COMET-22: Unbabel-IST 2022 Submission for the Metrics Shared Task

Ricardo Rei[*1,2,4], **José G. C. de Souza**[1], **Duarte M. Alves**[1,4], **Chrysoula Zerva**[3,4] **Ana C Farinha**[1], **Taisiya Glushkova**[3,4], **Alon Lavie**[1], **Luisa Coheur**[2,4], **André F. T. Martins**[1,3,4]
[1]Unbabel, Lisbon, Portugal, [2]INESC-ID, Lisbon, Portugal
[3]Instituto de Telecomunicações, Lisbon, Portugal
[4]Instituto Superior Técnico, University of Lisbon, Portugal

Abstract

In this paper, we present the joint contribution of Unbabel and IST to the WMT 2022 Metrics Shared Task. Our primary submission – dubbed COMET-22 – is an ensemble between a COMET estimator model trained with Direct Assessments and a newly proposed multitask model trained to predict sentence-level scores along with OK/BAD word-level tags derived from Multidimensional Quality Metrics error annotations. These models are ensembled together using a hyper-parameter search that weights different features extracted from both evaluation models and combines them into a single score. For the reference-free evaluation we present COMETKIWI. Similarly to our primary submission, COMETKIWI is an ensemble between two models. A traditional predictor-estimator model inspired by OPENKIWI and our new multitask model trained on Multidimensional Quality Metrics which can also be used without references. Both our submissions show improved correlations compared to state-of-the-art metrics from last year as well as increased robustness to critical errors.

1 Introduction

Automatic metrics for Machine Translation (MT) are a fundamental component of MT research and development. While human evaluation is still of great importance, automatic metrics allow the rapid evaluation and comparison of MT systems on large collections of text and facilitate expansion to low resource languages and domains. Neural fine-tuned metrics in particular, have shown the ability to leverage large multilingual data during training to better compare and assess the quality of state-of-the-art MT models, outperforming traditional lexical-based metrics. Hence, our research is targeted to and guided by the advancements in these metrics.

This paper presents the joint contribution of Unbabel and Instituto Superior Técnico (IST) to the WMT 2022 Metrics Shared Task (Freitag et al., 2022). We participated in the segment-level and system-level tracks, as well as the "QE-as-a-Metric". Similar to our participation last year (Rei et al., 2021), our models are based on the COMET framework[1] (Rei et al., 2020a).

Our efforts this year built on findings and observations from our participation in the WMT 2021 Metrics Shared Task (Rei et al., 2021; Freitag et al., 2021b) to further improve COMET for the Metrics task and to increase its robustness to translation errors such as deviation from named entities, reverse polarity and negation, deviation in numbers, etc. These types of fine-grained critical errors have been shown to be challenging for state-of-the-art metrics and QE systems (Amrhein and Sennrich, 2022; Kanojia et al., 2021). For that reason we aim to take advantage of finer-grained information, incorporating word-level supervision from Multidimensional Quality Metrics (MQM) annotations when available. This approach is motivated by the observed improvements in performance in WMT 2021 Metrics when fine-tuning on MQM data. Additionally, the importance of word-level supervision as an auxiliary task was established via our participation in the WMT 2022 Quality Estimation task (Zerva et al., 2022), where we found that we get a boost of performance across language pairs when we combine word- and sentence-level targets (Rei et al., 2022).

Overall, our main contributions are:

- We propose a new model architecture that is trained with a multitask objective to predict a sentence-level score along with word-level tags. This architecture is well suited for MQM data which comes in the form of sentence-level scores

*✉ ricardo.rei@unbabel.com

[1]Code and models available at: https://github.com/Unbabel/COMET

Proceedings of the Seventh Conference on Machine Translation (WMT), pages 578 - 585

alongside the annotation spans. Also, similarly to UniTE (Wan et al., 2022) we can use this architecture with and without access to a reference translation.

