Explaining with Contrastive Phrasal Highlighting:
A Case Study in Assisting Humans to Detect Translation Differences

Eleftheria Briakou[1]* **Navita Goyal**[2], **Marine Carpuat**[2]
[1] Google, [2] University of Maryland
ebriakou@google.com, {navita, marine}@cs.umd.edu

Abstract

Explainable NLP techniques primarily explain by answering *"Which tokens in the input are responsible for this prediction?"*. We argue that for NLP models that make predictions by comparing two input texts, it is more useful to explain by answering *"What differences between the two inputs explain this prediction?"*. We introduce a technique to generate *contrastive phrasal highlights* that explain the predictions of a semantic divergence model via phrase-alignment-guided erasure. We show that the resulting highlights match human rationales of cross-lingual semantic differences better than popular post-hoc saliency techniques and that they successfully help people detect fine-grained meaning differences in human translations and critical machine translation errors.

1 Introduction

A common strategy to explain the predictions of NLP models is to highlight salient tokens in their inputs (Li et al., 2016; Ribeiro et al., 2016; Lundberg and Lee, 2017). However, this is suboptimal for the many NLP tasks that require comparing and contrasting two or more pieces of text to predict a class or a similarity score, such as natural language inference (Bowman et al., 2015), semantic textual similarity (Agirre et al., 2012), or evaluation and quality estimation of text generation (Bojar et al., 2017; Ma et al., 2018, 2019; Mathur et al., 2020; Freitag et al., 2021b, 2022). At the same time, a long line of research in social sciences shows that human explanations are contrastive (Miller, 2019), i.e., humans do not explain why an event happened in a vacuum but instead compare what happened to a contrast case. For instance, to explain how a Greek sentence and a translation in English differ, it is more intuitive to show how they differ (as presented in Figure 1) than to highlight all salient Greek and English tokens without specifying how they relate to each other.

Figure 1: The color-coded contrastive phrasal highlights explain meaning differences between a Greek sentence (EL) and its English translation (EN) from Wikipedia.

In this work, we introduce a post-hoc explainability technique to answer *"What differences between two inputs explain a prediction"?* for NLP regression models that produce a score encoding a relationship between two text inputs. We apply it to the task of explaining meaning divergences between a text in one language and its translation. We design a phrase-alignment-guided erasure strategy to explain the predictions of a semantic divergence model and highlight minimal phrase pairs in the two inputs that result in meaning differences.

We extensively evaluate our approach with both proxy and human-centered evaluations. Comparing against human explanations for meaning differences in English-French human translations, our approach matches human rationales better than popular post-hoc methods that highlight salient tokens independently (Ribeiro et al., 2016; Lundberg and Lee, 2017). Acknowledging the limitations of such proxy evaluations (Buçinca et al., 2020), we conduct two IRB-approved human studies to assess the usefulness of our contrastive phrasal highlights when crowd-sourcing challenging bilingual annotation tasks. We find that our approach helps annotators reliably discover fine-grained meaning differences in English-French and English-Spanish human translations and detect critical errors in Portuguese-to-English machine trans-

---

*Work done while at the University of Maryland.

lation. We make our code and data publicly available: `https://github.com/Elbria/Ex-SemDiv`.

2 Background

We focus on *local* explanations which explain why a specific prediction was made, often to help humans rely on predictions appropriately in machine-in-the-loop use cases (Doshi-Velez and Kim, 2017).

Highlight Explanations Local explanations often consist of highlighting input features deemed important to a prediction. In glass-box settings, prior works seek to quantify feature saliency with model-internal signals such as gradients (Arras et al., 2016; Smilkov et al., 2017), contextual decompositions (Murdoch et al., 2018), or attention mechanisms (Choi et al., 2016). However, these highlights' ability to serve as explanations has been questioned (Jain and Wallace, 2019; Wang et al., 2020; Moradi et al., 2021), leading to work revisiting technical solutions and evaluations under which attention *could* be seen as an explanation (Wiegreffe and Pinter, 2019; Moradi et al., 2019; Tutek and Snajder, 2020). Another family of approaches relies on erasure strategies to explain the predictions of black box models. Li et al. (2016) quantify a token's importance as the difference in model confidence once it is erased from the input, with several following works exploring different token erasure schemes (Feng et al., 2018; Kim et al., 2020). LIME (Ribeiro et al., 2016) approximates the model's local decision boundary with linear models by erasing multiple tokens from a model's input, while SHAP (Lundberg and Lee, 2017) computes Shapley values by estimating the marginal contribution of each token across all possible erasures. We build on this line of work by designing an erasure strategy that identifies salient *contrastive phrasal pairs* rather than independent salient tokens.

