The Eval4NLP Shared Task on Explainable Quality Estimation: Overview and Results

Marina Fomicheva[*], **Piyawat Lertvittayakumjorn**[†], **Wei Zhao**[‡], **Steffen Eger**[‡], **Yang Gao**[◇]
∗ University of Sheffield, UK † Imperial College London, UK
‡ TU Darmstadt, Germany ◇ Royal Holloway, University of London, UK
m.fomicheva@sheffield.ac.uk pl1515@imperial.ac.uk
wei.zhao@h-its.org eger@aiphes.tu-darmstadt.de
yang.gao@rhul.ac.uk

Abstract

In this paper, we introduce the Eval4NLP-2021 shared task on explainable quality estimation. Given a source-translation pair, this shared task requires not only to provide a sentence-level score indicating the overall quality of the translation, but also to *explain* this score by identifying the words that negatively impact translation quality. We present the data, annotation guidelines and evaluation setup of the shared task, describe the six participating systems, and analyze the results. To the best of our knowledge, this is the first shared task on explainable NLP evaluation metrics. Datasets and results are available at https://github.com/eval4nlp/SharedTask2021.

1 Introduction

Recent Natural Language Processing (NLP) systems based on pre-trained representations from Transformer language models, such as BERT (Devlin et al., 2019) and XLM-Roberta (Conneau et al., 2020), have achieved outstanding results in a variety of tasks. This boost in performance, however, comes at the cost of efficiency and interpretability. Interpretability is a major concern in modern Artificial Intelligence (AI) and NLP research (Doshi-Velez and Kim, 2017; Danilevsky et al., 2020), as black-box models undermine users' trust in new technologies (Mercado et al., 2016; Toreini et al., 2020).

In the Eval4NLP 2021 shared task, we focus on evaluating machine translation (MT) as an example of this problem. Specifically, we look at the task of *quality estimation* (QE), where the aim is to predict the quality of MT output at inference time without access to reference translations (Blatz et al., 2004; Specia et al., 2018b).[1] Translation quality can be assessed at different levels of granularity: *sentence-level*, i.e. predicting the overall quality of translated sentences, and *word-level*, i.e. highlighting specific errors in the MT output. Those have traditionally been treated as two separate tasks, each one requiring dedicated training data.

In this shared task, we propose to address word-level translation error identification as an explainability task.[2] Explainability is a broad area aimed at explaining predictions of machine learning models. Rationale extraction methods achieve this by selecting a portion of the input that justifies model output for a given data point (Lei et al., 2016; Jain et al., 2020). A natural way to explain sentence-level quality assessment is to identify translation errors. Hence, we frame **error identification as a task of providing explanations for the predictions of sentence-level QE models**. We claim that this task represents a challenging new benchmark for testing explainability for NLP and provides a new way of addressing word-level QE.

On the one hand, QE is different from other explainable NLP tasks with existing datasets (DeYoung et al., 2020) in various important aspects. First, it is a regression task, as opposed to binary or multiclass text classification explored in previous work. Second, it is a multilingual task where the output score captures the relationship between source and target sentences. Finally, QE is fundamentally different from e.g. text classification, where clues are typically separate words or phrases (Zaidan et al., 2007) that can often be considered

[1] While QE is typically treated as a supervised task, a related research direction is *reference-free* evaluation, which refers to unsupervised cross-lingual metrics that assess MT quality by computing distances between cross-lingual semantic representations of the source and target sentences (Zhao et al., 2020; Song et al., 2021).

[2] A study on *global explainability* of MT evaluation metrics, disentangling them along linguistic factors such as syntax and semantics, has recently been conducted in Kaster et al. (2021). In contrast, our shared task addresses *local explainability* of individual input instances.

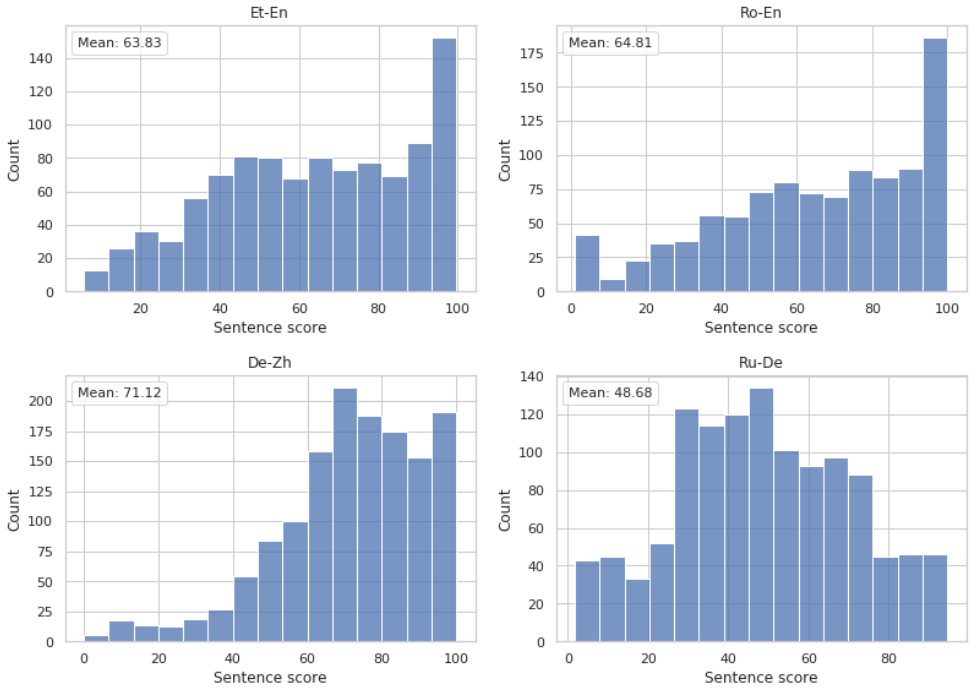

Figure 1: Distribution of sentence-level scores for each language pair.

independently of the rest of the text. By contrast, translation errors can only be identified given the context of the source and target sentences. Thus, this shared task provides a new benchmark for testing explainability methods in NLP.

On the other hand, treating word-level QE as an explainability problem offers some advantages compared to the current approaches. First, we can potentially avoid the need for supervised data at word level. Second, gold standard test sets can be made less expensive and more reliable. As we will show in Section 2, rationalized sentence-level evaluation can be a middle ground between relatively cheap but noisy annotations derived from post-editing (Fomicheva et al., 2020) and very informative but expensive explicit error annotation based on error taxonomies, such as the Multidimensional Quality Metrics (MQM) framework (Lommel et al., 2014b). For this shared task, we build a new test set with manually annotated explanations for sentence-level quality ratings. To the best of our knowledge, this is the first MT evaluation dataset annotated with human rationales.

