The Inside Story: Towards Better Understanding of Machine Translation Neural Evaluation Metrics

Ricardo Rei[*1,2,4], **Nuno M. Guerreiro**[*3,4], **Marcos Treviso**[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

Neural metrics for machine translation evaluation, such as COMET, exhibit significant improvements in their correlation with human judgments compared to traditional metrics based on lexical overlap, such as BLEU. Yet neural metrics are, to a great extent, "black boxes" that return a single sentence-level score without transparency about the decision-making process. In this work, we develop and compare several neural explainability methods and demonstrate their effectiveness for interpreting state-of-the-art fine-tuned neural metrics. Our study reveals that these metrics leverage token-level information that can be directly attributed to translation errors, as assessed through comparison of token-level neural saliency maps with *Multidimensional Quality Metrics* (MQM) annotations and with synthetically-generated critical translation errors. To ease future research, we release our code at https://github.com/Unbabel/COMET/tree/explainable-metrics.

1 Introduction

Reference-based neural metrics for machine translation evaluation are achieving evergrowing success, demonstrating superior results over traditional lexical overlap-based metrics, such as BLEU (Papineni et al., 2002) and CHRF (Popović, 2015), in terms of both their correlation with human ratings and their robustness across diverse domains (Callison-Burch et al., 2006; Smith et al., 2016; Mathur et al., 2020; Kocmi et al., 2021; Freitag et al., 2022). However, lexical overlap-based metrics remain popular for evaluating the performance and progress of translation systems and algorithms. Concerns regarding trust and interpretability may help explain this (Leiter et al., 2022): contrary to traditional metrics, neural metrics are considered "black boxes" as they often use increasingly large models (e.g., the winning metric of the WMT 22 Metrics shared task was a 10B parameter model (Freitag et al., 2022)).

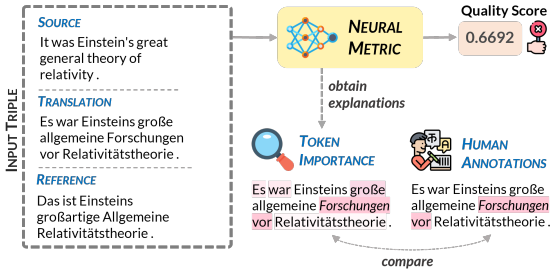


Figure 1: Illustration of our approach. In this example, the metric assigns the translation a low score. We aim to better understand this sentence-level assessment by examining the correspondence between our token-level explanations and human annotated error spans.

While some recent work has focus on explaining the predictions made by *reference-free* quality estimation (QE) systems (Fomicheva et al., 2021; Zerva et al., 2022), explaining *reference-based* metrics has remained a largely overlooked problem (Leiter et al., 2022). It is an open question whether the observations from studies of explainable QE carry over to this scenario. Thus, in this work, we fill that gap by turning to state-of-the-art reference-based metrics—we aim to interpret their decision-making process by exploiting the fact that these metrics show consistently good correlations with *Multidimentional Quality Metrics* (MQM) (Freitag et al., 2021, 2022; Sai et al., 2022), which are fine-grained quality assessments that result from experts identifying error spans in translation outputs (Lommel et al., 2014). We hypothesize that reference-based metrics leverage this token-level information to produce sentence-level scores. To test this hypothesis, we assess whether our explanations – measures of token-level importance obtained via attribution and input attribution methods such as attention weights and gradient scores (Treviso et al., 2021; Rei et al., 2022b) – align with

* Equal contribution. Corresponding author: ricardo.rei@unbabel.com

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 2: Short Papers, pages 1089–1105

human-annotated spans (Fomicheva et al., 2021, 2022; Zerva et al., 2022), as illustrated in Figure 1.

Our analysis focuses on two main vectors: (i) understanding the impact of the reference information on the quality of the explanations; and (ii) finding whether the explanations can help to identify potential weaknesses in the metrics. Our main contributions are:

- We provide a comparison between multiple explainability methods for different metrics on all types of evaluation: `src`-only, `ref`-only, and `src+ref` joint evaluation.
- We find that explanations are related to the underlying metric architecture, and that leveraging reference information improves the explanations.
- We show that explanations for critical translation errors can reveal weaknesses in the metrics.

2 Explaining Neural Metrics

We aim to explain sentence-level quality assessments of reference-based metrics by producing token-level explanations that align to translation errors. In what follows, we describe the metrics and how we produce the explanations that we study.

2.1 Metrics

We focus our analysis on two state-of-the-art neural metrics: COMET (Rei et al., 2020) and UNITE (Wan et al., 2022).[1] While both metrics use a multilingual encoder model based on XLM-R (Conneau et al., 2020), they employ distinct strategies to obtain sentence-level quality scores. On the one hand, COMET ***separately*** encodes the source, translation and reference to obtain their respective sentence embeddings; these embeddings are then combined to compute a quality score. On the other, UNITE ***jointly*** encodes the sentences to compute a contextualized representation that is subsequently used to compute the quality score. Interestingly, UNITE is trained to obtain quality scores for different input combinations: `[mt; src]` (SRC), `[mt; ref]` (REF), and `[mt; src; ref]` (SRC+REF). In fact, when the input is SRC, UNITE works like TransQuest (Ranasinghe et al., 2020); REF, like BLEURT (Sellam et al., 2020); and SRC+REF, like ROBLEURT (Wan et al., 2021).

[1]Ensembles composed of these two metrics were respectively ranked second and third in WMT 2022 Metrics shared task. The winner of WMT 2022 Metrics task — METRICXXL — is not publicly available (Freitag et al., 2022).