Contrastive Explanations Motivated by social science research suggesting that humans explain with respect to an (implicit) contrast case (Miller, 2019), recent work has sought to *generate* the latter by producing counterfactual explanations: minimally edited versions of the input that change the model prediction. Performing such edits usually requires training dedicated editors (Ross et al., 2021), prompting language models (Paranjape et al., 2021), and accessing knowledge bases (Chen et al., 2021), among others (Li et al., 2020; Chemmengath et al., 2022). Closer to our work, an-

other family of approaches extracts *highlights* contrastively: Jacovi and Goldberg (2021) extend erasure-based approaches to identify which input features lead to the actual prediction vs. a contrastive prediction, while Yin and Neubig (2022) extend saliency methods to explain the generation of contrast cases for language generation tasks. Our work explains the predictions of models that compare two inputs, where identifying salient phrase pairs that differ across them provides a natural mechanism for contrastive explanations.

Evaluation Most current work in NLP adopts *proxy* evaluations that compare automatic explanations with human explanations of the gold label, with numerous datasets encoding explanations using highlights or free-form text (Wiegreffe and Marasovic, 2021). Others rely on *simplified* human tasks, such as simulatability, evaluating whether explanations help people predict a model's prediction more accurately (Hase and Bansal, 2020; Nguyen, 2018). Despite the attractiveness of the above evaluations, evidence is growing that they do not reliably indicate how useful explanations are in practice (Buçinca et al., 2020). Boyd-Graber et al. (2022) call for application-grounded evaluations to directly assess whether explanations help people complete an actual task, as suggested by work in HCI (Suresh et al., 2021; Liao et al., 2020). We heed that call by complementing a proxy evaluation with two application-grounded user studies.

Explaining Translation Differences Detecting translation differences is a core task in multilingual NLP, e.g., to predict machine translation errors (Rei et al., 2020), to study how humans translate (Zhai et al., 2020), or to understand how multilingual resources such as Wikipedia differ across languages (Miquel-Ribé and Laniado, 2018; Jiang et al., 2017). Automatic approaches often score the degree of (dis)similarity between an input and its translation to quantify machine translation quality (Zhang et al., 2019; Sellam et al., 2020; Rei et al., 2020) or to detect meaning divergences in human translations (Vyas et al., 2018; Briakou and Carpuat, 2020; Wein and Schneider, 2021). However, sentence-level scores can be hard to interpret and do not pinpoint how translations differ. This has been addressed by tagging text with human translation processes (Zhai et al., 2018), with dedicated, supervised models or with word-level quality estimation (QE) tasks (Specia et al., 2018,

2020, 2021), often addressed by applying post-hoc explanation methods to sentence-level predictors (Treviso et al., 2021; Rei et al., 2022). We augment this work by contributing unsupervised contrastive explanations of translation differences and testing whether they help annotate them more reliably.

3 Explaining Divergences with Contrastive Phrasal Highlights

We introduce a highlighting approach to explain the predictions of a model that compares and contrasts two text inputs. We hypothesize that phrase pairs provide more appropriate *cognitive chunks* (Doshi-Velez and Kim, 2017) than tokens as basic units of explanations. For instance, instead of highlighting tokens that differ in the English-French pair of Figure 2, it is more natural to explain that the English text refers to the Maple Leaf Forever as a *national anthem* as opposed to a *chant patriotique* (patriotic song) in French and that the English phrase *"for many years"* has no equivalent in French.

