Are References Really Needed? Unbabel-IST 2021 Submission for the Metrics Shared Task

Ricardo Rei[1,2,4] **Ana C. Farinha**[1] **Chrysoula Zerva**[2,3] **Daan van Stigt**[1] **Craig Stewart**[1]
Pedro G. Ramos[1] **Taisiya Glushkova**[2,3] **André F. T. Martins**[1,2,3] **Alon Lavie**[1]
[1]Unbabel [2]Instituto Superior Técnico [3]Instituto de Telecomunicações [4]INESC-ID
[2,3,4]Lisbon, Portugal
{ricardo.rei, chrysoula.zerva, taisiya.glushkova, andre.t.martins}@tecnico.ulisboa.pt
{catarina.farinha, daan.stigt, craig.stewart, pedro.ramos, alon.lavie}@unbabel.com

Abstract

In this paper, we present the joint contribution of Unbabel and IST to the WMT 2021 Metrics Shared Task. With this year's focus on *Multidimensional Quality Metric* (MQM) as the ground-truth human assessment, our aim was to steer COMET towards higher correlations with MQM. We do so by first pretraining on *Direct Assessments* and then fine-tuning on z-normalized MQM scores. In our experiments we also show that reference-free COMET models are becoming competitive with reference-based models, even outperforming the best COMET model from 2020 on this year's development data. Additionally, we present COMETINHO, a light-weight COMET model that is 19x faster on CPU than the original model, while also achieving state-of-the-art correlations with MQM. Finally, in the "QE as a metric" track, we also participated with a QE model trained using the OPENKIWI framework leveraging MQM scores and word-level annotations.

1 Introduction

In this paper, we present the joint contribution of Unbabel and IST to the WMT 2021 Shared Task on Metrics. We participated in the segment-level and system-level tracks, as well as the "QE as a Metric" task.

Similar to our participation last year (Rei et al., 2020b), most of the models are based on the COMET framework[1] (Rei et al., 2020a). In last year's shared task (Mathur et al., 2020), COMET along with other trainable metrics such as PRISM (Thompson and Post, 2020) and BLEURT (Sellam et al., 2020) showed superior correlations with the *Direct Assessments* (DA) collected for the News Translation Shared Task. This year, we build on top of the models used last year to take into account that human assessments will be carried out using variants of the *Multidimensional Quality Metric* (MQM) (Lommel et al., 2014) framework and no longer based on DA (Graham et al., 2013). For this reason, we extended our training dataset to include DA evaluations from WMT ranging 2015 to 2020, with the exception of *en-de* and *zh-en* for which we do not include the 2020 data given that the same is included in the MQM development data (Freitag et al., 2021). Finally, we fine-tuned these new models on the z-normalized MQM scores provided for this year's shared task.

One of the remaining redeeming qualities of automated metrics such as BLEU (Papineni et al., 2002) is that they are incredibly light-weight. Despite the higher correlation with human judgement, trainable metrics tend to be slower to run. In an effort to close this gap we present COMETINHO, a light-weight model based on the COMET framework that replaces the original XLM-R large encoder with MiniLMv2 (Wang et al., 2020). This model is approximately 19x faster at inference time compared to the original COMET model (Rei et al., 2020a) and maintains state-of-the-art correlations with MQM in reference-based evaluations.

For the "QE as a metric" track, we show that reference-free evaluation models can reach surprisingly high correlations with human judgements and are competitive with their corresponding reference-based models. Last year we also participated with a similar model in the Metrics Shared Task, but here we elaborate in more detail on the primary differences between this model architecture and other COMET models.

Finally, and for the first time, we submit and describe a reference-free model that in addition to learning from MQM scores also makes use of word-level error annotations. This is possible this year given the shift in evaluation method from DA

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

Proceedings of the Sixth Conference on Machine Translation (WMT), pages 1030–1040
November 10–11, 2021.

Tags	OK	OK	OK	OK	OK	OK	OK	OK	OK	OK	OK	OK	BAD	BAD
MT	the	main	purpose	of	this	project	is	to	design	a	car	for	blind	driving.

Source: 这个项目的主要目的 是设计一辆盲人驾驶的车。
Reference: the main goal of this project is to develop a car for the blind.