The **main objective** of the shared task is threefold. First, it aims to explore the plausibility of explainable evaluation metrics (Wiegreffe and Pinter, 2019), by proposing a test set with manually annotated rationales. It helps the community better understand how similar the generated explanations are to the human explanations. Second, the shared task encourages research on unsupervised or semi-supervised methods for error identification, so as to reduce the cost on word-level MT error annotation. Last but not least, the shared task sheds light on how current NLP evaluation systems arrive at their predictions and to what extent this process is aligned with human reasoning.

2 Data

For this shared task, we collected a new test set with (i) manual assessment of translation quality at sentence level and (ii) word-level rationales that explain the sentence-level scores (Section 2.1). For training and development purposes, the participants were advised to use existing resources, which are briefly discussed in Section 2.2.

2.1 Eval4NLP Test Set

Language pairs and MT systems The test set contains four language pairs: Estonian-English (Et-En), Romanian-English (Ro-En), Russian-German (Ru-De) and German-Chinese (De-Zh). For Et-En and Ro-En, we use the source and translated sentences from the test21 partition of the MLQE-PE dataset (Fomicheva et al., 2020). For Ru-De, the source sentences were extracted from Wikipedia following the procedure described in Guzmán et al. (2019) and translated using the ML50 fairseq (Ott

et al., 2019) multilingual Transformer model (Tang et al., 2020). For De-Zh, the translations were produced using the Google Translate API, as the MT quality of the ML50 model was too low for this language pair according to our preliminary experiments.

Language pair	Tokens Source/Target	Sentences All/With rationales
Et-En	14,044/19,576	1,000/718
Ro-En	17,359/17,770	1,000/665
De-Zh	24,903/27,027	1,410/911
Ru-De	25,383/28,802	1,180/1,061

Table 1: Total number of source tokens, target tokens, sentences and sentences with lower-than-perfect sentence score (i.e. sentences with rationales) in the Eval4NLP 2021 test set.

Sentence- and word-level annotation For this annotation effort, we adapted the *Appraise* manual evaluation interface (Federmann, 2012). For sentence-level annotation, we follow the guidelines from the MLQE-PE dataset (Fomicheva et al., 2020), a variant of the so called *direct assessment* (DA) scores proposed by Graham et al. (2016). As illustrated in Figure 2, the annotators were asked to provide a sentence rating by moving a slider on the quality scale from left (worse) to right (best). They were additionally provided with instructions on what specific quality ranges represent. Following Graham et al. (2016), the numeric values were not visible to the annotators, but the scale is interpreted numerically as follows: *1-10* range represents a completely incorrect translation; *11-30*, a translation that contains a few correct keywords, but the overall meaning is different or lost; *31-50*, a translation that preserves parts of the original meaning; *51-70*, a translation which is understandable and conveys the overall meaning of the source but contains a few errors; *71-90*, a translation that closely preserves the semantics of the source and has only minor mistakes; and *91-100*, a perfect translation.

Crucially, besides the sentence-level rating, the annotators were asked to provide a rationale for their decisions. Specifically, for all translations except those they considered perfect, the annotators were required to highlight the words in the MT sentence corresponding to translation errors that would explain the assigned sentence score.[3] They were also asked to highlight the source words that caused the errors in the MT output, as shown in Figure 2. The missing contents was annotated by highlighting the source words that were not translated, whereas for the added (hallucinated) contents the annotators were only required to highlight the corresponding target words. We interpreted the highlighting as binary labels, indicating whether a given word is part of the rationale (positive class), or not (negative class). The annotators were provided with detailed annotation guidelines, which are available at https://github.com/eval4nlp/SharedTask2021/tree/main/annotation-guidelines.

The annotation was conducted by 3 annotators for Et-En and Ro-En, and by (up to) 4 annotators for Ru-De and De-Zh.[4] Et-En and Ro-En data was annotated by Estonian and Romanian native speakers with near native proficiency in English. De-Zh data was annotated by Chinese native speakers with strong proficiency in German. Finally, Ru-De data was annotated by native speakers of Russian with near native proficiency in German. The annotators for Ro-En, Et-En, and Ru-De are students specializing in Linguistics and Translation or are professional translators; the annotators for De-Zh are students specializing in computer science. The cost of annotation was approximately 4,000 Euro, with working times of 15 to 25 hours per annotator for De-Zh and Ru-De (Et-En and Ro-En annotators were compensated for the whole work, instead of on an hourly basis, and not all of them noted down their working times).

To produce a single sentence-level score, we take an average across the scores from individual annotators. To obtain a single binary label for each token, we use a majority voting mechanism, where the token is considered as part of the rationale if it was highlighted by the majority of the annotators.[5]

Inter-annotator agreement Table 2 shows average agreement levels between our annotators (on common sets of annotated data instances). We use Pearson correlation for sentence-level scores and

[3]For all languages except for Chinese, the source sentences and MT outputs were tokenized with Moses tokenizer available at https://github.com/moses-smt/mosesdecoder. For Chinese, the jieba tokenizer was used: https://github.com/fxsjy/jieba.

[4]Not all annotators annotated all sentences for Ru-De and De-Zh. Individual annotators did 1411, 871, 1101, 1026 sentences for De-Zh, and 601, 1002, 1181, 1001 sentences for Ru-De.

[5]When there is an even number of annotators, we weight the annotations by annotator reliability measured using their average agreement with the other annotators.

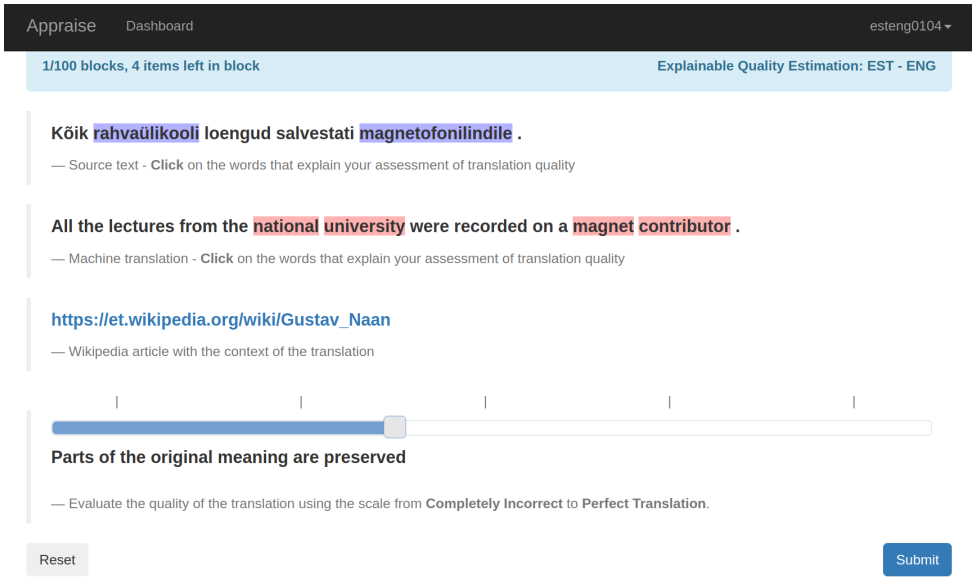

Figure 2: Screenshot of the annotation interface.

