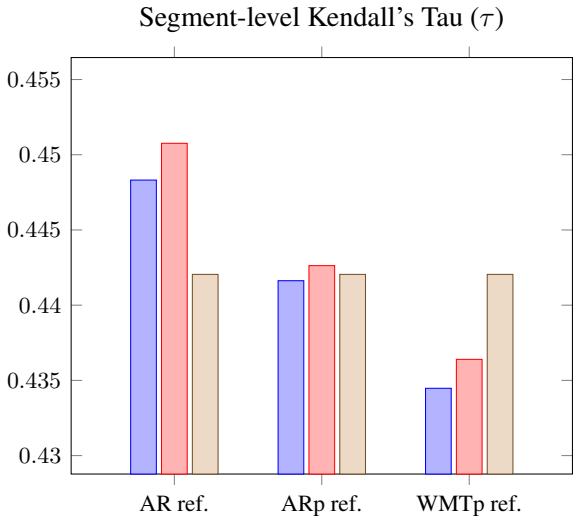

Figure 1: Performance impact of using different kinds of references in combination with the original WMT English-to-German reference. In — we observe the Kendall-Tau τ ranking correlation achieved by our multi-reference Estimator model (section 9). In — we present the Kendall-Tau τ ranking correlation of our "one-reference" Estimator model using the alternative reference. Finally, for comparison, in — we show the Kendall-Tau τ ranking correlation of our "one-reference" Estimator model using the original reference.

For de-en, using an alternative reference did not offer any gain in Pearson's r. We note that when using it alone we only achieve r=0.34 compared to using the original reference which achieves r=0.42. We speculate, based on our observations above, that this might be due to the alternative reference being of lower quality.

These results potentially show that for approaches such as COMET, quality is more important than quantity, and that lower-quality additional references can potentially hurt rather than help improve the correlations obtained using only one single high-quality reference.

With regard to the Kendall Tau measured at segment-level, by looking at Figure 1 (en-de), we see no significant differences in using the multi-reference technique. This suggests that having a higher Pearson's r score does not necessarily guarantee a better Kendall's Tau.

We note that by design, with an approach such as COMET that is based on a meaning-representation of references, extra references are expected to provide only minor additional value, especially versus lexical-based metrics such as BLEU (Papineni et al., 2002). Whereas the adequacy of the reference(s)

is (again by design) expected to have a more significant impact on the performance of the model. Our initial results seem to strongly support this hypothesis.

11 Conclusions

In this paper we present COMET, Unbabel's contribution to the WMT 2020 Metrics shared task. We leverage the framework outlined in Rei et al. (2020) to demonstrate state-of-the-art or otherwise competitive levels of correlation with human judgements in all tasks and introduce a novel method of making optimal use of alternative references and demonstrate that the quality of the reference used is relevant to the success of our framework. Further investigation of the latter, in particular how to better leverage different kinds of references, represent an interesting direction for future work.

12 Acknowledgments

We are grateful to Fabio Kepler, Daan Van Stigt, Miguel Vera, and the reviewers, for their valuable feedback and discussions. This work was supported in part by the P2020 Program through projects MAIA and Unbabel4EU, supervised by ANI under contract numbers 045909 and 042671, respectively.

References

Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019. Findings of the 2019 conference on machine translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy. Association for Computational Linguistics.

Aljoscha Burchardt and Arle Lommel. 2014. Practical Guidelines for the Use of MQM in Scientific Research on Translation quality. *Quality Translation 21*. (access date: 2020-05-26).

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.

Markus Freitag, David Grangier, and Isaac Caswell. 2020. BLEU might be Guilty but References are not Innocent. *ArXiv*, abs/2004.06063.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2013. Continuous measurement scales in human evaluation of machine translation. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 33–41, Sofia, Bulgaria. Association for Computational Linguistics.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339, Melbourne, Australia. Association for Computational Linguistics.

Julia Ive, Lucia Specia, Sara Szoc, Tom Vanallemeersch, Joachim Van den Bogaert, Eduardo Farah, Christine Maroti, Artur Ventura, and Maxim Khalilov. 2020. A post-editing dataset in the legal domain: Do we underestimate neural machine translation quality? In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 3692–3697, Marseille, France. European Language Resources Association.

Fabio Kepler, Jonay Trénous, Marcos Treviso, Miguel Vera, António Góis, M. Amin Farajian, António V. Lopes, and André F. T. Martins. 2019. Unbabel's participation in the WMT19 translation quality estimation shared task. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 78–84, Florence, Italy. Association for Computational Linguistics.

Arle Lommel, Aljoscha Burchardt, and Hans Uszkoreit. 2014. Multidimensional Quality Metrics (MQM): A framework for declaring and describing translation quality metrics. *Tradumàtica: tecnologies de la traducció*, 0:455–463.

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.

Nitika Mathur, Timothy Baldwin, and Trevor Cohn. 2020. Tangled up in BLEU: Reevaluating the evaluation of automatic machine translation evaluation metrics. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4984–4997, Online. 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.

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)*. 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.

Hiroki Shimanaka, Tomoyuki Kajiwara, and Mamoru Komachi. 2019. Machine Translation Evaluation with BERT Regressor. *arXiv preprint arXiv:1907.12679*.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *In Proceedings of Association for Machine Translation in the Americas*, pages 223–231.