2.2 Explanations via Attribution Methods

In this work, we produce explanations using attribution methods that assign a scalar value to each translation token (i.e. a token-level attribution) to represent its importance. While many input attribution methods exist and have been extensively studied in the literature (Ribeiro et al., 2016; Shrikumar et al., 2017; Sundararajan et al., 2017; Jain and Wallace, 2019; Atanasova et al., 2020; Zaman and Belinkov, 2022), we focus specifically on those that have been demonstrated to be effective for explaining the predictions of QE models (Treviso et al., 2021; Fomicheva et al., 2022; Fernandes et al., 2022; Zerva et al., 2022) and extend them to our reference-based scenario. Concretely, we use the following techniques to extract explanations:[2]

- **embed–align:** the maximum cosine similarity between each translation token embedding and the reference and/or source token embeddings (Tao et al., 2022);
- **grad** ℓ_2: the ℓ_2-norm of gradients with respect to the word embeddings of the translation tokens (Arras et al., 2019);
- **attention**: the attention weights of the translation tokens for each attention head of the encoder (Treviso et al., 2021);
- **attn** $\times$ **grad**: the attention weights of each head scaled by the ℓ_2-norm of the gradients of the value vectors of that head (Rei et al., 2022b).

3 Experimental Setting

MQM annotations. We use MQM annotations from the WMT 2021 Metrics shared task (Freitag et al., 2021),[3] covering three language pairs — English-German (en→de), English-Russian (en→ru), and Chinese-English (zh→en) —in two different domains: News and TED Talks. For each incorrect translation, human experts marked the corresponding error spans. In our framework, these error spans should align with the words that the attribution methods assign higher importance to.

[2]For all attention-based methods, we ensemble the explanations from the top 5 heads as this has shown to improve performance consistently over selecting just the best head (Treviso et al., 2021; Rei et al., 2022b). Moreover, we use the full attention matrix, instead of relying only on cross attention information.

[3]https://github.com/google/wmt-mqm-human-evaluation

Metric	Explainability Method	en→de AUC	en→de R@K	zh→en AUC	zh→en R@K	en→ru AUC	en→ru R@K	Avg. AUC	Avg. R@K
				src-only evaluation*					
UniTE Src	embed–align[mt, src]	0.587	0.339	**0.644**	0.281	0.583	0.167	0.604	0.262
	grad ℓ_2	0.572	0.293	0.535	0.200	**0.620**	0.169	0.576	0.221
	attention	0.636	0.322	0.612	0.253	0.612	0.189	0.620	0.254
	attn × grad	**0.707**	**0.376**	0.639	**0.294**	0.633	**0.211**	**0.660**	**0.294**
				ref-only evaluation					
UniTE Ref	embed–align[mt, ref]	0.658	0.396	0.667	0.328	0.635	0.218	0.653	0.314
	grad ℓ_2	0.596	0.319	0.571	0.260	0.661	0.202	0.609	0.261
	attention	0.637	0.344	**0.670**	0.335	0.652	0.224	0.653	0.301
	attn × grad	**0.725**	**0.425**	0.667	**0.380**	**0.660**	**0.248**	**0.684**	**0.351**
				src,ref joint evaluation					
UniTE Src+Ref	embed–align[mt, src; ref]	0.650	0.383	0.670	0.330	0.618	0.213	0.646	0.309
	grad ℓ_2	0.595	0.325	0.579	0.257	0.643	0.191	0.606	0.257
	attention	0.657	**0.421**	0.670	**0.383**	0.649	0.223	0.659	0.342
	attn × grad	**0.736**	**0.421**	**0.674**	**0.383**	**0.671**	**0.248**	**0.693**	**0.351**
Comet	embed–align[mt, src]	0.590	0.371	0.674	0.314	0.577	0.220	0.614	0.301
	embed–align[mt, ref]	0.694	**0.425**	0.696	0.355	0.647	0.275	**0.679**	**0.352**
	embed–align[mt, src; ref]	0.688	0.416	**0.697**	**0.357**	0.622	**0.279**	0.669	0.350
	grad ℓ_2	0.603	0.312	0.540	0.252	0.604	0.185	0.582	0.250
	attention	0.604	0.351	0.592	0.259	0.633	0.209	0.608	0.268
	attn × grad	**0.710**	0.365	0.633	0.278	**0.662**	0.244	0.669	0.295

Table 1: AUC and Recall@K of explanations obtained via different attribution methods for Comet and UniTE models on the MQM data. *Although UniTE Src is a *src-only evaluation* metric, it was trained with reference information (Wan et al., 2022).

Models. For Comet, we use the latest publicly available model: `wmt22-comet-da` (Rei et al., 2022a).[4] For UniTE, we train our own model using the same data used to train Comet in order to have a comparable setup[5]. We provide full details (training data, correlations with human annotations, and hyperparameters) in Appendix A. Overall, the resulting reference-based UniTE models (Ref and Src+Ref) are on par with Comet.

Evaluation. We want our explanations to be directly attributed to the annotated error spans, in the style of an error detection task. Thus, we report Area Under Curve (AUC) and Recall@K.[6] These metrics have been used as the main metrics in previous works on explainable QE (Fomicheva et al., 2021, 2022; Zerva et al., 2022).