- We show that ensembling scores from different models, trained with different annotations (e.g DA and MQM) can lead to improved correlations and more robust metrics.
- We corroborate our findings from last year (Rei et al., 2021) showing that reference-free evaluation is becoming competitive with reference-based evaluation.

Our submitted metrics, compared to two of the best submissions from last year, improve by a considerable margin in terms of correlations with MQM ($\approx 4\%$ in Kendall-Tau at segment level and $\approx 6\%$ on system-level accuracy) and in the ability of detecting critical errors ($\approx 30\%$ in accuracy on SMAUG challenge set (Alves et al., 2022), a newly proposed challenge set built to test the robustness of MT metrics to errors in named entities, numbers, meaning, inserted content and missing content.).

2 Corpora

Every year, since 2017, the organisers of WMT News translation tasks collect annotations in the form of Direct Assessments (DA) (Graham et al., 2013). Recently, Freitag et al. (2021a) showed that DA annotations, collected by non-professional crowd-source workers, are noisy and unfit to measure the quality of high performing MT systems. For high quality MT evaluation the authors suggest the use of MQM annotations performed by professionals; they released annotations for English→German and Chinese→English on the WMT 2020 translation outputs. Since then, along with the DA annotations coming from the News translation task, the metrics task organizers collect additional MQM data for English→German (en-de), Chinese→English (zh-en), and English→Russian (en-ru) to evaluate metrics against a more reliable ground-truth.

With that said, to test the performance of our new systems we will use the MQM annotations from 2021 News domain. For training we will use all DA ranging from 2017 to 2020 and the remaining MQM annotations.

One of the findings from last years shared task is that metrics struggle to accurately penalize translations with errors in reversing negation (Freitag et al., 2021b). Also, Amrhein and Sennrich (2022) showed that using COMET as a utility function for Minimum Bayes Risk decoding is more likely to lead to errors in named entities and numbers when compared to lexical metrics such as CHRF. In order to measure progress on capturing those errors we will test our new metrics on SMAUG (Alves et al., 2022).

3 Implemented Systems

Our goal this year is to build more robust **C**ross-lingual **O**ptimized **M**etrics for **E**valuation of **T**ranslations by ensembling systems that model different aspects of MT evaluation. For that purpose we used three different systems: a COMET Estimator (Rei et al., 2020a) trained on DA, a newly proposed Sequence Tagger, trained with MQM data, that works with and without references, and a COMETKIWI (Rei et al., 2022) model trained on DA.

3.1 COMET Estimator

For a more comprehensive description of the Estimator architecture we direct the reader to the original paper (Rei et al., 2020a). Compared to our COMET-DA model from last year (Rei et al., 2021) we only changed hyper-parameters in order to maximize Kendall-Tau correlations with the MQM annotations from 2021 News domain.

3.2 QE Predictor-Estimator Model

For a more comprehensive description of the implemented Predictor-Estimator architecture we direct the reader to our QE system description paper (Rei et al., 2022). In summary, for this year QE shared task we combine the strengths of COMET and OPENKIWI, leading to models that adopt COMET training features, useful for multilingual generalization, along with the predictor-estimator architecture of OPENKIWI.

3.3 Extending COMET for Sequence Tagging:

Following our experiments for the Quality Estimation shared task we implemented a multitask COMET model that is trained to perform sequence tagging along with sentence-level regression.

Inspired by UniTE (Wan et al., 2022), our model receives three inputs:

1. Source-only (src): machine translated sentence concatenated with its source sentence.

		zh-en		en-de		en-ru			
	Nº Segments	9750		8959		8432			
	Correlations	ρ	τ	ρ	τ	ρ	τ	Avg. ρ	Avg. τ
Baselines	BLEU	0.215	0.153	0.086	0.065	0.123	0.094	0.141	0.104
	CHRF	0.116	0.088	0.116	0.088	0.213	0.165	0.192	0.143
	BLEURT	0.456	0.331	0.309	0.236	0.345	0.267	0.370	0.278
	COMET-20	0.463	0.336	0.270	0.206	0.330	0.256	0.355	0.266
	COMET-21	0.513	0.377	0.309	0.237	0.345	0.263	0.389	0.292
Primary Sub.	COMET-22	0.537	0.395	**0.366**	**0.281**	**0.407**	**0.315**	**0.437**	**0.330**
	MQM Sequence Tagger								
	$\hookrightarrow \hat{y}_{\text{tags}}$	0.311	0.222	0.302	0.237	0.362	0.314	0.325	0.258
	$\hookrightarrow \hat{y}_{\text{src}}$	0.487	0.356	0.347	0.266	0.359	0.276	0.398	0.299
	$\hookrightarrow \hat{y}_{\text{ref}}$	0.535	0.394	0.358	0.275	0.386	0.297	0.427	0.322
	$\hookrightarrow \hat{y}_{\text{uni}}$	**0.538**	**0.396**	0.360	0.277	0.382	0.294	0.427	0.322
	DA Estimator	0.495	0.362	0.289	0.221	0.369	0.285	0.384	0.289
QE metric	COMETKIWI	0.471	0.343	0.348	0.266	0.366	0.283	0.395	0.297
	MQM Sequence Tagger								
	$\hookrightarrow \hat{y}_{\text{tags}}$	0.431	0.312	0.279	0.218	0.332	0.257	0.313	0.245
	$\hookrightarrow \hat{y}_{\text{src}}$	0.283	0.201	0.347	0.266	0.310	0.268	0.348	0.262
	DA Pred-Estimator	0.487	0.356	0.286	0.219	0.359	0.276	0.377	0.284

Table 1: Segment-level Spearman R (ρ) and Kendall-Tau (τ) correlations for `zh-en`, `en-de` and `en-ru` 2021 MQM annotations for the News Domain.

2. Reference-only (ref): machine translated sentence concatenated with its reference.

3. Unified input (uni): machine translated sentence concatenated with both source and reference.

These inputs can be seen as a sequence with two parts: 1) the machine translated sentence $\boldsymbol{t} = \langle t_1, ..., t_n \rangle$ and 2) additional support information such as source and/or reference $\hat{\boldsymbol{r}} = \langle r_1, ..., r_m \rangle$.

Given that, for each input, our model works exactly like COMETKIWI. We run three forward passes and we store the corresponding sentence-level scores $\hat{y}_{\text{src}}$, $\hat{y}_{\text{ref}}$ and $\hat{y}_{\text{uni}}$. Additionally, we average the obtained word-level logits to derive a single sequence S of $\{\text{OK}, \text{BAD}\}$ tags from which we compute an additional sentence score by using a similar formula to MQM:

$$\hat{y}_{\text{tags}} = 1 - \frac{w \times \sum_i^{N_S} \mathbb{1}[S_i = \text{BAD}]}{N_S} \quad (1)$$

where w is a severity penalty for BAD tags which we set to 1.

In sum, after running our new model we obtain 4 different scores: $\hat{y}_{\text{tags}}$, $\hat{y}_{\text{src}}$, $\hat{y}_{\text{ref}}$, $\hat{y}_{\text{uni}}$ which we can combine into a single quality score. Also, since this model is trained with a reference-less input it can be used, during inference, as a QE system. In those cases we run a single forward pass with the reference-less input and instead of 4 quality score we only get 2 ($\hat{y}_{\text{tags}}$ and $\hat{y}_{\text{src}}$).

Training Data. Since the MQM training data is not abundant and only covers 3 language pairs we start by training the above model without word-level information for 2 full epochs on DA ranging the shared task data from 2017 to 2020. Then, we fine-tune the model using the multitask setting described above with the available MQM training data for `zh-en`, `en-de` and `en-ru`.