Lucia Specia, Kim Harris, Frédéric Blain, Aljoscha Burchardt, Viviven Macketanz, Inguna Skadina, Matteo Negri, , and Marco Turchi. 2017. Translation quality and productivity: A study on rich morphology languages. In *Machine Translation Summit XVI*, pages 55–71, Nagoya, Japan.

Kosuke Takahashi, Katsuhito Sudoh, and Satoshi Nakamura. 2020. Automatic machine translation evaluation using source language inputs and cross-lingual language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3553–3558, Online. Association for Computational Linguistics.

Brian Thompson and Matt Post. 2020. Automatic machine translation evaluation in many languages via zero-shot paraphrasing. *ArXiv*, abs/2004.14564.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations*.

A Appendix A

Table 6: Segment-level Kendall's Tau (τ) correlations on language pairs not involving English for the WMT19 Metrics DARR corpus. COMET-RANK (base) scores are to be replaced with results of the large model.

	de-cs	**de-fr**	**fr-de**	
Nº Tuples	23194	4862	1369	**avg.**
BLEU	0.222	0.226	0.173	0.207
CHRF	0.341	0.287	0.274	0.301
BERTSCORE (F1)	0.356	0.330	0.277	0.321
PRISM	0.452	0.443	**0.421**	0.439
COMET-MQM (large)	0.413	0.422	0.327	0.387
COMET-HTER (large)	0.425	0.449	0.381	0.418
COMET-DA (large)	**0.471**	**0.469**	**0.420**	**0.453**
COMET-RANK (base)	0.389	0.444	0.331	0.388

Table 7: System-level Pearson correlation (r) for the from-English language pairs from WMT19 DA corpus. DARR Ranker (base) scores are to be replaced with results of the large model.

	en-cs	**en-de**	**en-fi**	**en-gu**	**en-kk**	**en-lt**	**en-ru**	**en-zh**	
Nº Systems	11	22	12	11	10	12	12	12	**avg.**
BLEU	**0.988**	0.952	0.978	0.780	0.864	0.979	0.973	0.762	0.910
CHRF	**0.986**	0.983	**0.988**	0.839	0.969	0.964	0.979	0.822	0.941
BERTSCORE (F1)	0.983	0.990	0.969	0.907	0.983	0.972	**0.989**	0.927	0.965
PRISM	0.964	0.987	0.947	-	0.978	0.929	0.914	0.900	0.946
COMET-MQM (large)	0.943	0.968	0.949	0.946	0.979	**0.985**	0.966	0.958	0.962
COMET-HTER (large)	0.948	**0.991**	0.959	0.948	0.965	**0.982**	0.973	0.943	0.964
COMET-DA (large)	0.964	**0.995**	0.969	**0.964**	**0.989**	**0.982**	**0.987**	**0.969**	**0.977**
COMET-RANK (base)	0.943	0.937	0.914	0.817	0.963	0.973	0.861	0.942	0.919

Table 8: System-level Pearson correlation (r) for the into-English language pairs from WMT19 DA corpus. DARR Ranker (base) scores are to be replaced with results of the large model.

	de-en	**fi-en**	**gu-en**	**kk-en**	**lt-en**	**ru-en**	**zh-en**	
Nº Systems	16	11	9	7	11	13	15	**avg.**
BLEU	0.879	0.984	0.975	0.959	0.969	0.840	0.895	0.929
CHRF	0.916	**0.988**	0.967	0.982	0.938	0.942	0.952	0.955
BERTSCORE (F1)	0.949	0.984	**0.990**	**0.995**	0.961	0.901	0.982	0.966
BLEURT (large-512)	0.939	0.984	0.989	0.989	**0.992**	**0.980**	**0.994**	**0.981**
PRISM	**0.954**	0.981	**0.992**	**0.992**	**0.994**	0.905	**0.992**	0.973
COMET-MQM (large)	0.926	0.974	0.972	0.971	0.986	0.889	0.959	0.954
COMET-HTER (large)	0.918	0.953	0.958	0.951	0.983	0.924	0.978	0.952
COMET-DA (large)	0.946	0.983	**0.993**	**0.996**	**0.993**	0.970	**0.993**	**0.982**
COMET-RANK (base)	0.922	0.981	0.963	0.932	0.987	0.674	0.967	0.918

Table 9: System-level Pearson correlation (r) for language pairs not involving English from WMT19 DA corpus.

	de-cs	**de-fr**	**fr-de**	
N^o Systems	9	11	10	**avg.**
BLEU	0.936	0.934	0.869	0.913
CHRF	**0.994**	0.933	0.908	0.945
BERTSCORE (F1)	0.988	0.953	0.942	0.961
PRISM	0.988	0.924	0.922	0.945
COMET-MQM (large)	0.936	0.950	0.885	0.924
COMET-HTER (large)	0.951	0.901	0.924	0.925
COMET-DA (large)	0.973	**0.972**	**0.954**	**0.966**
COMET-RANK (base)	0.819	0.941	0.927	0.896

Table 10: Document-level Pearson correlation (r) for micro average and macro average for English-to-German and English-to-Czech.

	en-cs		**en-de**	
	Micro-avg.	Macro-avg.	Micro-avg.	Macro-avg.
COMET-DA (large)	**0.667**	0.660	0.528	**0.529**
COMET-MQM (large)	0.638	**0.639**	0.516	**0.519**
COMET-HTER (large)	**0.655**	0.650	**0.558**	0.552
	0.653	0.649	**0.534**	0.533