4 Results

4.1 High-level analysis

Explanations are tightly related to the underlying metric architecture. The results in Table 1 show that the predictive power of the attribution methods differ between UniTE and Comet: attn × grad is the best method for UniTE-based models, while embed–align works best for Comet.[7] This is expected as UniTE constructs a joint representation for the input sentences, thus allowing attention to flow across them; Comet, in contrast, encodes the sentences separately, so it relies heavily on the separate contextualized embeddings that are subsequently combined via element-wise operations such as multiplication and absolute difference. Interestingly, embed–align and attn × grad were the winning explainability approaches of the WMT 2022 Shared-Task on Quality Estimation (Zerva et al., 2022). This suggests that explainability methods developed for QE systems can translate well to reference-based metrics. We provide examples of explanations in Appendix C.

Reference information boosts explainability power. Table 1 also shows that, across all metrics, using reference information brings substantial improvements over using only the source information. Moreover, while reference-based attributions significantly outperform source-based attributions, combining the source and reference information to

[4] https://huggingface.co/Unbabel/wmt22-comet-da

[5] Our implementation differs from the original work by Wan et al. (2022), See Appendix A for full details.

[6] In this setup, Recall@K is the proportion of words with the highest attribution that correspond to translation errors against the total number of errors in the annotated error span.

[7] In Appendix B, we provide a comparison between the explanations obtained via embed–align with Comet and with its pretrained encoder model, XLM-R.

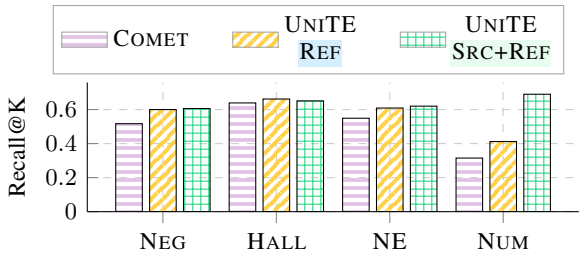


Figure 2: Performance of the best attribution methods for COMET, UNITE REF and UNITE SRC+REF in terms of Recall@K on translations with critical errors: negations (NEG), hallucinations (HALL), named entity errors (NE), and errors in numbers (NUM).

obtain token-level attributions does not consistently yield superior results over using the reference alone. Notably, the best attribution method for COMET does not require any source information. This is interesting: in some cases, reference-based metrics may largely ignore source information, relying heavily on the reference instead.

4.2 How do the explanations fare for critical translation errors?

The MQM data analyzed until now consists primarily of high quality translations, with the majority of annotated errors being non-critical. However, it is important to assess whether our explanations can be accurately attributed to critical errors, as this may reveal potential metric shortcomings. To this end, we employ SMAUG (Alves et al., 2022)[8], a tool designed to generate synthetic data for stress-testing metrics, to create corrupted translations that contain critical errors. Concretely, we generate translations with the following pathologies: negation errors, hallucinations via insertions, named entity errors, and errors in numbers.[9]

Explanations identify critical errors more easily than non-critical errors. Figure 2 shows that explanations are more effective in identifying critical errors compared to other non-critical errors (see Table 1). Specifically, we find significant performance improvements up to nearly 30% in Recall@K for certain critical errors. Overall, hallucinations are the easiest errors to identify across all neural metrics. This suggests that neural metrics appropriately identify and penalize hallucinated translations, which aligns with the findings of Guerreiro et al. (2022). Moreover, explanations for both UNITE models behave similarly for all errors except numbers, where the source information plays a key role in improving the explanations. Notably, contrary to what we observed for data with non-critical errors, COMET explanations are less effective than those of UNITE REF and UNITE SRC+REF for identifying critical errors.

Explanations can reveal potential metric weaknesses. Figure 2 suggests that COMET explanations struggle to identify localized errors (negation errors, named entity errors and discrepancies in numbers). We hypothesize that this behavior is related to the underlying architecture. Unlike UNITE-based metrics, COMET does not rely on soft alignments via attention between the sentences in the encoding process. This process may be key to identify local misalignments during the encoding process. In fact, the attention-based attributions for UNITE metrics can more easily identify these errors. COMET, however, encodes the sentences separately, which may result in grammatical features (e.g. numbers) being encoded similarly across sentences (Chi et al., 2020; Chang et al., 2022). As such, explanations obtained via embedding alignments will not properly identify these misalignments on similar features. Importantly, these findings align with observations made in (Amrhein and Sennrich, 2022; Raunak et al., 2022). This showcases how explanations can be used to diagnose and reveal shortcomings of neural-based metrics.

5 Conclusions and Future Work

In this paper, we investigated the use of explainability methods to better understand widely used neural metrics for machine translation evaluation, such as COMET and UNITE. Concretely, we analyzed how explanations are impacted by the reference information, and how they can be used to reveal weaknesses of these metrics. Our analysis shows that the quality of the explanations is tightly related to the underlying metric architecture. Interestingly, we also provide evidence that neural metrics like COMET may heavily rely on reference information over source information. Additionally, we show that explanations can be used to reveal reference-based metrics weaknesses such as failing to severely penalize localized critical errors. This opens up promising opportunities for future

[8] https://github.com/Unbabel/smaug

[9] We corrupt fully correct translations that are not an exact copy of the reference translation. Moreover, as the full suit of SMAUG transformations can only be applied to English data, we focus solely on zh→en translations. Overall, the synthetic dataset consists of 2610 translations. Full statistics about the corrupted data and examples are shown in Appendix A.2.

research on leveraging explanations to diagnose reference-based metrics errors. To support these studies, we call for future datasets illustrating critical errors (e.g., challenge sets (Karpinska et al., 2022)) to be accompanied by annotated error spans.

Limitations

We highlight three main limitations of our work.