Table 1: Example of word-level OK and BAD tags produced by our OPENKIWI model trained with word-level annotation spans. This translation received an overall sentence score of 0.2 and the model was able to identify that the words "blind driving" are translation errors giving a good insight on why the sentence score is low.

to MQM. This model uses the OPENKIWI [2] architecture and its word-level tagging feature to predict OK/BAD word tags along with a sentence-level quality score.

2 The COMET Framework

For a more comprehensive description of the COMET architecture we direct the reader to the original paper (Rei et al., 2020a). Below we will highlight some relevant features that contrast with the COMET reference-free model (COMET-QE). In COMET we encode segment-level representations using the pretrained, cross-lingual, model XLM-RoBERTa (Conneau et al., 2020). Even though we encode the source, the hypothesis, and the reference (i.e. the human curated translation of the source) separately, their embeddings are mapped into a shared feature space. Subsequently, we obtain combined features using the three embeddings ($\boldsymbol{s}$, $\boldsymbol{h}$, and $\boldsymbol{r}$, for the source, hypothesis, and reference, respectively): $\boldsymbol{h} \odot \boldsymbol{s}$, $\boldsymbol{h} \odot \boldsymbol{r}$, $|\boldsymbol{h} - \boldsymbol{s}|$, and $|\boldsymbol{h} - \boldsymbol{r}|$. These features, concatenated to $\boldsymbol{r}$ and $\boldsymbol{h}$ and the resulting vector is the input to a feed-forward regressor.

2.1 Reference-free COMET

The architecture of the COMET model used in the "QE as a metric" task (COMET-QE) is very similar to the main COMET model (Rei et al., 2020a) briefly described above and RUSE (Shimanaka et al., 2018). The biggest difference being that in the COMET-QE model the reference is not used and, consequently, the combination of features used as input to the feed-forward regressor are also different from reference-based COMET. In this case, the combined features are simply: $\boldsymbol{h} \odot \boldsymbol{s}$ and $|\boldsymbol{h} - \boldsymbol{s}|$; the final vector to the feed-foward regressor being the concatenation of the latter features together with $\boldsymbol{h}$ and $\boldsymbol{s}$. A schematic representation is shown in Figure 1.

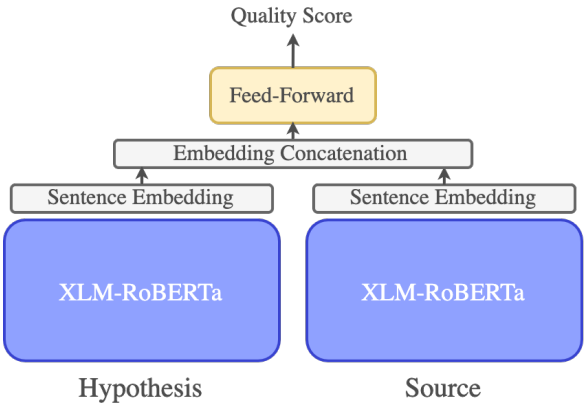

Figure 1: The COMET-QE model follows the dual encoder architecture proposed in RUSE (Shimanaka et al., 2018) but replacing the reference translation with the source sentence.

3 Lightweight COMET: COMETINHO

Our light-weight version of the original COMET model is almost an exact replica in terms of architecture save that we replaced the underlying pre-trained encoder with MiniLMv2 (Wang et al., 2020) which is a distilled version of XLM-R large (Conneau et al., 2020). This distilled model is made available by HuggingFace Transformers (Wolf et al., 2020): `nreimers/mMiniLMv2-L6-H384-distilled-from-XLMR-Large`

Our COMETINHO models are 19x faster on CPU and 14.3x times faster on GPU than COMET models based on XLM-R large. Also, in terms of disk footprint, these models are 5x smaller [3].

4 The OPENKIWI Framework

When using the MQM framework for the calculation of the quality score, human annotators seek to identify and annotate error spans at the word-level, as well as the severity of those errors. We

[2] **OpenKiwi** hosted at: https://github.com/Unbabel/OpenKiwi

[3] Contrastive inference times were tested using a 2.3 GHz Intel Core i5 for CPU, and using a Nvidia T4 for GPU.

leveraged these word-level annotations using the OpenKiwi framework (Kepler et al., 2019), by transforming each word into an OK or BAD tag. In the OpenKiwi architecture, in contrast with COMET-QE, source and hypothesis are jointly encoded. A sentence pair representation is then obtained using average pooling over the hypothesis word embeddings and then used as features to a feed-forward regression layer that learns to produce a sentence level score. At the same time, the word embeddings from the hypothesis are used to predict OK/BAD tags and therefore, the model is trained in a multitask setting (regression and sequence labelling).