Building on token-level erasure-based explanation methods (Li et al., 2016), we design a phrase-alignment-guided erasure strategy to explain the predictions of a model (Briakou and Carpuat, 2020) that ranks bilingual sentence pairs based on the granularity of their semantic similarity (assuming $\mathbf{S}$ is an equivalent pair and $\hat{\mathbf{S}}$ a pair containing semantic divergences, the model $\mathcal{R}(\cdot)$ ranks them such that $\mathcal{R}(\mathbf{S}) > \mathcal{R}(\hat{\mathbf{S}})$). As shown in Figure 2, given two input texts, we first extract a set of candidate counterfactual inputs by masking a subset of phrase pairs (§3.1). Then, we explain the model's prediction by selecting the phrase pair whose erasure maximally increases the similarity score between the two inputs (§3.2).

3.1 Alignment-Guided Phrasal Erasure

We propose an erasure approach that takes into account the input's structure. Given a sentence pair $\mathbf{S}$, we start by extracting a set of candidate counterfactual instances by deleting a single phrase pair from $\mathbf{S}$. Given that erasing all possible phrase pairs for each sentence is computationally prohibitive, we restrict our search to deleting phrases that belong to an aligned phrase table, $\mathcal{P}$.

Our phrase pair candidates for erasure are derived from classical statistical machine translation techniques developed for extracting phrase translation dictionaries from parallel corpora (Och et al., 1999; Koehn et al., 2007). Given two texts, we de-

rive word alignments based on multilingual word embeddings (Jalili Sabet et al., 2020) and then extract all phrase pairs consistent with the word alignments. A phrase pair (p_1, p_2) is consistent with the word alignment if p_1 and p_2 are each contiguous sequences of words in each language, where all alignment links that start within p_1 map to tokens within p_2 and vice versa. Unaligned words (i.e., words aligned to a `null` token) can be included in either the source (p_1) or the target (p_2) phrases. As a result, the extracted phrase pairs comprise not only equivalent but also include related but divergent phrase pairs (*"national anthem"* and *"chant patriotique"*) and unaligned phrase (*"served for many years"*), as seen in Figure 2.

3.2 Explaining by Minimal Contrast

Given the sentence pair $\mathbf{S}$ and its aligned phrase table $\mathcal{P}$, we first extract a set of contrast cases $\tilde{\mathcal{P}}$ consisting of all the phrasal pairs that, once erased, make the two inputs more equivalent, as measured by an increase in score larger than a margin ϵ:

$$\tilde{\mathcal{P}} = \left\{ p \in \mathcal{P}, \text{ s.t. } \mathcal{R}\Big(\textbf{DEL}\big[\mathbf{S}; p\big]\Big) > \mathcal{R}\Big(\mathbf{S}\Big) + \epsilon \right\}$$

Since explanations should be minimal (Hilton, 2017; Lipton, 1990), covering only the most relevant causes to reduce cognitive load, we select as contrastive phrasal highlights the phrase pairs that correspond to minimal edits with a large impact on predictions, as follows:

$$\arg\max \left\{ \mathcal{R}\Big(\textbf{DEL}\big[\mathbf{S}; p\big]\Big) \cdot \mathcal{BR}\Big(\mathbf{S}, p\Big) \right\}$$

$$\mathcal{BR}(\mathbf{S}, p) = \begin{cases} e^{-\frac{|p|}{|\mathbf{S}|}}, & \text{if } \mathcal{R}\Big(\textbf{DEL}\big[\mathbf{S}; p\big]\Big) \geq 0 \\ e^{+\frac{|p|}{|\mathbf{S}|}}, & \text{otherwise} \end{cases}$$