Cohen's kappa coefficient for word-level annotations. To be more precise, we measure Pearson correlation among all common instances between two annotators and then report the average across annotators; we measure average kappa agreement (averaged over all sentences) between any two annotators and then report the average across all annotators. We observe that Ro-En and Ru-De are most consistently annotated and De-Zh and Et-En have least agreement on average. Overall, our agreements are acceptable, however, in all cases, ranging from 0.42 to 0.67 kappa on word-level and ∼0.6 to 0.8 Pearson on sentence-level. For comparison, the average kappa reported by Lommel et al. (2014a) for the fine-grained MQM error annotation ranges from 0.25 to 0.34.

	Sentence-Level	**Word-Level**	
		Source	Target
De-Zh	0.59	0.48	0.55
Ru-De	0.71	0.56	0.59
Ro-En	0.81	0.54	0.67
Et-En	0.68	0.42	0.45

Table 2: Sentence-level (Pearson) and word-level (kappa) agreements for different language pairs.

Data statistics The number of annotated sentences, as well as the number of the source and target tokens in the test set are shown in Table 1. In addition, we show the number of sentences with lower-than-perfect translation quality. This is the final subset of sentences that was used to evaluate the submissions to the shared task, since in our manual evaluation setup no rationales were required for the MT outputs with perfect quality. As shown in Table 1, for all the language pairs the vast majority of translations has a lower-than-perfect score, where the percentage of such sentences is the lowest for De-Zh (65%) and the highest for Ru-De (90%).

Figure 1 shows the distribution of sentence-level scores for each language pair. The language pair with the highest average quality is De-Zh, whereas Ru-De has the lowest average score. For Et-En, Ro-En and Ru-De, the scores cover the whole quality range, while the distribution for De-Zh is highly skewed, which makes the task more challenging for this language pair (see Section 6).

Table 4 shows the proportion of words annotated as rationales. The numbers in Table 4 are consistent with the average sentence-level quality, as De-Zh and Ru-De have the lowest and the highest percentage, respectively. This is expected given that lower quality translations should contain a higher number of errors. In general, the proportion of tokens considered relevant for explaining sentence-level ratings is fairly low. This is consistent with the annotation guidelines which stipulate that all and only the words necessary to justify the sentence score must be highlighted. Finally, we observe that, for Et-En and Ro-En, the percentage of annotated tokens is higher for the target than for the source sentences. This can be related to the presence of hallucinations, where the target contains words that do not have a clear correspondence with any part

Src	Pe 20 august , trupele sârbe au început urmărirea austriecilor în retragere .											
PE	On 20 August , the Serbian troops began pursuing the retreating Austrians .											
MT		Serbian	troops	started	pursuing	Austria	on	20	August	in	withdrawal	.
Ann-EXPL		Serbian	troops	started	pursuing	Austria	on	20	August	in	withdrawal	.
Ann-EXPL*		Serbian	troops	started	pursuing	Austria	on	20	August	in	withdrawal	.
Predictions		Serbian	troops	started	pursuing	Austria	on	20	August	in	withdrawal	.
Ann-PE		Serbian	troops	started	pursuing	Austria	on	20	August	in	withdrawal	.
Color scale	0.0					0.5					1.0	

Table 3: Example of the target-side annotation from the Ro-En test set and the output expected from the participants. "Src" stands for the source sentence, "MT" is the MT output, "PE" is the post-edited version of the MT output taken from the MLQE-PE dataset. "Ann-EXPL*" is the mean of the binary scores for each word averaged across the annotators. "Ann-EXPL" corresponds to the binary scores obtained by aggregating individual annotations through majority voting (official gold standard of the shared task). "Ann-PE" is the word-level annotation derived from post-editing. "Predictions" contains the predictions (after min-max normalization) for this sentence from the IST-Unbabel submission to the constrained track.

		Et-En	Ro-En	De-Zh	Ru-De
Ours	Source	0.14	0.09	0.11	0.21
	MT	0.18	0.13	0.12	0.21
MLQE	Source	0.22	0.20	-	-
	MT	0.26	0.22	-	-

Table 4: Percentage of source and MT tokens annotated as rationales. For comparison, the percentage of source and target tokens annotated as errors in the same test partition of the MLQE-PE dataset is provided.

of the source sentence, as well as to typological differences between languages, whereby there tends to be a one-to-many correspondence between the source and target words.

Difference to existing QE datasets with word-level annotation The test set collected for this shared task is different from existing QE datasets with word-level annotation. A popular approach to building QE datasets is based on measuring post-editing effort (Bojar et al., 2017; Specia et al., 2018a; Fonseca et al., 2019; Specia et al., 2020). This can be done at sentence level, by computing the so called HTER score (Snover et al., 2006) that represents the minimum number of edits a human language expert is required to make in order to correct the MT output; or at word level, by aligning the MT output to its post-edited version and annotating the misaligned source and target words. An important limitation of this strategy is that the annotated words do not necessarily correspond to translation errors, as correcting a specific error may involve changing multiple related words in the sentence. This is exacerbated by the limitations of the heuristics used to automatically align the MT and its post-edited version. Indeed, as shown in Table 4, the percentage of error tokens on the same data for Ro-En and Et-En language pairs is considerably higher in the MLQE-PE dataset, where word-level annotation is derived from post-editing.