5 Corpora

In this year's shared task the organisers provided a development set with MQM annotations for the *en-de* and *zh-en* participating systems on WMT20 (Freitag et al., 2021). Apart from the official development data we used all the Direct Assessments available from previous years.

5.1 Multi-dimensional Quality Metric Corpus

In this corpus, for each language pair, each translation was annotated by 3 raters from a pool of 6. Following what is a common practice for the DA's we convert the segment-level scores of each annotator into a z-normalized score and the final translation quality score is an average of the 3 z-scores. Also, because the sign of these MQM annotations is the opposite of the Direct Assessments we invert the score. Subsequently we generate a train and test split leaving 20% of the documents for each language pair for testing. This results in a total of 11230 *en-de* training samples and 15600 *zh-en* training samples, with testsets of 2950 and 4400 samples, respectively. All results reported in this paper are with respect to the above train and test split. The documents contained in each split are listed in the Appendix of this paper.

Annotators are not always consistent and the annotations of one annotator might differ from another (Graham et al., 2017). With this in mind, we decided to calculate the Kendall's Tau correlation between all annotators as a measure of inter-annotator agreement (Figure 2). The inter-annotator Kendall Tau can then be used as a ceiling effect for the developed metrics which ideally should behave as an additional annotator.

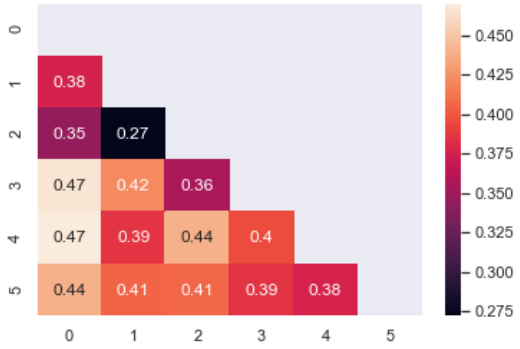

Figure 2: Kendall Tau Correlations between the *en-de* annotators used to develop the shared task development set (Freitag et al., 2021).

For training of the OpenKiwi model described herein we used proprietary MQM data from the customer support domain, covering several industries such as tech industry and travel industry. This data is composed by 1.1M (source, hypothesis) pairs with corresponding MQM annotations from 38 language pairs mostly out-of-english.

5.2 Direct Assessments

Each year, the WMT News Translation shared task organisers collect human judgements in the form of Direct Assessments. Those assessments are then used in the Metrics task to measure the correlation between metrics and therefore decide which metric works best. In recent years researchers have been using these annotations to create trainable metrics that regress on these scores (Shimanaka et al., 2018; Sellam et al., 2020; Rei et al., 2020a). We follow the same approach and use Direct Assessments ranging from 2015 to 2020 for training. The collective corpora contain a total of 33 language pairs including low-resource languages such as English-Tamil (*en-ta*) and a total of 795269 tuples with source, hypothesis, reference and direct assessment z-score. The only exception to this data is that we did not include the *en-de* and *zh-en* assessment from 2020 because they overlap with the MQM development data described in section 5.1.

6 Segment-level task

The COMET framework is highly flexible and easy to adapt to different types of human judgements (Rei et al., 2020a). This year we first pre-trained on the DA collected from 2015 to 2020 except for *en-de* and *zh-en* as described above. Like in Glushkova et al. (2021) we trained 5 models for 1 epoch each using 5 different seeds and created an ensembled

	Nº Segments	zh-en 4400		en-de 2950			
		Pearson	Kendall	Pearson	Kendall	Pearson Avg.	Kendall Avg.
Baselines	BLEURT	0.492	0.405	0.107	0.060	0.299	0.232
	PRISM	0.399	0.337	0.072	0.020	0.235	0.178
	BERTSCORE	0.441	0.344	0.116	0.060	0.279	0.202
	BLEU	0.196	0.275	0.062	0.004	0.129	0.140
	CHRF	0.267	0.219	0.119	0.059	0.193	0.139
	COMET-DA (2020)	0.538	0.435	0.425	0.282	0.481	0.359
Ref. based	COMET-DA (2021)	0.559	0.454	0.464	0.309	0.511	0.382
	COMET-MQM (2021)	**0.717**	**0.546**	**0.488**	**0.361**	**0.602**	**0.454**
	COMETINHO-DA	0.484	0.386	0.299	0.204	0.392	0.295
	COMETINHO-MQM	0.670	0.496	0.311	0.237	0.490	0.367
Ref. Free	COMET-QE-DA (2021)	0.567	0.436	**0.497**	0.308	0.532	0.372
	COMET-QE-MQM (2021)	**0.720**	**0.531**	0.470	**0.359**	**0.595**	**0.445**
	OPENKIWI	0.522	0.385	0.448	0.287	0.485	0.336