3.4 Primary Submission

Our primary submission is an ensemble between a COMET Estimator model trained on top of XLM-R using DA from 2017 to 2020 and a sequence tagging model, such as the one described above, trained on top of InfoXLM (Chi et al., 2021). The final score is computed by a weighted average of the model outputs (5 scores), where the weights for each language pair were tuned with Optuna (Akiba et al., 2019)[2].

3.5 QE-as-a-metric Submission

Our primary submission is an ensemble between a COMETKIWI model trained on top of RemBERT (Chung et al., 2021) and the same sequence tagger from the primary submission but using a reference-less input during inference. The final score is computed by a weighted average in the

[2]We tuned weights specifically for the 3 MQM language pairs. For all other language pairs the weights were tuned by concatenating the MQM annotations for all three language pairs

same way as for our primary submission but using only the obtained 3 reference-less scores.

	Nº Systems	zh-en 15	en-de 17	en-ru 16	avg.
Baselines	BLEU	45.71	66.91	46.66	53.10
	CHRF	43.81	65.44	55.00	54.75
	BLEURT	48.57	83.09	70.83	67.50
	COMET-20	53.33	74.26	64.17	63.92
	COMET-21	53.33	78.68	70.83	67.61
Primary Sub.	COMET-22	64.76	**86.76**	70.83	74.12
	MQM Sequence Tagger				
	↪ $\hat{y}_{\text{tags}}$	60.95	80.88	**75.83**	72.55
	↪ $\hat{y}_{\text{src}}$	**72.38**	**86.76**	70.83	76.66
	↪ $\hat{y}_{\text{ref}}$	71.42	**86.76**	**75.83**	**78.00**
	↪ $\hat{y}_{\text{uni}}$	69.52	85.29	72.5	75.77
	DA Estimator	50.47	71.32	66.66	62.82
QE metric	COMETKIWI	68.57	86.02	70.83	75.14
	MQM Sequence Tagger				
	↪ $\hat{y}_{\text{tags}}$	47.61	**86.76**	59.16	64.51
	↪ $\hat{y}_{\text{src}}$	67.61	78.68	72.50	72.93
	DA Pred-Estimator	**72.38**	57.35	70.83	66.85

Table 2: System-level accuracy for `zh-en`, `en-de` and `en-ru` 2021 MQM annotations for the News Domain.

4 Experimental Results

As we have seen in Section 2, for our experiments we use WMT 2021 News MQM annotations from last year shared task (Freitag et al., 2021b) for testing our metrics. As for baselines we used lexical metrics such as CHRF (Popović, 2015) and BLEU (Papineni et al., 2002) and three state-of-the-art metrics: BLEURT (Sellam et al., 2020; Pu et al., 2021), COMET-20 (Rei et al., 2020b) and COMET-21 (Rei et al., 2021)[3].

4.1 Segment-level

Segment-level correlations for 2021 MQM annotations on the News domain are shown in Table 1. We used both Spearman R (ρ) and Kendall-Tau (τ) correlation metrics to evaluate our models.

From this table we can observe that some individual scores from the MQM Sequence Tagger already outperform state-of-the-art metrics such as BLEURT and (COMET-20/21). Also, our newly trained DA Estimator is able to outperform BLEURT achieving results close to COMET-21 without ever seeing MQM data. Finally, when ensembled together, we are able to improve correlations by $\approx 1\%$ for both reference-based and reference-free metrics.

4.2 System-level

System-level results for 2021 MQM annotations for the News domain are shown in Table 2. To evaluate how our metrics perform we used the pairwise accuracy proposed in (Kocmi et al., 2021), which simulates a real world scenario where we are interested in comparing two systems and deciding which one is better.

Similarly to segment-level results, from Table 2, we can observe that the accuracy of individual scores from the MQM Sequence Tagger outperform, on average, strong baselines such as BLEURT and COMET-20/21. Nonetheless, when ensembled together, these scores do not improve the overall accuracy which seems to be obtained by using the MQM Sequence Tagger with references only. Another interesting finding is that our QE submission (COMET-KIWI) achieves higher accuracy than our primary submission COMET-22. Also, depending on the language pair the best accuracy is either achieved by $\hat{y}_{\text{src}}$ or $\hat{y}_{\text{ref}}$ but not by $\hat{y}_{\text{uni}}$. This seems to indicate that the unified score is not learning to take the best out of the source and reference and that there might be a best way to combine these two signals.