An alternative approach is the explicit annotation of translation errors by human experts. This is typically done based on fine-grained error taxonomies such as the Multidimensional Quality Metrics (MQM) framework (Lommel et al., 2014b). While such annotations provide very informative labelled data, the annotator agreement for this style of annotation is fairly low (Lommel et al., 2014a) and the annotation is very time-consuming.[6]

The example in Table 3 shows a sample of the annotated data from the Ro-En test set. The first three rows correspond to the source (Src), the MT output (MT), and the post-edited MT output (PE). "Ann-EXPL*" shows the mean of the binary scores assigned by each annotator to a given word. Thus, the words "20" and "August" were included in the rationale by 1 out of 3 annotators for this example, the words "Austria" and "in" were highlighted by 2 out of 3 annotators; finally, the word "withdrawal" was included in the rationale by all 3 annotators. We can interpret this information as an indirect indication of error severity, as the most serious errors are expected to be noted by all of the annotators. "Ann-EXPL" shows the binary scores that

[6]The interest towards MQM has recently increased due to a higher overall quality of MT (Freitag et al., 2021), but the aforementioned issues still remain unsolved.

we obtain through a majority voting mechanism, as described above. These binary scores were used for the official evaluation reported in Section 6. "Predictions" illustrates the predicted scores from one of the participants of the shared task.[7] The predictions almost perfectly correspond to the human rationale, as in both cases the words "Austria" and "withdrawal" receive the highest scores. Finally, for comparison, "Ann-PE" shows the word labels for this sentence taken from the MLQE-PE dataset. In this case all tokens are considered as errors since re-orderings (or "shifts") are not included in the set of possible edit operations used to compute minimum edit distance, from which the alignment between MT output and its PE is derived.

To the best of our knowledge, this test set is the first MT evaluation dataset annotated with human rationales. The proposed annotation scheme has certain advantages for the QE task, as it allows to explicitly annotate translation errors, and at the same time results in higher agreement and less effort than fine-grained error annotation.[8]

2.2 Training and development data

As discussed above, we use the same sentence-level annotation scheme as the one used in the MLQE-PE dataset. Therefore, for Ro-En and Et-En the participants could use the train and development partitions of MLQE-PE to build their sentence-level models. The De-Zh and Ru-De language pairs represent a fully zero-shot scenario where no sentence-level training data is available.

3 Task and Evaluation

The task consisted of building a QE system that (i) predicts the quality score for an input pair of source text and MT hypothesis, (ii) provides word-level evidence for its predictions. An example of the test data used for evaluation is shown in Table 3. The participants were expected to provide explanations for each sentence pair in the form of continuous scores, with the highest scores corresponding to the tokens considered as relevant by human annotators. The participants could submit to either **constrained** or **unconstrained** track. For the constrained track, the participants were expected to use no supervision at word level, while in the unconstrained track they were allowed to use any word-level data for training.

Explanations can be obtained either by building inherently interpretable models (Yu et al., 2019) or by using post-hoc explanation methods which extract explanations from an existing model (Ribeiro et al., 2016; Lundberg and Lee, 2017; Sundararajan et al., 2017a; Schulz et al., 2020), for example by analysing the values of the gradient on each input feature. In this shared task, we provide both sentence-level training data and strong sentence-level models (see the TransQuest-LIME baseline in Section 4), and thus encourage the participants to either train their own inherently interpretable models or use post-hoc techniques on top of our existing sentence-level models.

We accommodate the evaluation scheme to be suitable both for approaches that return continuous scores, and for supervised approaches that can return binary scores. Namely, we use evaluation metrics based on class probabilities that have been previously adapted for assessing the plausibility of rationale extraction methods (Atanasova et al., 2020). Since explainability methods typically proceed on instance-by-instance basis, and the scores produced for different instances are not necessarily comparable, we compute the evaluation metrics for each instance separately and average the results across all instances in the test set. Following Fomicheva et al. (2021), we define the following evaluation metrics to assess the performance of the submissions to the shared task at the word-level:

AUC score For each instance, we compute the area under the receiver operating characteristic curve (AUC score) to compare the continuous attribution scores against binary gold labels.

Average Precision AUC scores can be overly optimistic for imbalanced data. Therefore, we also use Average Precision (AP). AP summarizes a precision-recall curve as the weighted mean of precisions achieved at each threshold, with the increase in recall from the previous threshold used as the weight (Zhu, 2004).

Recall at Top-K In addition, we report the Recall-at-Top-K metric commonly used in information retrieval. Applied to our setting, this metric computes the proportion of words with the

[7] We did not ask the participants to normalize the scores, as we are only interested in the ranking of tokens according to their relevance for sentence-level quality.

[8] As shown by McDonnell et al. (2017), rationales increase the reliability of human annotation when judging the relevance of webpages for information retrieval. In the future, we plan to investigate whether this also applies to MT evaluation and providing word-level explanations increases the consistency of sentence-level assessments.

Team ID	Participating team	
NICT-Kyoto	National Institute of Information and Communications Technology	Rubino et al. (2021)
IST-Unbabel	IST/University of Lisbon & Unbabel	Treviso et al. (2021)
CLIP-UMD	Department of Computer Science, University of Maryland	Kabir and Carpuat (2021)
Gringham	Technical University of Darmstadt	Leiter (2021)
HeyTUDa	Technical University of Darmstadt	Eksi et al. (2021)
CUNI-Prague	Charles University	Polák et al. (2021)

Table 5: Participants of the Eval4NLP Shared Task on Explainable Quality Estimation.

highest attribution that correspond to translation errors against the total number of errors in the MT output. The code for computing the evaluation metrics can be found in the shared task github repository: https://github.com/eval4nlp/SharedTask2021/tree/main/scripts. The shared task used CodaLab as the submission platform.

4 Baseline systems

Random baseline is built by sampling scores uniformly at random from a continuous $[0..1)$ range for each source and target token in a given sentence pair as well as for the sentence-level QE score.

Transquest-LIME uses TransQuest QE models described in Ranasinghe et al. (2020) to produce sentence-level scores. TransQuest follows the current standard practice of building task-specific NLP models by fine-tuning pre-trained multilingual language models, such as XLM-Roberta, on task-specific data. For Ro-En and Et-En, the Ro-En and Et-En TransQuest models are used, whereas for the zero-shot language pairs we use the multilingual variant of TransQuest, which was trained on a concatenation of MLQE-PE data. The post-hoc LIME explanation method (Ribeiro et al., 2016) is then applied to generate relevance scores for the source and target words. LIME is a simplification-based explanation technique, which fits a linear model in the vicinity of each test instance, to approximate the decision boundary of the complex model. Since in our sentence-level gold standard higher scores mean better quality, we invert LIME explanations so that higher values correspond to errors.

XMover-SHAP uses the reference-free metric XMoverScore (Zhao et al., 2020) to rate translations and uses the (likewise post-hoc) SHAP explainer (Lundberg and Lee, 2017) to explain the ratings. In particular, given a source-translation pair, XMoverScore provides a real number to indicate the quality of the translation, in terms of its semantic overlapping with the source sentence, using re-mapped multilingual BERT embeddings and a target-side language model.[9] To explain the contribution of each word in the rating, SHAP creates perturbations of the source/translation sentence by masking out some words and estimates the average marginal contribution of each word across all possible perturbations. The source code for all the baseline systems is available at https://github.com/eval4nlp/SharedTask2021/tree/main/baselines.