Table 2: Segment-level correlations on the *en-de* and *zh-en* testset.

model (COMET-DA). During our experiments we tested two ensembling techniques; averaging the different model predictions and averaging the parameters from the 5 models. Those two approaches had similar results but in the end we decided to use the later one for performance.

Subsequently, we fine-tuned each of the 5 models on the MQM data provided as development for another epoch. As before, we performed weight averaging to obtain an ensemble of those models (COMET-MQM). In both the pre-training and fine-tuning we only perform 1 training epoch in order to ensure that the final models are able to generalise to many language pairs and do not overfit to the News domain. This is especially important since the MQM dataset only contains *en-de* and *zh-en*.

For COMETINHO, as previously mentioned, we used the distilled version of XLM-R (MiniLMv2), available through Hugging Face, and we followed the same training recipe where we pre-train the model using DA's for 1 epoch and then we adapt the model to the MQM data for another epoch.

7 System-level task

For the System-level task we compute the system-level score for each system by averaging the segment-level scores obtained. This follows the same approach used to compute system-level scores based on segment-level human annotations such as DA's and MQM which means that a metric that achieves strong segment-level correlation should also achieve strong system-level performances.

8 QE as a Metric Task

We trained a reference-free model (COMET-QE) in the same way we did for reference-based COMET models described in section 6. As described in section 2.1, the primary difference between the two models is the inclusion or exclusion of the source as input.

9 Experimental Results

9.1 Segment-level task

Reference-based segment-level correlations on the *en-de* and *zh-en* testsets are shown in Table 2. We used both Pearson and Kendall Tau correlation metrics to evaluate our models. As baselines we used lexical metrics such as CHRF (Popović, 2015) and BLEU (Papineni et al., 2002), an embedding-based metric BERTSCORE (Zhang et al., 2020) and three trainable-metrics; BLEURT (Sellam et al., 2020), PRISM (Thompson and Post, 2020) and COMET-DA (2020) (Rei et al., 2020b).

The fact that the COMET-DA (2021) gives higher correlations than the COMET-DA (2020) shows that adding more training data and combining checkpoints trained on different seeds already provides a boost in performance. However, fine-

		All systems			Human vs MT		
		en-de	en-zh		en-de	en-zh	
Nº Comparisons		45	45		21	16	
		Kendall		Avg	Kendall		Avg
Baselines	BLEU	0.378	0.311	0.345	0.095	0.077	0.086
	CHRF	0.444	0.422	0.433	0.143	0.000	0.072
	BERTSCORE (F1)	0.356	0.356	0.356	0.143	0.000	0.072
	PRISM	0.444	0.422	0.433	0.143	0.077	0.110
	COMET-DA (2020)	0.822	0.533	0.678	0.714	0.231	0.473
Ref. based	COMET-DA (2021)	0.844	0.489	0.667	0.761	0.231	0.496
	COMET-MQM (2021)	0.867	0.778	0.823	0.762	0.875	0.819
	COMETINHO-DA	0.533	0.378	0.456	0.238	0.000	0.119
	COMETINHO-MQM	0.355	0.311	0.333	0.095	0.000	0.048
Ref. Free	COMET-QE-DA (2021)	0.778	0.778	0.778	0.667	0.938	0.803
	COMET-QE-MQM (2021)	0.933	0.800	0.867	**1.000**	**1.000**	**1.000**
	OPENKIWI	0.822	0.733	0.778	0.762	0.769	0.766

Table 3: System-level Kendall's Tau (τ) correlations for all system combinations (on the left) and Human vs MT (on the right).

tuning on the MQM development data was the most significant addition to previous work: the COMET-MQM (2021) model increased on average more than 0.1 Pearson correlation. This improvement is consistent with regard to the two COMETINHO models (with COMETINHO-MQM having notably higher correlations than COMETINHO-DA). Nevertheless, the fact that COMETINHO-DA has competitive or state-of-the-art performance with all the other metrics such as BLEURT, PRISM, and BERTSCORE, while also being much faster, presents an ideal opportunity for future work to investigate the incorporation of trainable metrics into the training objectives of MT systems.