4.3 Robustness to Critical Errors

The `SMAUG` challenge set was built to specifically test the robustness of metrics in capturing 5 different phenomena: deviation in Named Entities (NE), deviation in Numbers (NUM), deviation in meaning (MEAN), insertion of content (INS) and removal of content (DEL). The goal of this challenge set is to check if metrics correctly penalize an incorrect translation that was created by perturbing a reference. To do so, the perturbed translation (t) is scored using source (s) and reference (r) against an alternative reference ($\hat{r}$). The goal of a metric f is to score $\hat{r}$ above t ($f(s, \hat{r}, r) > f(s, t, r)$). To measure f's performance we will look at the accuracy over the entire challenge set for each phenomena:

$$Acc^P = \frac{\sum_i^{N_P} \mathbb{1}[f(s, \hat{r}, r) > f(s, t, r)]}{N_P} \quad (2)$$

where P denotes a phenomena and N_P the number of examples for that specific phenomena.

[3]For all neural fine-tuned metrics we used the checkpoints that were used as primary submissions for the WMT20 and WMT21 Metric tasks, more precisely, `BLEURT20`, `wmt20-comet-da` and `wmt21-comet-mqm`

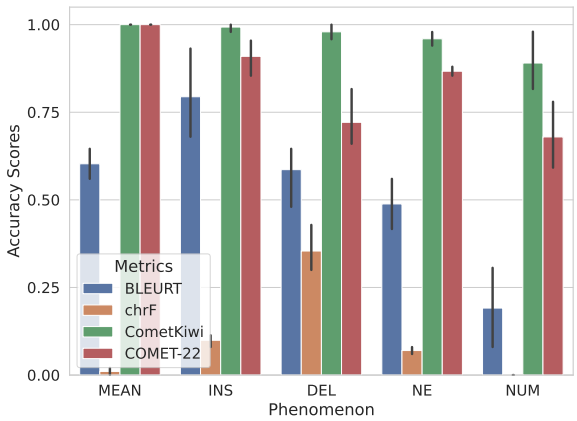

Figure 1: Accuracy Scores on the SMAUG Challenge Set for the baseline and submitted metrics.

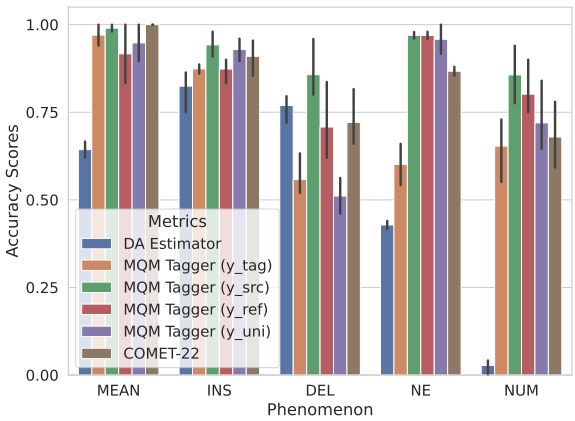

Figure 2: Accuracy Scores on the SMAUG Challenge Set for Primary Submission and respective individual scores.

Figure 1 presents the accuracy of our submissions against a lexical baseline (CHRF) and a learnt baseline (BLEURT) [4]. From these figure we can observe that our reference-free submission seems to be more robust than our primary submission which indicates that, when the reference is present, models look at lexical overlap and can be oblivious to critical errors that were derived from small perturbations. Also, we can observe that our submissions achieve a perfect accuracy on detecting deviations in meaning (which tests phenomena such as negation), above 0.65 accuracy in detecting wrong numbers and above 0.86 accuracy in detecting incorrect named entities. All of which were not correctly detected by previous state-of-the-art metrics such as BLEURT and COMET-20/21.