5 Participants

For this first edition of the shared task, we had a total of 6 participating teams listed in Table 5.[10] Below, we briefly describe the submitted approaches.

NICT-Kyoto use synthetic data to fine-tune the XLM-Roberta language model for the QE task. To produce synthetic sentence-level scores, they translate publicly available parallel corpora using SOTA neural MT systems and compute three reference-based metrics: ChrF (Popović, 2015), TER (Snover et al., 2006) and BLEU (Papineni et al., 2002). To simulate word-level annotation, they derive word-level labels from the alignment between the MT outputs and human reference translations. The QE model is then jointly trained to predict the scores from different metrics as well as word-level tags. A metric embedding component is proposed where each metric is represented with a set of learnable parameters. An attention mechanism between the metric embeddings and the input representations is employed to obtain word-level scores as explanations for the sentence-level predictions.

IST-Unbabel participated in the constrained and unconstrained tracks of the shared task. For the constrained track ("IST-Unbabel" in Table 6), they

[9] Note that XMoverScore is an unsupervised reference-free metric, in contrast to the supervised TransQuest QE model.

[10] Initially, there were seven participating teams, but one of them opted out after the competition ended.

used a set of explainability methods to extract the relevance of the input tokens from sentence-level QE models built on top of XLM-Roberta and RemBERT. The explainability methods explored in this work include attention-based, gradient-based and perturbation based approaches, as well as rationalization by construction. The best performing method which was submitted to the competition relies on the attention mechanism of the pre-trained Transformers in order to obtain the relevance scores for each token. In addition, scaling attention weights by the L2 norm of value vectors as suggested in Kobayashi et al. (2020) resulted in a further boost in performance.

For the unconstrained track ("IST-Unbabel*" in Table 6), they add a word-level loss to the sentence-level models and train jointly using the annotated data from the MLQE-PE dataset.

HeyTUDa use the TransQuest QE models (Ranasinghe et al., 2020) for sentence-level prediction and a set of explainability techniques to estimate the relevance of each source and target word. Specifically, they explore three perturbation-based methods: LIME, SHAP, and occlusion (Zeiler and Fergus, 2014), as well as three gradient-based methods: DeepLift (Shrikumar et al., 2017), Layer Gradient x Activation (Shrikumar et al., 2016) and Integrated Gradients (Sundararajan et al., 2017b). They further use an unsupervised ensembling method to combine the different explainability approaches.

Gringham use the reference-free metrics XBERTScore (i.e., BERTScore (Zhang et al., 2020) with cross-lingual embeddings) and XMoverScore and make them inherently interpretable by considering the token alignments produced by the models. The intuition is that words that are not well-aligned are most likely erroneous. Specifically, they explore XBERTScore and XMoverScore as sentence-level models and use the corresponding similarity (or distance) matrices to produce token-level scores.

CLIP-UMD propose an ensemble of two approaches: (1) the LIME explanation technique applied to the TransQuest sentence-level model; (2) Divergent mBERT (Briakou and Carpuat, 2020), which is a BERT-based model that can detect cross-lingual semantic divergences. Divergent mBERT is trained using synthetic data where semantic divergences are introduced automatically following a set of pre-defined perturbations. To produce a combination of the two methods, the predictions from each approach are averaged.

CUNI-Prague participated in the unconstrained track. They fine-tune the XLM-R model for word-level and sentence-level QE. To map sentence piece tokenization from XLM-R to Moses tokenization, they ignore all sentence piece tokens corresponding to a given Moses token except the first one.

6 Results

Table 6 shows the results of the shared task. We report the word-level metrics presented in Section 3, as well as Pearson correlation at sentence level. The values of the "Rank" columns are computed by first ranking the participants according to each of the three word-level metrics and then averaging the resulting rankings.[11] First, we note that all of the submissions outperform the three baselines for all the language pairs,[12] which indicates that error detection can indeed be approached as rationale extraction.

Approaches Overall, the submitted approaches vary a lot in the way they addressed the task. The following trends can be identified:

- Following the recent standard in QE and similar multilingual NLP tasks, all the approaches rely on multilingual Transformer-based language models.
- Submissions to the unconstrained track use the SOTA approach to word-level supervision explored previously by Lee (2020).
- The use of synthetic data produced by aligning MT outputs and reference translations from existing parallel corpora proves an efficient strategy to identify translation errors. Supervising the predictions based on Transformer attention weights with the labels derived from synthetic data was used by the winning submission to the shared task.
- The approaches that rely on attention weights to predict human rationales (NICT-Kyoto and IST-Unbabel) achieve the best results for the constrained track.

[11] This ranking is slightly different from the Codalab results, as one of the teams retracted from the competition.

[12] The only exception is HeyTUDa, which is outperformed by XMover-SHAP for De-Zh and by TransQuest-LIME for Et-En and Ro-En.