For reference-free metrics, the fine-tuning on the MQM data, on average, gave a boost in performance (the only exception being the Pearson correlation for the *en-de* where COMET-QE-DA has a slightly higher correlation than COMET-QE-MQM). Overall, it is somewhat surprising that COMET-QE-* (2021) and COMET-* (2021) show relatively comparable correlations, suggesting that using the reference as input for MT evaluation might be less useful than expected and could feasibly become redundant. This surprising result was also reported by Kocmi et al. (2021) and is especially important since curating reference sentences is usually costly and time consuming and can introduce undesired bias in the evaluation (Freitag et al., 2020).

Finally, the OPENKIWI model has competitive correlations when looking to other trainable metrics and to COMET models that were not fine-tuned on the MQM development data. This add further weight to the suggestion above that references might not add substantial value to MT evaluation. Its performance is even more surprising when considering the fact that this model was train with data from a completely different domain.

It is worth highlighting that the Kendall's Tau correlations for all models (with exception of the two reference-based COMETINHO models) are in the range obtained for correlations between different annotators, for *en-de*, Figure 1. This further validates the value of our models.

9.2 System-level task

System-level results are presented in Table 3 where we report a Kendall Tau correlation defined as follows:

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

where *Concordant* defined as the number of times a metric agrees with humans that a given system x is better than a given system y and *Discordant* is the opposite. These decisions are the computed for all combinations of systems in the testset.

Due to the low number of systems and the relative proximity of the ground-truth MQM system scores we also compare metrics on their ability to distinguish human references from MT outputs. With reference to table 7 in the appendix we note that, for *zh-en*, all 8 MT systems demonstrate comparable performance but that there is a clear separation of human translations. For that reason Table 3 also presents the Kendall Tau correlations considering only "Human" systems against MT systems where we can observe that reference-free metrics achieve better performance. This results confirms the finding from last year's shared task (Mathur et al., 2020) where COMET-QE was highlighted as being the only metric able to differentiate human translations from MT.

10 Related work

Classic n-gram matching MT evaluation metrics such as BLEU (Papineni et al., 2002) have been adopted by the MT community as a primary form of MT evaluation, yet, in the recent years of the WMT Metrics shared task (Bojar et al., 2017; Ma et al., 2018, 2019; Mathur et al., 2020) these classic metrics have been outperformed first by embedding-based alternatives and more recently by trainable metrics based on pre-trained models.

With the rise of word embeddings (Pennington et al., 2014; Peters et al., 2018; Devlin et al., 2019), metrics such as BLEU2VEC (Tättar and Fishel, 2017) and MEANT 2.0 (Lo, 2017) replaced the typical word/n-gram matching by fuzzy matches based on distributional word representations. These metrics appeared for the first time at the WMT Metrics task in 2017 with MEANT 2.0-SRL achieving the highest results at segment-level. In 2018 and 2019 YISI-1 (Lo, 2019), a successor of MEANT 2.0 (Lo, 2017), was among the winners of the WMT Metrics task. YISI-1 (Lo, 2019) mostly takes advantage of BERT embeddings (Devlin et al., 2019) to create soft alignments between hypothesis and reference.

Trainable metrics started as simple regressions based on lexical features (e.g BLEND (Ma et al., 2017)) but nowadays these metrics also use embeddings to extract features that are then used to regress on quality assessments. The first of such metrics were RUSE (Shimanaka et al., 2018) and ESIM (Mathur et al., 2019) which were based on RNN encoders and worked mostly for English. In 2020, BLEURT (Sellam et al., 2020) and COMET (Rei et al., 2020a) were proposed. Both metrics used pre-trained transformer based encoders to extract sentence-level features that are then passed to a regression model; the difference is that COMET also extracts features for the source segment which was something overlooked by predecessor metrics. In the 2020 Metrics Shared task both COMET and BLEURT achieved some of the highest correlations with human judgements and shared the podium with PRISM (Thompson and Post, 2020)

11 Conclusions

In this paper we present the Unbabel-IST's contribution to the WMT 2021 Metrics shared task which for the first time, introduced evaluation using MQM. Our specific contributions include; the fine-tuning of Direct Assessment based models on MQM data which yields impressive gains on the described test sets and a new, lightweight COMET model which achieves comparable performance to its predecessors. Such a light model can provide interesting opportunities for future work into the incorporation of modern metrics into MT training. Finally, but perhaps our most important contributions; we further validate the observations in (Kocmi et al., 2021) that QE as a metric is becoming competitive as an alternative to reference-based evaluation, and, we show that a word-level QE system can be successfully trained on MQM annotations and be competitive with current trainable metrics while providing some intuition about "what" is wrong with a specific translation.