First, although we have explored gradient-based explanations that take the whole network into consideration and have been shown to be faithful in previous work (Bastings et al., 2021), we do not explicitly explore how COMET combines the sentence representations in the feed-forward that precedes the encoder model to produce the sentence-level score.

Second, we have shown that combining attention with gradient information results in the best explanations for UNITE-based metrics. However, from a practical standpoint, running inference and extracting the explainability scores simultaneously may be more computationally expensive than other techniques: gradient-based metrics benefit from GPU infrastructure and require storing all gradient information.

Third, we have not explored extracting explanations in low-resource settings. That is because high-quality MQM annotations for such language pairs are not yet available. Nevertheless, further research in those settings is needed to access the broader validity of our claims.

Acknowledgements

This work was partially supported by the P2020 programs (MAIA, contract 045909), the Portuguese Recovery and Resilience Plan (PRR) through project C645008882-00000055, Center for Responsible AI, by the European Research Council (ERC StG DeepSPIN, 758969), by EU's Horizon Europe Research and Innovation Actions (UTTER, contract 101070631), and by the Fundação para a Ciência e Tecnologia (contracts UIDB/50021/2020 and UIDB/50008/2020).

References

Duarte Alves, Ricardo Rei, Ana C Farinha, José G. C. de Souza, and André F. T. Martins. 2022. Robust MT Evaluation with Sentence-level Multilingual Augmentation. In *Proceedings of the Seventh Conference on Machine Translation*, pages 469–478, 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 (Volume 1: Long Papers)*, pages 1125–1141, Online only. Association for Computational Linguistics.

Leila Arras, Ahmed Osman, Klaus-Robert Müller, and Wojciech Samek. 2019. Evaluating recurrent neural network explanations. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 113–126, Florence, Italy. Association for Computational Linguistics.

Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. 2020. A diagnostic study of explainability techniques for text classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3256–3274, Online. Association for Computational Linguistics.

Jasmijn Bastings, Sebastian Ebert, Polina Zablotskaia, Anders Sandholm, and Katja Filippova. 2021. "will you find these shortcuts?" a protocol for evaluating the faithfulness of input salience methods for text classification.

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Re-evaluating the role of Bleu in machine translation research. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 249–256, Trento, Italy. Association for Computational Linguistics.

Tyler A. Chang, Zhuowen Tu, and Benjamin K. Bergen. 2022. The geometry of multilingual language model representations.

Ethan A. Chi, John Hewitt, and Christopher D. Manning. 2020. Finding universal grammatical relations in multilingual BERT. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5564–5577, Online. 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.

Daniel Deutsch, Rotem Dror, and Dan Roth. 2021. A statistical analysis of summarization evaluation metrics using resampling methods. *Transactions of the Association for Computational Linguistics*, 9:1132–1146.

Patrick Fernandes, Marcos Treviso, Danish Pruthi, André F. T. Martins, and Graham Neubig. 2022. Learning to scaffold: Optimizing model explanations for teaching.

Marina Fomicheva, Piyawat Lertvittayakumjorn, Wei Zhao, Steffen Eger, and Yang Gao. 2021. The Eval4NLP shared task on explainable quality estimation: Overview and results. In *Proceedings of the 2nd Workshop on Evaluation and Comparison of NLP Systems*, pages 165–178, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Marina Fomicheva, Lucia Specia, and Nikolaos Aletras. 2022. Translation error detection as rationale extraction. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 4148–4159, Dublin, Ireland. Association for Computational Linguistics.

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*, pages 46–68, Abu Dhabi. Association for Computational Linguistics.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, George Foster, Alon Lavie, and Ondřej Bojar. 2021. 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.

Nuno M. Guerreiro, Elena Voita, and André F. T. Martins. 2022. Looking for a needle in a haystack: A comprehensive study of hallucinations in neural machine translation.

Sarthak Jain and Byron C. Wallace. 2019. Attention is not Explanation. 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 3543–3556, Minneapolis, Minnesota. Association for Computational Linguistics.

Marzena Karpinska, Nishant Raj, Katherine Thai, Yixiao Song, Ankita Gupta, and Mohit Iyyer. 2022. Demetr: Diagnosing evaluation metrics for translation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, page 9540–9561, Abu Dhabi, United Arab Emirates. 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.

Christoph Leiter, Piyawat Lertvittayakumjorn, Marina Fomicheva, Wei Zhao, Yang Gao, and Steffen Eger. 2022. Towards explainable evaluation metrics for natural language generation.

Arle Lommel, Hans Uszkoreit, and Aljoscha Burchardt. 2014. Multidimensional Quality Metrics (MQM) : A Framework for Declaring and Describing Translation Quality Metrics. *Tradumàtica*, pages 0455–463.

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.

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.

Tharindu Ranasinghe, Constantin Orasan, and Ruslan Mitkov. 2020. TransQuest: Translation Quality Estimation with Cross-lingual Transformers. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5070–5081, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Vikas Raunak, Matt Post, and Arul Menezes. 2022. Salted: A framework for salient long-tail translation error detection.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022a. COMET-22: Unbabel-IST 2022 Submission for the Metrics Shared Task. In *Proceedings of the Seventh Conference on Machine Translation*, pages 578–585, Abu Dhabi. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte

Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. 2022b. CometKiwi: IST-Unbabel 2022 Submission for the Quality Estimation Shared Task. In *Proceedings of the Seventh Conference on Machine Translation*, pages 634–645, Abu Dhabi. Association for Computational Linguistics.

Marco Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 97–101, San Diego, California. Association for Computational Linguistics.