where the first term $\mathcal{R}(\textbf{DEL}\big[\mathbf{S}; p\big])$ encourages the extraction of a contrastive phrasal highlight p corresponding to a high score under the model (i.e., deleting this phrase pair yields a contrast case), while the second term, $\mathcal{BR}(\mathbf{S}, p)$ corresponds to a *brevity reward* that encourages extraction of shorter highlights. $|\mathbf{S}|$ is computed by adding the length of each of the two sentences, and $\textbf{DEL}[\cdot]$ is a function that erases a phrase pair p from $\mathbf{S}$.
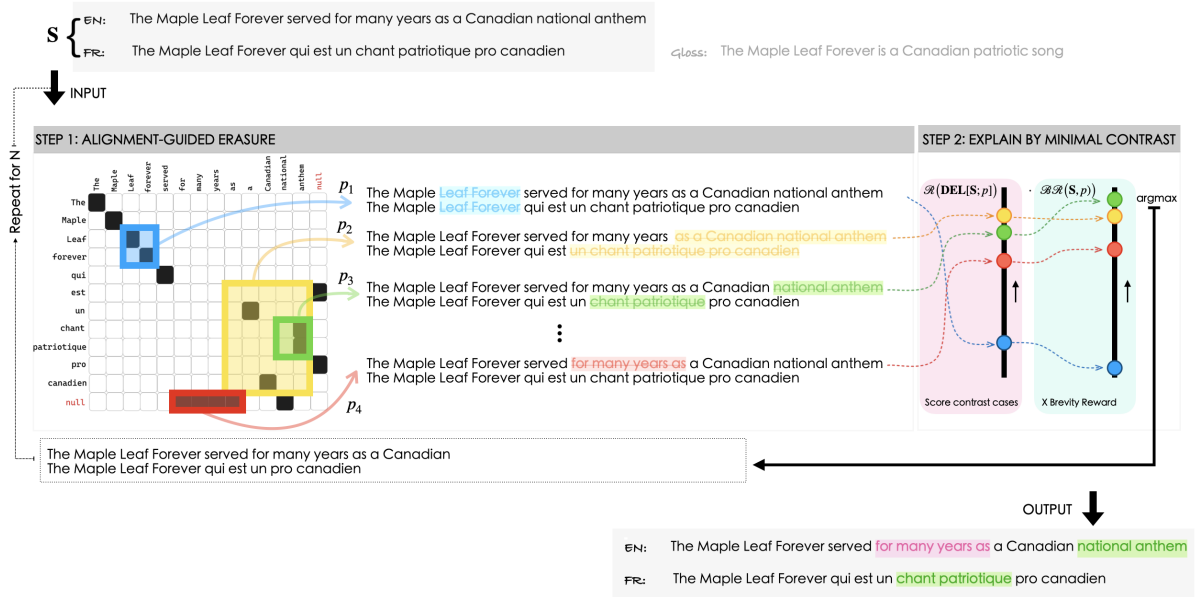
Figure 2: Our approach takes as input a sentence pair (**S**) and extracts a set of perturbed inputs by erasing phrasal pairs guided by word alignments (Step 1). Then it explains the prediction of a regressor $\mathcal{R}(\mathbf{S})$ by highlighting the phrasal pair p that, once deleted, maximizes the model's prediction $\mathcal{R}(\text{DEL}[\mathbf{S}; p])$ multiplied by a brevity reward $\mathcal{BR}(\mathbf{S}, p)$ that encourages the extraction of short phrasal pairs (Step 2).

The above approach identifies a single phrase pair that explains the divergence prediction for the original sentence **S**. To extract multiple explanations, we repeat this process by iteratively erasing the extracted contrastive phrasal highlight from the current input sentence $\mathbf{S}' = \text{DEL}[\mathbf{S}; p]$, and repeat the steps in §3.1 and §3.2. This iterative process ends when, for a given input, none of the extracted counterfactual instances yield a more equivalent pair, i.e., $\tilde{\mathcal{P}} = \varnothing$, or we reach an equivalent input under the divergent ranker, i.e., $\mathcal{R}(\mathbf{S}') > 0$.

4 Proxy Evaluation

In this section, we describe our proxy evaluation based on human-provided highlights. We acknowledge that proxy evaluations encounter validity issues (Boyd-Graber et al., 2022) and use them primarily to guide system development and validate against standard highlighting methods.

4.1 Experimental Setup

Explanandum We seek to explain the prediction of a divergence ranking model that is trained as recommended in prior work (see Appendix A.1).

Explainers We contrast our contrastive phrasal highlights against a RANDOM baseline that highlights tokens in each sentence at random with 0.5 probability, and two standard post-hoc explanation

methods, LIME and SHAP, that seek to explain the predictions of the explanandum.