	Target				Source				Sentence
	Rank	AUC	AP	Rec	Rank	AUC	AP	Rec	Pearson
Estonian-English									
IST-Unbabel*	1.0	**0.92**	**0.85**	**0.76**	1.3	**0.94**	**0.86**	**0.77**	0.86
CUNI Prague*	2.0	0.92	0.84	**0.75**	3.0	0.93	**0.85**	**0.76**	0.80
NICT Kyoto†	3.0	**0.90**	0.82	**0.73**	1.7	0.93	0.85	**0.77**	0.85
IST-Unbabel	4.3	0.82	0.74	0.63	4.0	0.86	0.76	0.64	0.82
Gringham	4.7	0.84	0.71	0.60	5.0	0.86	0.72	0.59	0.71
CLIP-UMD	6.0	0.74	0.63	0.53	6.0	0.76	0.51	0.45	0.77
Baseline: TransQuest-LIME	7.3	0.62	0.54	0.43	7.0	0.54	0.44	0.31	0.77
HeyTUDa	7.7	0.66	0.52	0.41	10.	N/A	N/A	N/A	0.77
Baseline: XMover-SHAP	9.0	0.62	0.44	0.34	8.0	0.54	0.37	0.23	0.49
Baseline: Random	10.	0.50	0.36	0.25	9.0	0.49	0.34	0.19	-0.03
Romanian-English									
NICT Kyoto†	1.0	**0.95**	**0.87**	**0.78**	1.0	**0.95**	**0.85**	**0.75**	0.92
IST-Unbabel*	2.0	0.94	0.84	0.75	2.0	0.93	0.81	0.71	0.87
CUNI Prague*	3.0	0.94	0.83	0.73	3.0	0.93	0.81	0.70	0.89
IST-Unbabel	4.0	0.88	0.78	0.68	4.0	0.86	0.73	0.62	0.90
Gringham	5.0	0.87	0.73	0.61	5.0	0.84	0.61	0.45	0.78
CLIP-UMD	6.0	0.73	0.60	0.49	6.0	0.72	0.41	0.37	0.90
Baseline: TransQuest-LIME	7.7	0.63	0.52	0.42	7.7	0.48	0.35	0.24	0.90
HeyTUDa	7.7	0.68	0.50	0.38	10.	N/A	N/A	N/A	0.90
Baseline: XMover-SHAP	8.7	0.67	0.44	0.30	8.0	0.53	0.29	0.15	0.70
Baseline: Random	10.	0.52	0.31	0.19	8.3	0.50	0.28	0.15	0.02
Russian-German									
NICT Kyoto†	1.0	**0.93**	**0.83**	**0.74**	1.0	**0.92**	**0.80**	**0.71**	0.68
IST-Unbabel*	2.0	0.80	0.64	0.52	2.0	0.85	0.71	0.59	0.67
CUNI Prague*	3.3	0.76	0.61	0.50	3.3	0.80	0.67	0.56	0.61
Gringham	4.3	0.79	0.57	0.46	3.7	0.84	0.67	0.56	0.60
IST-Unbabel	4.3	0.75	0.58	0.47	5.0	0.77	0.63	0.52	0.64
CLIP-UMD	6.0	0.65	0.46	0.36	6.3	0.66	0.41	0.37	0.30
HeyTUDa	7.3	0.54	0.33	0.23	10.	N/A	N/A	N/A	0.50
Baseline: XMover-SHAP	7.7	0.52	0.33	0.23	8.0	0.52	0.36	0.26	0.25
Baseline: Random	9.0	0.49	0.31	0.22	9.0	0.51	0.34	0.24	-0.02
Baseline: TransQuest-LIME	10.	0.40	0.26	0.16	6.7	0.53	0.43	0.32	0.50
German-Chinese									
NICT Kyoto†	1.0	**0.85**	**0.68**	**0.57**	1.0	**0.85**	**0.64**	**0.51**	0.29
IST-Unbabel	2.3	0.68	0.50	0.37	3.0	0.67	0.47	0.32	0.33
IST-Unbabel*	2.7	0.71	0.47	0.34	2.0	0.73	0.50	0.35	0.27
CUNI Prague*	4.3	0.61	0.44	0.30	5.0	0.62	0.42	0.27	0.25
CLIP-UMD	4.7	0.63	0.40	0.27	6.3	0.61	0.31	0.25	0.50
Gringham	6.3	0.55	0.36	0.22	4.0	0.64	0.42	0.27	-0.04
Baseline: XMover-SHAP	6.7	0.55	0.33	0.22	9.0	0.47	0.29	0.16	0.18
HeyTUDa	8.0	0.51	0.31	0.18	10.	N/A	N/A	N/A	0.34
Baseline: Random	9.0	0.50	0.29	0.17	7.7	0.50	0.30	0.17	0.00
Baseline: TransQuest-LIME	10.	0.46	0.27	0.14	7.0	0.49	0.32	0.20	0.34

Table 6: Official results of the Eval4NLP Shared Task on Explainable Quality Estimation. Submissions to the unconstrained track are marked with *. We mark the NICT Kyoto submissions with a †, as they submitted to the constrained track, but use synthetic data for word-level supervision. Submissions not significantly outperformed by any other submission according to paired t-test for each metric are marked in bold. N/A means that the participating team did not submit the word-level scores for the source sentences.

- Both IST-Unbabel and HeyTUDa explore a wide set of explanation methods. The differences in performance are likely due to the method used for the final submission. While IST-Unbabel submission explores normalized attention weights, HeyTUDa use an ensemble of gradient-based approaches. A possible reason for the inferior performance of HeyTUDa is that the gradient is computed with respect to the embedding layer. As noted by Fomicheva et al. (2021), attribution to the embedding layer in the Transformer-based QE models does not provide strong results for the error detection task since word representations

at the embedding layer do not capture contextual information, which is crucial for predicting translation quality.

- Gringham follow an entirely different strategy where they modify an existing reference-free metric to obtain both sentence score and word-level explanations in an unsupervised way. A similar approach is explored in our XMover-SHAP baseline, but the difference is that we apply SHAP explainer on top of XMover, while Gringham makes the XMover-Score inherently interpretable, which leads to better results.

Winners The overall winner of the competition is the submission to the constrained track from NICT-Kyoto, which wins on 3 out of 4 language pairs, according to the source and target ranking. Fine-tuning on large amounts of synthetic data as well as the use of attention mechanism between the evaluation metric embeddings and the contextualized input representations seem to be the key to their performance. We note, however, that they offer a mixed approach with word-level supervision on synthetic data. Among the constrained approaches that do not use any supervision at word level, the best performing submission is IST-Unbabel, which outperforms other constrained submissions for all language pairs, except Ru-De, where they perform on par with Gringham on the target side and are surpassed by Gringham on the source side. For the unconstrained track we received only two submissions, from which IST-Unbabel* performs the best.

Sentence-level correlation is not predictive of the performance of the submissions at detecting relevant tokens. This is due to the fact that submitted approaches vary in the role played by the sentence-level model. In fact, if we look at the submissions that follow comparable strategies, we do observe a correspondence between sentence-level and token-level results. For example, among the approaches that build upon a sentence-level QE model and use post-hoc methods to explain the predictions, IST-Unbabel tends to achieve higher performance both in terms of the token-level results and in terms of the Pearson correlation with sentence ratings, compared to HeyTUDa and the TransQuest-LIME baseline.

The performance on zero-shot language pairs is lower than for Et-En and Ro-En. This is the case for all approaches except NICT-Kyoto on Ru-De, where the performance at word-level is comparable to the results for Et-En and Ro-En, even though the Pearson correlation for sentence scores is inferior. We attribute this outcome to the use of supervision with synthetic data, which helps boost performance for word-level QE when no manually labelled data is available, as has been shown by Tuan et al. (2021). Performance degradation for De-Zh is considerably larger than Ru-De. De-Zh was among the language pairs with the lowest inter-annotator agreement and, in addition, had a different distribution of sentence-level scores, with many high-quality translations, according to the annotators (see Section 2.1).