Ananya B. Sai, Vignesh Nagarajan, Tanay Dixit, Raj Dabre, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh M. Khapra. 2022. IndicMT Eval: A Dataset to Meta-Evaluate Machine Translation metrics for Indian Languages.

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.

Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. 2017. Learning Important Features Through Propagating Activation Differences. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3145–3153. PMLR.

Aaron Smith, Christian Hardmeier, and Joerg Tiedemann. 2016. Climbing mont BLEU: The strange world of reachable high-BLEU translations. In *Proceedings of the 19th Annual Conference of the European Association for Machine Translation*, pages 269–281.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic Attribution for Deep Networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3319–3328. PMLR.

Shimin Tao, Su Chang, Ma Miaomiao, Hao Yang, Xiang Geng, Shujian Huang, Min Zhang, Jiaxin Guo, Minghan Wang, and Yinglu Li. 2022. CrossQE: HW-TSC 2022 Submission for the Quality Estimation Shared Task. In *Proceedings of the Seventh Conference on Machine Translation*, pages 646–652, Abu Dhabi. Association for Computational Linguistics.

Marcos Treviso, Nuno M. Guerreiro, Ricardo Rei, and André F. T. Martins. 2021. IST-unbabel 2021 submission for the explainable quality estimation shared task. In *Proceedings of the 2nd Workshop on Evaluation and Comparison of NLP Systems*, pages 133–145, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yu Wan, Dayiheng Liu, Baosong Yang, Tianchi Bi, Haibo Zhang, Boxing Chen, Weihua Luo, Derek F. Wong, and Lidia S. Chao. 2021. RoBLEURT submission for WMT2021 metrics task. In *Proceedings of the Sixth Conference on Machine Translation*, pages 1053–1058, 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.

Kerem Zaman and Yonatan Belinkov. 2022. A Multilingual Perspective Towards the Evaluation of Attribution Methods in Natural Language Inference.

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*, pages 69–99, Abu Dhabi. Association for Computational Linguistics.

A Model Details

In Section 2.1, we employed the latest publicly available model (`wmt22-comet-da`) for COMET, which emerged as a top-performing metric in the WMT 2022 Metrics task (Freitag et al., 2022). To ensure a comparable setting for UNITE (Wan et al., 2022), we trained our own model. In doing so, we utilized the same data employed in the development of the COMET model by (Rei et al., 2022a), without pretraining any synthetic data, as originally suggested. Additionally, our implementation did not incorporate monotonic regional attention, as our preliminary experiments revealed no discernible benefits from its usage. The hyperparameters used are summarized in Table 3, while Table 4 presents the number of Direct Assessments utilized during training. Furthermore, Table 5 displays the segment-level correlations with WMT 2021 MQM data for the News and TED domains.

Regarding computational infrastructure, a single NVIDIA A10G GPU with 23GB memory was used. The resulting UNITE model has 565M parameters while COMET has 581M parameters.

A.1 Output Distribution

To better understand the output of the models and what scores are deemed low, we plotted the output distributions for the two models we used in our study. The average score for English→German data is 0.856 for the COMET model and 0.870 for the UNITE model we trained. From Figure 3 we can observe the distribution of scores. This means that the 0.6692 score from the example in Figure 1 corresponds to a low quality output (5th percentile).

A.2 SMAUG Corpus

As we have seen in Section 4.2, we have created synthetic translation errors for the following pathologies: negation errors, hallucinations via insertions, named entity errors, and errors in numbers. Table 7 presents a summary of the examples created using SMAUG and in Table 8 we show examples of each error category.

B Comparison between COMET and XLM-R Alignments

From Table 1, it is evident that the alignments between the reference and/or source and the translation yield effective explanations for COMET. This raises the question of how these alignments compare to the underlying encoder of COMET before the fine-tuning process with human annotations. To investigate this, we examine the results for XLM-R without any fine-tuning, as presented in Table 2.

Overall, the explanations derived from the alignments of COMET prove to be more predictive of error spans than those obtained from XLM-R alignments. This suggests that during the fine-tuning phase, COMET models modify the underlying XLM-R representations to achieve better alignment with translation errors.

C Examples

In Tables 9 and 10, we show examples of COMET explanations for Chinese→English and English→German language pairs, respectively. We highlight in gray the corresponding MQM annotation performed by an expert linguist and we sort the examples from highest to lowest COMET scores. From these examples we can observe the following:

- Highlights provided by COMET explanations have a high recall with human annotations. In all examples, subword tokens corresponding to translation errors are highlighted in red but we often see that not everything is incorrect.

- Explanations are consistent with scores. For example, in the third example from Table 10, the red highlights do not correspond to errors and in fact the translation only has a major error griffen. Nonetheless, the score assigned by COMET is a low score of 0.68 which is faithful to the explanations that was given even if the assessment does not agree with human experts.