We also compare the performance of each ensemble with the individual models that compose it. In Figure 2, we observe that the DA Estimator has the worst overall performance. Also, the MQM Sequence Tagger $\hat{y}_{\text{src}}$ achieves the highest scores over all individual models, further suggesting that reference-free evaluation is more robust to these errors. Our final submission, while not reaching the highest accuracy for all phenomena, obtains good results in all cases. Regarding our QE-as-a-metric submission, Figure 3 shows that both the ensemble and individual systems achieve very high scores. Our submission outperforms the MQM Taggers and obtains a performance similar to the DA Predictor-Estimator.

[4]Performance of COMET-20 and 21 is similar to the performance shown by BLEURT while BLEU accuracy is close to 0.

5 Related work

For years, **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, recently, these classic metrics have been outperformed by metrics based on large pretrained models such as BERT (Bojar et al., 2017; Ma et al., 2018, 2019; Mathur et al., 2020).

Metrics based on large pretrained models can be divided into two categories: 1) **Embedding-distance metrics** and, 2) **Fine-tuned metrics**. **Embedding-distance metrics** replaced the typical word/n-gram matching by fuzzy matches based on dense representations. Examples of such metrics are BERTSCORE (Zhang et al., 2020) and YISI-1 (Lo, 2019), which has been a top performing metric since WMT 2019 Metrics task (Ma et al., 2019). Note that these metrics used the embedding models without any further fine-tuning relying only on their ability to capture semantic similarity. On the other hand, **fine-tuned metrics** such as RUSE (Shimanaka et al., 2018), BLEURT (Sellam et al., 2020) and COMET (Rei et al., 2020a) modify the underlying embedding models in order to learn how to produce quality scores such as DA and/or MQM, and thus to achieve higher correlations with human judgements of MT quality.

Recently, the evaluation of metrics has been extended to consider not only correlations with human judgements but also sensitivity to specific errors in translations. Namely, several works focused on translations with critical errors, which Specia et al. (2021) defines as translations that deviate in

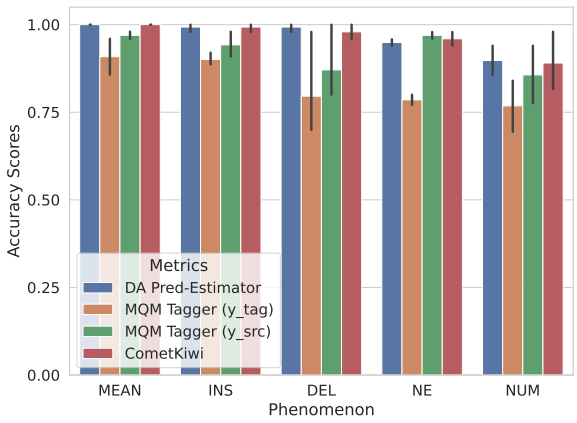

Figure 3: Accuracy Scores on the SMAUG Challenge Set for QE-as-a-metric Submission and respective individual scores.

meaning from their source in such way that they are misleading and can carry health, safety, legal, reputation, religious or financial implications. Amrhein and Sennrich (2022) show that COMET is less sensitive to errors in named entities and numbers than CHRF; Freitag et al. (2021b) found that several metrics struggle with negation and sentiment polarity errors; and Kanojia et al. (2021) showed that several reference-free metrics fail to detect errors related to omitting negation markers.

6 Conclusions

We present the joint contribution of Unbabel and IST to the WMT 2022 Metrics shared task. We propose a new architecture trained in a multitask setting which takes advantage of sentence-level scores along with supervision from MQM annotation spans. Inspired by UNITE, our new model can be used with and without references showing promising results when references are not available.

Our primary submission ensembles our new model along with the COMET Estimator architecture showing both higher correlations and improved robustness to phenomena that was deemed challenging in previous shared task editions.