Acknowledgments

We are grateful to Fabio Kepler and José G. C. de Souza for their valuable feedback and discussions. This work was supported by the P2020 programs MAIA (contract 045909) and Unbabel4EU (contract 042671), by the European Research Council (ERC StG DeepSPIN 758969), and by the Fundação para a Ciência e Tecnologia through contract UIDB/50008/2020.

References

Ondřej Bojar, Yvette Graham, and Amir Kamran. 2017. Results of the WMT17 metrics shared task. In *Proceedings of the Second Conference on Machine Translation*, pages 489–513, Copenhagen, Denmark. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco

Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 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.

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.

Markus Freitag, George F. Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *CoRR*, abs/2104.14478.

Markus Freitag, David Grangier, and Isaac Caswell. 2020. BLEU might be guilty but references are not innocent. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 61–71, Online. Association for Computational Linguistics.

Taisiya Glushkova, Chrysoula Zerva, Ricardo Rei, and André F. T. Martins. 2021. Uncertainty-Aware Machine Translation Evaluation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, Online. Association for Computational Linguistics.

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

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.

Fabio Kepler, Jonay Trénous, Marcos Treviso, Miguel Vera, and André F. T. Martins. 2019. OpenKiwi: An open source framework for quality estimation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 117–122, Florence, Italy. 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. *CoRR*, abs/2107.10821.

Chi-kiu Lo. 2017. MEANT 2.0: Accurate semantic MT evaluation for any output language. In *Proceedings of the Second Conference on Machine Translation*, pages 589–597, Copenhagen, Denmark. 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, 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, Ondřej Bojar, and Yvette Graham. 2018. Results of the WMT18 metrics shared task: Both characters and embeddings achieve good performance. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 671–688, Belgium, Brussels. Association for Computational Linguistics.

Qingsong Ma, Yvette Graham, Shugen Wang, and Qun Liu. 2017. Blend: a novel combined MT metric based on direct assessment — CASICT-DCU submission to WMT17 metrics task. In *Proceedings of the Second Conference on Machine Translation*, pages 598–603, Copenhagen, Denmark. Association for Computational Linguistics.

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. 2019. Putting evaluation in context: Contextual embeddings improve machine translation evaluation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2799–2808, Florence, Italy. 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.

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.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. 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 2227–2237, New Orleans, Louisiana. 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, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020a. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020b. Unbabel's participation in the WMT20 metrics shared task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 911–920, Online. Association for Computational Linguistics.

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

Hiroki Shimanaka, Tomoyuki Kajiwara, and Mamoru Komachi. 2018. RUSE: Regressor using sentence embeddings for automatic machine translation evaluation. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 751–758, Belgium, Brussels. Association for Computational Linguistics.

Andre Tättar and Mark Fishel. 2017. bleu2vec: the painfully familiar metric on continuous vector space steroids. In *Proceedings of the Second Conference on Machine Translation*, pages 619–622, Copenhagen, Denmark. 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.

Wenhui Wang, Hangbo Bao, Shaohan Huang, Li Dong, and Furu Wei. 2020. Minilmv2: Multi-head self-attention relation distillation for compressing pretrained transformers. *CoRR*, abs/2012.15828.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

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.1 COMET Hyper-Parameters

In Table 5 is an excerpt of the training configuration used for training the COMET-DA model and Table 5 for the COMET-QE-DA. Then these models are fine-tuned for 1 extra epoch with same hyperparameters except the `learning_rate` that is decreased to $1.0e-05$ and the `nr_frozen_epochs` which we increase to 1 to completely freeze the encoder model.

A.2 OPENKIWI Hyper-Parameters

The hyperparameters used for the OpenKiwi model are expressed in Table 4 and follows the configurations proposed in the sample file of the `github` repository[4].