METRIC	EXPLAINABILITY METHOD	en→de AUC	en→de R@K	zh→en AUC	zh→en R@K	en→ru AUC	en→ru R@K	Avg. AUC	Avg. R@K
XLM-R	embed–align$^{[mt, src]}$	0.587	0.359	0.668	0.311	0.576	0.199	0.610	0.289
	embed–align$^{[mt, ref]}$	0.671	0.405	0.689	0.345	0.634	0.244	0.664	0.331
	embed–align$^{[mt, src; ref]}$	0.666	0.395	0.690	0.347	0.616	0.242	0.657	0.328
COMET	embed–align$^{[mt, src]}$	0.590	0.371	0.674	0.314	0.577	0.220	0.614	0.301
	embed–align$^{[mt, ref]}$	**0.694**	**0.425**	0.696	0.355	**0.647**	0.275	**0.679**	**0.352**
	embed–align$^{[mt, src; ref]}$	0.688	0.416	**0.697**	**0.357**	0.622	**0.279**	0.669	0.350

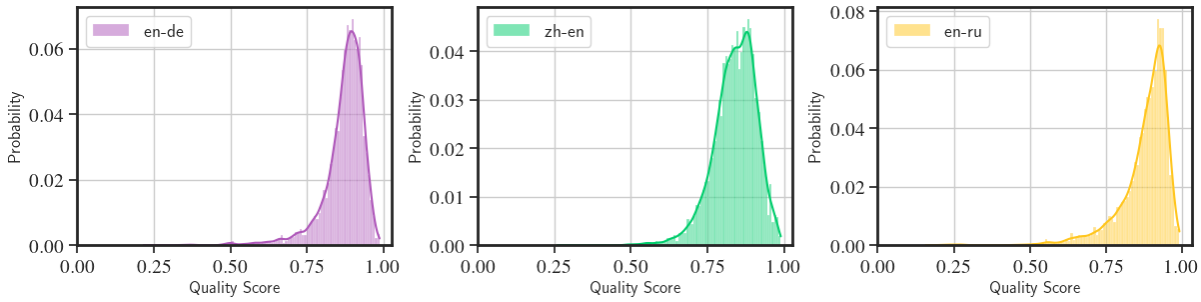
Table 2: AUC and Recall@K of explanations obtained via alignments for COMET and XLM-R without any further fine-tuning on human annotations.

Hyperparameter	UNITE	COMET
Encoder Model	XLM-R (large)	
Optimizer	AdamW	
No. frozen epochs	0.3	
Learning rate (LR)	1.5e-05	
Encoder LR.	1.0e-06	
Layerwise Decay	0.95	
Batch size	16	
Loss function	MSE	
Dropout	0.1	
Hidden sizes	[3072, 1024]	
Embedding layer	Frozen	
FP precision	16	
No. Epochs	1	2

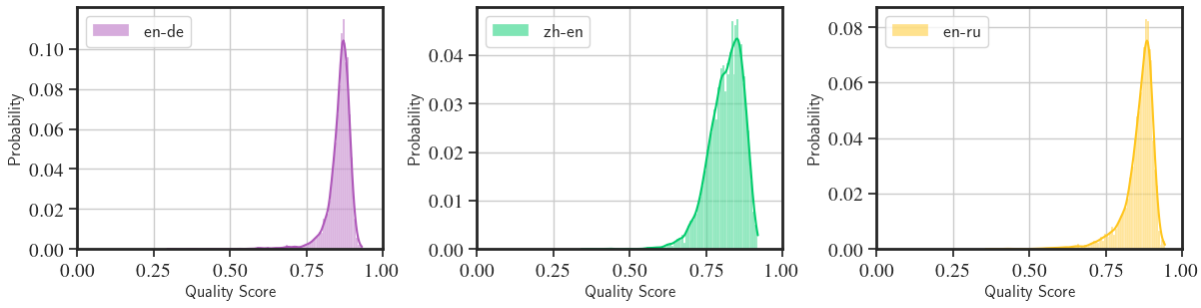
Table 3: Hyperparameters used to train UNITE and COMET checkpoints used in this work. The only difference between the two is the number of training epochs due to the fact that, for UNITE, the best validation checkpoint is the first one.

Language Pair	SIZE
zh-en	126947
en-de	121420
de-en	99183
en-zh	90805
ru-en	79280
en-ru	62749
en-cs	60937
fi-en	46145
en-fi	34335
tr-en	30186
et-en	29496
cs-en	27847
en-mr	26000
de-cs	13804
en-et	13376
pl-en	11816
en-pl	10572
lt-en	10315
en-ja	9578
gu-en	9063
si-en	9000
ro-en	9000
ne-en	9000
en-lt	8959
ja-en	8939
en-kk	8219
en-ta	7890
ta-en	7577
en-gu	6924
kk-en	6789
de-fr	6691
en-lv	5810
en-tr	5171
km-en	4722
ps-en	4611
fr-de	3999
Total	1027155

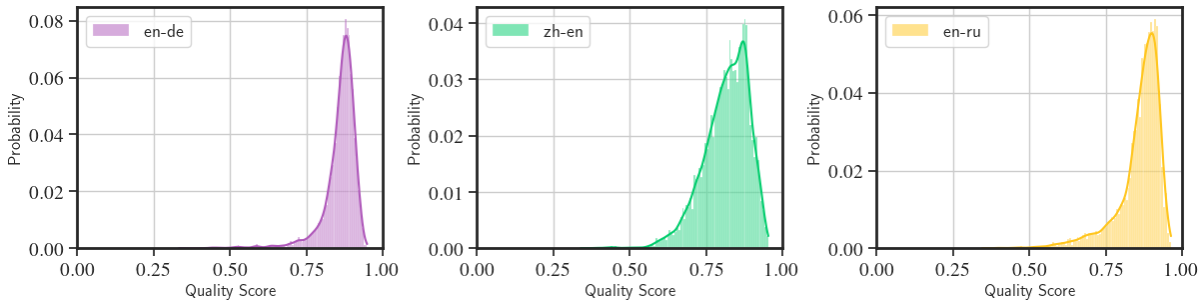
Table 4: Number of direct assessments per language pair used to train COMET (Rei et al., 2022a) and the UNITE model used in this work.