Reference Highlights We evaluate our approach using the REFRESD dataset,[1] which is manually annotated with divergences of fine-grained and coarse-grained granularity, along with rationales that justify the sentence label. We compare our contrastive phrasal highlights with the human rationales on a subset of 418 challenging instances annotated as having "Some Meaning Differences" at the sentence level, where we expect the more subtle differences in meaning to be found.

Evaluation Metrics Following prior proxy evaluations of explanations (Ross et al., 2021; DeYoung et al., 2020), we compute: 1. *Agreement with Human Rationales*: Precision, Recall, and F-1 scores computed against human rationales; 2. *Minimality*: the length of the contrastive phrasal highlights, measured as the number of highlighted tokens.

4.2 Results

As seen in Table 1, explaining divergence predictions by extracting contrastive phrasal highlights significantly outperforms both LIME and SHAP. Those standard highlighting baselines even under-

---

[1] https://github.com/Elbria/xling-SemDiv/tree/master/REFreSD

	ENGLISH				FRENCH			
	PR.	RE.	F1	DEL.	PR.	RE.	F1	DEL.
ORACLE	100	100	100	39%	100	100	100	43%
RANDOM	39	48	43	50%	42	49	45	50%
LIME	45	37	41	30%	44	37	40	34%
SHAP	53	34	41	25%	50	32	39	26%
OURS (− BR)	52	76	62	54%	56	74	64	55%
OURS	58	61	62	37%	60	55	64	37%

Table 1: Proxy evaluation results with respect to human rationales on REFRESD.

HUMAN	The older generation turbines generate kilowatts , and the modern turbines installed generate up to 3 megawatts , depending on the specific turbine and manufacturer ..
	Les turbines d' ancienne génération génèrent des kilowatts , alors que les éoliennes modernes ont une puissance pouvant aller jusqu' à 3 mégawatts .
	(gloss) The turbines of the old generation generate kilowatts, while the modern wind turbines generate up to 3 megawatts of power.
LIME	The older generation turbines generate kilowatts , and the modern turbines installed generate up to 3 megawatts , depending on the specific turbine and manufacturer ..
	Les turbines d' ancienne génération génèrent des kilowatts , alors que les éoliennes modernes ont une puissance pouvant aller jusqu' à 3 mégawatts .
SHAP	The older generation turbines generate kilowatts , and the modern turbines installed generate up to 3 megawatts , depending on the specific turbine and manufacturer ..
	Les turbines d' ancienne génération génèrent des kilowatts , alors que les éoliennes modernes ont une puissance pouvant aller jusqu' à 3 mégawatts .
OURS	The older generation turbines generate kilowatts , and the modern turbines installed generate up to 3 megawatts, depending on the specific turbine and manufacturer ..
	Les turbines d' ancienne génération génèrent des kilowatts , alors que les éoliennes modernes ont une puissance pouvant aller jusqu' à 3 mégawatts .

Table 2: Examples of divergence explanations (**HUMAN** corresponds to REFRESD rationales).

perform the RANDOM baseline.[2] A closer look at the outputs (Table 2) indicates that both LIME and SHAP suffer from two major issues: sparsity and accuracy. We attribute those issues to the fact that both approaches operate in token space, ignoring the interdependent relationships between tokens across the two languages. By explicitly modeling such relationships, our approach produces highlights that match the ones in REFRESD better.

We measure the impact of the brevity reward by ablation (i.e., OURS (− BR)). This achieves the highest recall at the expense of precision: dropping the brevity reward produces fewer and longer highlights per instance. It highlights more than 50% of each sentence pair on average, while humans only highlight about 40%. The brevity reward thus helps match reference highlights better, supporting the benefits of producing minimal explanations.

In sum, these results are promising indicators that contrastive phrasal highlights could be helpful when detecting semantic divergences. However, as highlighted by recent work on explanation evalu-

ation, proxy evaluations can be misleading. Given that prior work evaluates highlight-based explanations of (machine) translation errors based *solely on automatic evaluations*, we chose to initiate a human-centered discussion that evaluates the usefulness of such highlights directly with an in-depth exploration of multiple small-scale *application-grounded* evaluations as detailed below.