Limitations of the evaluation settings Our current evaluation settings can be further improved in various ways. First, the submissions were ranked according to the global statistics, i.e. by comparing the mean AUC, AP and Rec-TopK scores of different submissions over a common set of test instances. However, such aggregation mechanisms ignore how many of its competitors a given submission outperforms and on how many test instances. In the future we plan to follow a more rigorous approach suggested by Peyrard et al. (2021) and use the Bradley–Terry (BT) model (Bradley and Terry, 1952), which leverages the instance-level pairing of metric scores.

Second, the metrics used for evaluation are tailored for unsupervised explainability approaches that produce continuous scores, but they do not allow a direct comparison with the SOTA work on word-level QE, which is evaluated using F-score and Matthews correlation coefficient (Specia et al., 2020). One way to address this would be to require the participants to submit binary scores, but we discarded this option in this first edition of the shared task, as it would substantially limit the exploration of the explainability approaches.

Finally, the binary rationales obtained from our pool of annotators through majority voting do not capture the fact that some words are more relevant for sentence-level quality than others. As shown in Table 3, an alternative version of the data can be produced by averaging the scores assigned to each word by individual annotators, as an indication of the severity of translated errors. In the future, we plan to study to what extent such scores agree with the continuous explanation scores produced by the participants. Another limitation of our annotation

scheme is that sometimes a word may be missing in the machine translation, which can then not be highlighted (e.g., Russian does often not use determiners and the MT system may wrongly omit it when translating into English or German).

7 Conclusions

In this paper, we presented the findings of the Eval4NLP-2021 shared task on *explainable Quality Estimation (QE)*, where the goal is to not only produce a sentence-level score for an MT output, given a source sentence, but also highlight erroneous words in the target (and source) sentence explaining the score. We detailed the data annotation, involving two novel non-English language pairs, our baselines (post-hoc explanation techniques on top of state-of-the-art QE models), as well as the participants' approaches to the task. These include supervised approaches, training on synthetic data as well as genuine post-hoc and inherent explainability techniques.

The scope for future research is huge: for example, we aim to include new language pairs, especially low-resource ones, address explainability for metrics in other NLP tasks, e.g. semantic textual similarity (Agirre et al., 2016) and summarization (Gao et al., 2020), and identify error categories of highlighted words, ideally in an unsupervised manner.

Acknowledgments

Marina Fomicheva was supported by funding from the Bergamot project (EU H2020 Grant No. 825303). Piyawat Lertvittayakumjorn was supported by a scholarship from Anandamahidol Foundation. We would like to thank Lisa Yankovskaya and Mark Fishel from the University of Tartu for helping organize and monitor the manual quality annotation. We also thank Anton Malinovskiy for adapting the Appraise interface for quality annotation with rationales. Finally, we gratefully thank the Artificial Intelligence Journal (https://aij.ijcai.org/) and Salesforce Research for their financial support enabling our human annotations.

References

Eneko Agirre, Aitor Gonzalez-Agirre, Iñigo Lopez-Gazpio, Montse Maritxalar, German Rigau, and Larraitz Uria. 2016. SemEval-2016 task 2: Interpretable semantic textual similarity. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 512–524, San Diego, California. Association for Computational Linguistics.

Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. 2020. A diagnostic study of explainability techniques for text classification. *arXiv preprint arXiv:2009.13295*.

John Blatz, Erin Fitzgerald, George Foster, Simona Gandrabur, Cyril Goutte, Alex Kulesza, Alberto Sanchis, and Nicola Ueffing. 2004. Confidence estimation for machine translation. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, pages 315–321, Geneva, Switzerland. COLING.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Raphael Rubino, Lucia Specia, and Marco Turchi. 2017. Findings of the 2017 conference on machine translation (wmt17). In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Tasks Papers*, Copenhagen, Denmark. Association for Computational Linguistics.

Ralph Allan Bradley and Milton E. Terry. 1952. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345.

Eleftheria Briakou and Marine Carpuat. 2020. Detecting Fine-Grained Cross-Lingual Semantic Divergences without Supervision by Learning to Rank. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1563–1580, 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.

Marina Danilevsky, Kun Qian, Ranit Aharonov, Yannis Katsis, Ban Kawas, and Prithviraj Sen. 2020. A survey of the state of explainable AI for natural language processing. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 447–459, Suzhou, China. 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.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. 2020. ERASER: A benchmark to evaluate rationalized NLP models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4443–4458, Online. Association for Computational Linguistics.

Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

Melda Eksi, Erik Gelbing, Jonathan Stieber, and Chi Viet Vu. 2021. Explaining errors in machine translation with absolute gradient ensembles. In *Proceedings of the 2nd Workshop on Evaluation and Comparison for NLP systems*.

Christian Federmann. 2012. Appraise: An open-source toolkit for manual evaluation of machine translation output. *The Prague Bulletin of Mathematical Linguistics*, 98:25–35.

Marina Fomicheva, Lucia Specia, and Nikolaos Aletras. 2021. Translation error detection as rationale extraction.

Marina Fomicheva, Shuo Sun, Erick Fonseca, Frédéric Blain, Vishrav Chaudhary, Francisco Guzmán, Nina Lopatina, Lucia Specia, and André F. T. Martins. 2020. MLQE-PE: A multilingual quality estimation and post-editing dataset. *arXiv preprint arXiv:2010.04480*.

Erick Fonseca, Lisa Yankovskaya, André F. T. Martins, Mark Fishel, and Christian Federmann. 2019. Findings of the WMT 2019 shared tasks on quality estimation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 1–10, Florence, Italy. Association for Computational Linguistics.

Markus Freitag, George 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.

Yang Gao, Wei Zhao, and Steffen Eger. 2020. SUPERT: Towards new frontiers in unsupervised evaluation metrics for multi-document summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1347–1354, Online. Association for Computational Linguistics.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2016. Can machine translation systems be evaluated by the crowd alone. *Natural Language Engineering*, FirstView:1–28.

Francisco Guzmán, Peng-Jen Chen, Myle Ott, Juan Pino, Guillaume Lample, Philipp Koehn, Vishrav Chaudhary, and Marc'Aurelio Ranzato. 2019. The FLORES evaluation datasets for low-resource machine translation: Nepali–English and Sinhala–English. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6098–6111, Hong Kong, China. Association for Computational Linguistics.

Sarthak Jain, Sarah Wiegreffe, Yuval Pinter, and Byron C. Wallace. 2020. Learning to faithfully rationalize by construction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4459–4473, Online. Association for Computational Linguistics.

Tasnim Kabir and Marine Carpuat. 2021. The umd submission to the explainable mt quality estimation shared task: Combining explanation models with sequence labeling. In *Proceedings of the 2nd Workshop on Evaluation and Comparison for NLP systems*.