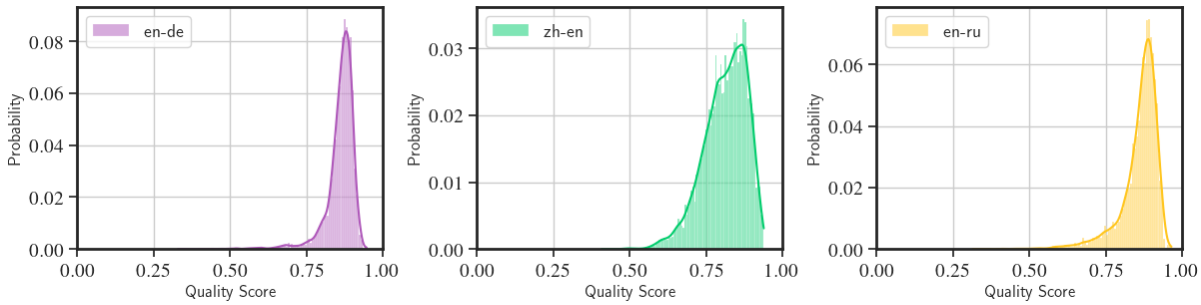
(a) COMET



(b) UNITE SRC



(c) UNITE REF



(d) UNITE SRC+REF

Figure 3: Distribution of scores for all metrics obtained on the MQM data (for all language pairs).

			BLEU	CHRF	YISI-1	BLEURT	UNITE SRC	UNITE REF	UNITE SRC+REF	COMET wmt22-comet-da
EN→DE	News	ρ	0.077	0.092	0.163	0.307	0.274	**0.321**	0.304	0.297
		τ	0.069	0.092	0.144	0.240	0.222	**0.248**	0.241	0.232
	TED	ρ	0.151	0.158	0.236	**0.325**	0.311	**0.335**	**0.338**	**0.329**
		τ	0.113	0.146	0.212	0.283	0.264	**0.301**	**0.298**	0.278
EN→RU	News	ρ	0.153	0.252	0.263	0.359	0.333	**0.391**	**0.382**	0.363
		τ	0.106	0.178	0.216	0.276	0.276	**0.298**	0.297	0.293
	TED	ρ	0.154	0.268	0.235	0.286	0.239	0.289	**0.318**	**0.308**
		τ	0.112	0.189	0.204	0.255	0.232	**0.262**	**0.264**	**0.268**
ZH→EN	News	ρ	0.215	0.231	0.301	0.428	0.413	0.438	0.426	**0.445**
		τ	0.165	0.188	0.289	0.341	0.331	0.358	0.352	**0.371**
	TED	ρ	0.155	0.181	0.287	0.295	0.244	0.301	**0.310**	**0.307**
		τ	0.113	0.144	0.216	0.246	0.224	0.265	**0.266**	**0.269**

Table 5: Segment-level correlations for WMT 2021 MQM annotations over News and TED domains (Freitag et al., 2021). The metrics are Pearson (ρ) and Kendall Tau (τ). Results in bold indicate which metrics are top-performing for that specific language pair, domain and metric according to Perm-Both hypothesis test (Deutsch et al., 2021), using 500 re-sampling runs, and setting $p = 0.05$.

Error Type	NUM EXAMPLES
NE	978
NEG	669
HALL	530
NUM	432
Total	2609

Table 6: Number of examples for each category, synthetically-created using SMAUG (Alves et al., 2022).

Language Pair	TOKENS / SENT.	ERRORS / SPANS
en-de	528704 / 15310	25712 / 3567
en-ru	525938 / 15074	17620 / 7172
zh-en	603258 / 16506	43984 / 10042

Table 7: Statistics about MQM data from WMT 2021 Metrics task (Freitag et al., 2021) used in our experiments.

Source:
格里沃里表示，分析人士对越南所提出的和平倡议给予认可。

Translation:
Grivory said that analysts recognize the peace initiative proposed by Vietnam.

Reference:
Grigory said that analysts endorse the peace initiative proposed by Vietnam.

NE Error:
Grivory said that analysts recognize the peace initiative proposed by Russia.

Source:
英国的这一决定预计将会使西班牙的旅游业大受影响。

Translation:
This decision by the United Kingdom is expected to greatly affect Spain's tourism industry.

Reference:
This decision by the UK is expected to have a significant impact on tourism in Spain.

NEG Error:
This decision by the United Kingdom is expected to greatly benefit Spain's tourism industry.

Source:
由于疫情，人们开始在互联网上花费更多的时间。”

Translation:
Because of the epidemic, people are starting to spend more time on the Internet."

Reference:
For reason of the pandemic, people are starting to spend more time on the Internet. ”

HALL Error:
Because we have a lot of friends around during the epidemic, people are starting to spend more time on the mobile devices than on the Internet."

Source:
展销区将展至7月29日。

Translation:
The exhibition and sales area will be open until July 29.

Reference:
The exhibition will last until July 29.

NUM Error:
The exhibition and sales area will be open until July 2018

Table 8: Synthetically-generated critical errors (highlighted in gray) created with SMAUG (Alves et al., 2022) to assess whether our explanations can be accurately attributed to critical errors.

Source:
And yet, the universe is not a silent movie because the universe isn't silent.

Translation:
Und dennoch ist das Universum kein Stummfilm, weil das Universum nicht still ist.

COMET score: 0.8595
COMET explanation:
_Und _dennoch _ist _das _Univers um _kein _Stu mm film , _weil _das _Univers um _nicht _still _ist .