Finally, our "QE-as-a-metric" submission yet again has shown that reference-free is competitive to reference-based evaluation not only at segment-level but also at system-level and in terms of detecting critical errors.

Acknowledgements

This work was supported by the P2020 programs (MAIA, contract 045909), by EU's Horizon Europe Research and Innovation Actions (UTTER, contract 101070631), by the European Research Council (ERC StG DeepSPIN 758969), and by the Fundação para a Ciência e Tecnologia through contract UIDB/50008/2020.

References

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Duarte M. Alves, Ricardo Rei, Ana C. Farinha, José G. C. de Souza, and André F. T. Martins. 2022. Robust MT evaluation with Sentence-level Multilingual data Augmentation. In *Proceedings of the Seventh Conference on Machine Translation*, Abu Dhabi. Association for Computational Linguistics.

Chantal Amrhein and Rico Sennrich. 2022. Identifying Weaknesses in Machine Translation Metrics Through Minimum Bayes Risk Decoding: A Case Study for COMET. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing*, Online. Association for Computational Linguistics.

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.

Zewen Chi, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-Ling Mao, Heyan Huang, and Ming Zhou. 2021. InfoXLM: An information-theoretic framework for cross-lingual language model pre-training. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3576–3588, Online. Association for Computational Linguistics.

Hyung Won Chung, Thibault Fevry, Henry Tsai, Melvin Johnson, and Sebastian Ruder. 2021. Rethinking Embedding Coupling in Pre-trained Language Models. In *International Conference on Learning Representations*.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021a. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André F. T. Martins. 2022. Results of WMT22 Metrics Shared Task: Stop Using BLEU – Neural Metrics Are Better and More Robust. In *Proceedings of the Seventh Conference on Machine Translation*, Abu Dhabi. Association for Computational Linguistics.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, George Foster, Alon Lavie, and Ondřej Bojar. 2021b. Results of the WMT21 metrics shared task: Evaluating metrics with expert-based human evaluations on TED and news domain. In *Proceedings of the Sixth Conference on Machine Translation*, pages 733–774, Online. Association for Computational Linguistics.

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.

Diptesh Kanojia, Marina Fomicheva, Tharindu Ranasinghe, Frédéric Blain, Constantin Orăsan, and Lucia Specia. 2021. Pushing the right buttons: Adversarial evaluation of quality estimation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 625–638, Online. 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.

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.

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, 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, 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.

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.

Amy Pu, Hyung Won Chung, Ankur Parikh, Sebastian Gehrmann, and Thibault Sellam. 2021. Learning compact metrics for MT. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 751–762, Online and Punta Cana, Dominican Republic. 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. 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.

Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C. Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte M. Alves, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. CometKiwi: IST-Unbabel 2022 Submission for the Quality Estimation Shared Task. In *Proceedings of the Seventh Conference on Machine Translation*, Abu Dhabi. 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.

Lucia Specia, Frédéric Blain, Marina Fomicheva, Chrysoula Zerva, Zhenhao Li, Vishrav Chaudhary, and André F. T. Martins. 2021. Findings of the WMT 2021 shared task on quality estimation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 684–725, Online. Association for Computational Linguistics.

Yu Wan, Dayiheng Liu, Baosong Yang, Haibo Zhang, Boxing Chen, Derek Wong, and Lidia Chao. 2022. UniTE: Unified translation evaluation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8117–8127, Dublin, Ireland. Association for Computational Linguistics.

Chrysoula Zerva, Frédéric Blain, Ricardo Rei, Piyawat Lertvittayakumjorn, José G. C. de Souza, Steffen Eger, Diptesh Kanojia, Duarte Alves, Constantin Orăsan, Marina Fomicheva, André F. T. Martins, and Lucia Specia. 2022. Findings of the WMT 2022 Shared Task on Quality Estimation. In *Proceedings of the Seventh Conference on Machine Translation*, Abu Dhabi. 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*.