System	
batch_size	2
Encoder	
hidden_size	1024
Decoder	
bottleneck_size	1024
dropout	0.05
hidden_size	1024
Optimizer	
class_name	adam
encoder_learning_rate	0.0001
learning_rate_decay	1.0
learning_rate_decay_start	0
learning_rate	0.0001
Trainer	
training_steps	2180
early_stop_patience	10
validation_steps	0.5
gradient_accumulation_steps	4
gradient_max_norm	1.0

Table 4: Hyperparameters for OPENKIWI MQM model

nr_frozen_epochs	0.3
keep_embeddings_frozen	True
optimizer	AdamW
encoder_learning_rate	1.0e-05
learning_rate	3.1e-05
layerwise_decay	0.95
encoder	XLM-RoBERTa
pretrained_model	xlm-roberta-large
pool	avg
layer	mix
dropout	0.15
batch_size	4
gradient_accumulation_steps	4
hidden_sizes	[3072, 1024]
epochs	1

Table 5: Hyper-parameters for fine-tuning Reference-based COMET model on Direct Assessments.

nr_frozen_epochs	0.3
keep_embeddings_frozen	True
optimizer	AdamW
encoder_learning_rate	1.0e-05
learning_rate	3.1e-05
layerwise_decay	0.95
encoder	XLM-RoBERTa
pretrained_model	xlm-roberta-large
pool	avg
layer	mix
dropout	0.15
batch_size	4
gradient_accumulation_steps	4
hidden_sizes	[2048, 1024]
epochs	1

Table 6: Hyper-parameters for fine-tuning Reference-free COMET model on Direct Assessments.

[4] https://github.com/Unbabel/OpenKiwi/blob/master/config/xlmroberta.yaml

en-de		zh-en	
System	MQM	System	MQM
Human-B.0	0.794	Human-A.0	3.114
Human-A.0	0.933	Human-B.0	3.149
Human-P.0	1.547	Huoshan_Translate.919	5.077
Tohoku-AIP-NTT.890	2.043	Tencent_Translation.1249	5.163
OPPO.1535	2.284	OPPO.1422	5.309
Tencent_Translation.1520	2.333	THUNLP.1498	5.389
Online-B.1590	2.516	DeepMind.381	5.442
eTranslation.737	2.530	WeChat_AI.1525	5.469
Huoshan_Translate.832	2.600	DiDi_NLP.401	5.484
Online-A.1574	3.189	Online-B.1605	5.512

Table 7: System-level Ranking and corresponding MQM scores for the test split described in section 5.1

A.3 Train/Test Split Documents

In our train/test split described in section 5.1 we leave the following documents for testing:

- *reuters.276709*
- *cnbc.com.33889*
- *cnn.385672*
- *aj-english.8643*
- *express.co.uk.10983*
- *cbsnews.302258*
- *sky.com.20683*, *chicago_defender.80*
- *sciencedaily.com.75569*
- *seattle_times.7141*
- *huffingtonpost.com.19389*
- *huffingtonpost.com.19385*
- *upi.205721*
- *dailymail.co.uk.365293*
- *upi.205735*
- *standard.co.uk.14562*
- *foxnews.100085*
- *allafrica.15342*
- *abcnews.364021*
- *kcal.279*
- *sky.com.20667*
- *en.ndtv.com.13143*
- *reuters.276541*
- *heraldscotland.com.7318*
- *foxnews.100073*
- *upi.205695*
- *tsrus.cn.2113*
- *chinanews.com.102574*
- *chinanews.com.102805*
- *chinanews.com.102708*
- *xinhua-zh-01.6415*
- *chinanews.com.102657*
- *chinanews.com.102700*
- *chinanews.com.102573*
- chinanews.com.102534
- *chinanews.com.102914*
- *tsrus.cn.2112*
- *xinhua-zh-01.6608*
- *australian-zh.104*
- *chinanews.com.102580*
- *xinhua-zh-01.6520*
- *chinanews.com.102767*
- *chinanews.com.102748*
- *chinanews.com.102807*
- *international_times-zh.165*
- *chubun-zh.1066*
- *international_times-zh.160*
- *international_times-zh.150*
- *xinhua-zh-01.6434*
- *xinhua-zh-01.6586*
- *xinhua-zh-01.6307*
- *xinhua-zh-01.6529*
- *chinanews.com.102780*
- *hunan_ribao-zh.199*
- *chinanews.com.102737*
- *chinanews.com.102722*
- *chinanews.com.102709*