Source:
And yet black holes may be heard even if they're not seen, and that's because they bang on space-time like a drum.

Translation:
Und dennoch werden Schwarze Löcher vielleicht gehört, auch wenn sie nicht gesehen werden, und das liegt daran, dass sie wie eine Trommel auf die Raumzeit schlagen.

COMET score: 0.7150
COMET explanation:
_Und _dennoch _werden _Schwarz e _Lö cher _vielleicht _gehört , _auch _wenn _sie _nicht _gesehen _werden , _und _das _liegt _daran , _dass _sie _wie _eine _Tro mmel _auf _die _Raum zeit schlagen .

Source:
Ash O'Brien and husband Jarett Kelley say they were grabbing a bite to eat at Dusty Rhodes dog park in San Diego on Thursday, with their three-month-old pug in tow.

Translation:
Ash O'Brien und Ehemann Jarett Kelley sagen, dass sie am Donnerstag im Hundepark Dusty Rhodes in San Diego einen Happen zu essen griffen, mit ihrem drei Monate alten Mops im Schlepptau.

COMET score: 0.6835
COMET explanation:
_Ash _O ' Bri en _und _Ehe mann _Ja rett _Kel ley _sagen , _dass _sie _am _Donnerstag _im _Hunde park _Du sty _Rhod es _in _San _Diego _einen _Happ en _zu _essen _ griff en _ , _mit _ihrem _drei _Monate _alten _M ops _im _Schle ppt au .

Source:
It was Einstein's great general theory of relativity.

Translation:
Es war Einsteins große allgemeine Forschungen vor Relativitätstheorie.

COMET score: 0.6692
COMET explanation:
_Es _war _Einstein s _große _allgemein e _Forschung en _vor _Relativ ität s the ori e .

Source:
There's mask-shaming and then there's full on assault.

Translation:
Es gibt Maskenschämen und dann ist es voll bei Angriff.

COMET score: 0.2318
COMET explanation:
_Es _gibt _Mask en schä men _und _dann _ist _es _voll _bei _Angriff _ .

Table 9: Saliency map for COMET explanation scores for a set of en→de examples. Comparing the token-level explanations with the MQM annotation (highlighted in gray) reveals the source of correspondence between specific token-level translation errors and the resulting scores.

Source:
我想告诉大家 宇宙有着自己的配乐， 而宇宙自身正在不停地播放着。 因为太空可以想鼓一样振动。

Translation:
I want to tell you that the universe has its own iconic soundtrack and the universe itself is constantly playing non-stop because space can vibrate like a drum.

COMET score: 0.8634
COMET explanation:
_I _want _to _tell _you _that _the _univers e _has _its _own _icon ic _soundtrack _and _the _univers e _itself _is _constantly _playing _non - stop _because _space _can _vibra te _like _a _drum .

Source:
另外,吉克隽逸和刘烨作为运动助理,也围绕运动少年制造了不少爆笑话题。

Translation:
In addition, as sports assistants, Ji Kejunyi and Liu Ye have also created a lot of hilarious topics around sports teenagers.

COMET score: 0.8214
COMET explanation:
_In _addition , _as _sports _assistant s , _Ji _Ke ju nyi _and _Li u _Ye _have _also _created _a _lot _of _ hila rious _topic s _around _sports _teenager s .

Source:
一番言论让场上的少年和运动领队们都倒吸一口凉气。

Translation:
The remarks made the teenagers and the sports leaders on the field gasp a sigh of relief.

COMET score: 0.7793
COMET explanation:
_The _re marks _made _the _teenager s _and _the _sports _leaders _on _the _field _gas p _a _sig h _of _relief _ .

Source:
强烈的阳光是如此地刺眼，

Translation:
The intense sunlight is so harsh;

COMET score: 0.7561
COMET explanation:
_The _intense _sun light _is _so _har sh ;

Source:
如今，我们希望能够 给这部关于宇宙的 宏伟的视觉作品 配上声音。

Translation:
Today, we hope to be able to give this magnificent visual work of the universe a sound.

COMET score: 0.7073
COMET explanation:
_Today , _we _hope _to _be _able _to _give _this _magnific ent _visual _work _of _the _univers e _a _sound .

Table 10: Saliency map for COMET explanation scores for a set of zh→en examples. Comparing the token-level explanations with the MQM annotation (highlighted in gray) reveals the source of correspondence between specific token-level translation errors and the resulting scores.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Yes. Section 6

☐ A2. Did you discuss any potential risks of your work?
Not applicable. Left blank.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Abstract and Section 1

☑ A4. Have you used AI writing assistants when working on this paper?
Assistance purely with the language of the paper along every section. Grammarly and DeepL write

B ☑ Did you use or create scientific artifacts?

Section 3 explains the methods we used. We will release the adaptations required to use the explainability methods over COMET framework, the UniTE model we trained, and all data synthetically-generated data.

☑ B1. Did you cite the creators of artifacts you used?
Section 2

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
footnote on the first page. The License will be Apache 2.0

☐ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Not applicable. Left blank.

☐ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Not applicable. Left blank.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
in the Appendix we have several statistics for training and testing data.

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Left blank.

C ☑ Did you run computational experiments?

Appendix provides detailed information about the trained model.

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Appendix provides detailed information about the trained model including GPU infrastructure and total number of parameters.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☐ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Not applicable. Left blank.

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Appendix has all information needed about test data and performance of the models.

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Section 2 and Appendix

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Not applicable. Left blank.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Not applicable. Left blank.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Not applicable. Left blank.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Not applicable. Left blank.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Not applicable. Left blank.