Marvin Kaster, Wei Zhao, and Steffen Eger. 2021. Global explainability of bert-based metrics by disentangling along linguistic factors. In *EMNLP 2021*, Online. Association for Computational Linguistics.

Goro Kobayashi, Tatsuki Kuribayashi, Sho Yokoi, and Kentaro Inui. 2020. Attention is not only a weight: Analyzing transformers with vector norms. *arXiv preprint arXiv:2004.10102*.

Dongjun Lee. 2020. Two-phase cross-lingual language model fine-tuning for machine translation quality estimation. In *Proceedings of the Fifth Conference on Machine Translation, Volume 2: Shared Tasks Papers*, Online. Association for Computational Linguistics.

Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2016. Rationalizing neural predictions. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 107–117, Austin, Texas. Association for Computational Linguistics.

Christoph Wolfgang Leiter. 2021. Reference-free word- and sentence-level translation evaluation with token-matching metrics. In *Proceedings of the 2nd Workshop on Evaluation and Comparison for NLP systems*.

Arle Lommel, Maja Popovic, and Aljoscha Burchardt. 2014a. Assessing inter-annotator agreement for translation error annotation. In *LREC Workshop on Automatic and Manual Metrics for Operational Translation Evaluation*.

Arle Richard Lommel, Aljoscha Burchardt, and Hans Uszkoreit. 2014b. Multidimensional quality metrics (MQM): A framework for declaring and describing translation quality metrics. *Tradumàtica: tecnologies de la traducció*, 0(12):455–463.

Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc.

Tyler McDonnell, Mücahid Kutlu, Tamer Elsayed, and Matthew Lease. 2017. The many benefits of annotator rationales for relevance judgments. In *IJCAI*, pages 4909–4913.

Joseph E Mercado, Michael A Rupp, Jessie YC Chen, Michael J Barnes, Daniel Barber, and Katelyn Procci. 2016. Intelligent agent transparency in human–agent teaming for multi-uxv management. *Human factors*, 58(3):401–415.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling.

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.

Maxime Peyrard, Wei Zhao, Steffen Eger, and Robert West. 2021. Better than average: Paired evaluation of NLP systems. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2301–2315, Online. Association for Computational Linguistics.

Peter Polák, Muskaan Singh, and Ondřej Bojar. 2021. Explainable quality estimation: Cuni eval4nlp submission. In *Proceedings of the 2nd Workshop on Evaluation and Comparison for NLP systems*.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 372–375, Lisboa, Portugal. Association for Computational Linguistics.

Tharindu Ranasinghe, Constantin Orasan, and Ruslan Mitkov. 2020. Transquest at wmt2020: Sentence-level direct assessment. In *Proceedings of the Fifth Conference on Machine Translation, Volume 2: Shared Tasks Papers*, Online. Association for Computational Linguistics.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 1135–1144, New York, NY, USA. Association for Computing Machinery.

Raphael Rubino, Atsushi Fujita, and Benjamin Marie. 2021. Error identification for machine translation with metric embedding and attention. In *Proceedings of the 2nd Workshop on Evaluation and Comparison for NLP systems*.

Karl Schulz, Leon Sixt, Federico Tombari, and Tim Landgraf. 2020. Restricting the flow: Information bottlenecks for attribution. *arXiv preprint arXiv:2001.00396*.

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*, ICML'17, page 3145–3153. JMLR.org.

Avanti Shrikumar, Peyton Greenside, Anna Shcherbina, and Anshul Kundaje. 2016. Not just a black box: Learning important features through propagating activation differences. *CoRR*, abs/1605.01713.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231.

Yurun Song, Junchen Zhao, and Lucia Specia. 2021. SentSim: Crosslingual semantic evaluation of machine translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3143–3156, Online. Association for Computational Linguistics.

Lucia Specia, Frédéric Blain, Marina Fomicheva, Erick Fonseca, Vishrav Chaudhary, Francisco Guzmán, and André F. T. Martins. 2020. Findings of the WMT 2020 shared task on quality estimation. In *Proceedings of the Fifth Conference on Machine Translation*, pages 743–764, Online. Association for Computational Linguistics.

Lucia Specia, Frédéric Blain, Varvara Logacheva, Ramón F. Astudillo, and André F. T. Martins. 2018a. Findings of the WMT 2018 shared task on quality estimation. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 689–709, Belgium, Brussels. Association for Computational Linguistics.

Lucia Specia, Carolina Scarton, and Gustavo Henrique Paetzold. 2018b. Quality estimation for machine translation. *Synthesis Lectures on Human Language Technologies*, 11(1):1–162.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017a. Axiomatic attribution for deep networks. In *International Conference on Machine Learning*, pages 3319–3328. PMLR.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017b. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 3319–3328. JMLR.org.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning.

Ehsan Toreini, Mhairi Aitken, Kovila Coopamootoo, Karen Elliott, Carlos Gonzalez Zelaya, and Aad Van Moorsel. 2020. The relationship between trust in ai and trustworthy machine learning technologies. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 272–283.

Marcos V. 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 for NLP systems*.

Yi-Lin Tuan, Ahmed El-Kishky, Adithya Renduchintala, Vishrav Chaudhary, Francisco Guzmán, and Lucia Specia. 2021. Quality estimation without human-labeled data. *arXiv preprint arXiv:2102.04020*.

Sarah Wiegreffe and Yuval Pinter. 2019. Attention is not not explanation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 11–20, Hong Kong, China. Association for Computational Linguistics.

Mo Yu, Shiyu Chang, Yang Zhang, and Tommi S Jaakkola. 2019. Rethinking cooperative rationalization: Introspective extraction and complement control. *arXiv preprint arXiv:1910.13294*.

Omar Zaidan, Jason Eisner, and Christine Piatko. 2007. Using "annotator rationales" to improve machine learning for text categorization. In *Human language technologies 2007: The conference of the North American chapter of the association for computational linguistics; proceedings of the main conference*, pages 260–267.

Matthew D. Zeiler and Rob Fergus. 2014. Visualizing and understanding convolutional networks. In *Computer Vision – ECCV 2014*, pages 818–833, Cham. Springer International Publishing.

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*.

Wei Zhao, Goran Glavaš, Maxime Peyrard, Yang Gao, Robert West, and Steffen Eger. 2020. On the limitations of cross-lingual encoders as exposed by reference-free machine translation evaluation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1656–1671, Online. Association for Computational Linguistics.

Mu Zhu. 2004. Recall, precision and average precision. *Department of Statistics and Actuarial Science, University of Waterloo, Waterloo*, 2(30):6.