5 Application-Grounded Evaluation I: Annotation of Semantic Divergences

As shown by prior work, annotating fine-grained divergences is a challenging task and requires dedicated annotation protocols based on human rationales (Briakou and Carpuat, 2020) or abstract meaning representation frameworks (Wein et al., 2022) to achieve moderate agreement. As a result, such annotations are usually hard to collect with crowd workers since they require explicit annotator training. In this section, we ask whether contrastive phrasal highlights help crowd-workers annotate divergences more reliably, which could ease the need for complex and expensive annotation protocols. Concretely, we test the following hypothesis: *Contrastive phrasal highlights improve the annotation of fine-grained semantic divergences in terms of accuracy and agreement.*

---

[2]Since LIME and SHAP are feature attribution methods that assign a continuous score to each word, we checked that changing the threshold from the default $t = 0$ does not improve the results, as both methods suffer from a precision-recall tradeoff.
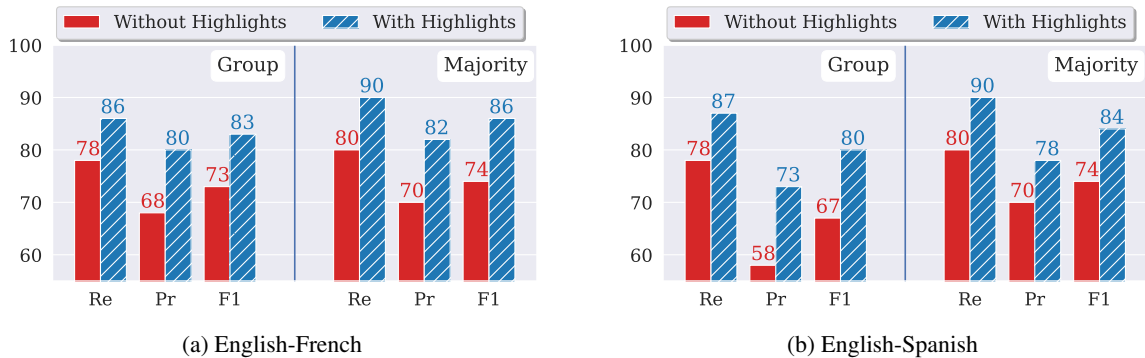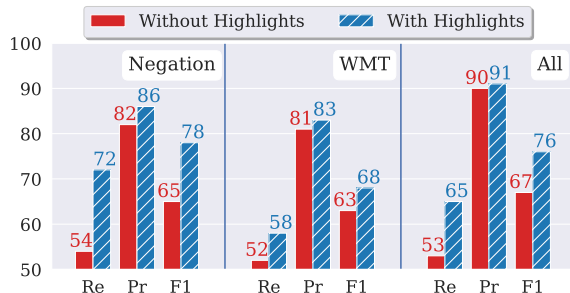
Figure 3: Annotation-grounded evaluation comparing with- vs. without-highlights annotation conditions. Contrastive phrasal highlights significantly ($p = 0.1$) improve the annotation of fine-grained semantic divergences.

5.1 Experimental Setup

Explanandum We seek to explain the predictions of Divergent mBERT. We train separate models for English-French and English-Spanish following the approach detailed in Appendix A.1.

Study Data To provide controlled yet realistic test samples, we construct a dataset of fine-grained divergences that mimic translation processes used by human translators (Zhai et al., 2019). We start with English-French and English-Spanish human translations from the multi-parallel FLORES benchmark (Goyal et al., 2022). We randomly select 50 *seed* translations among those identified as semantically equivalent pairs by Divergent mBERT. Then, we introduce fine-grained meaning differences by editing the English references of 10 of those samples. We perform edits motivated by translation processes, such as modulation, explicitation, and reduction (Zhai et al., 2019), by interaction with ChatGPT.[3] For instance, we introduce a generalization as follows:

> Despite leaving the show in 1993 he kept *[the title of executive producer] → [a senior role]* and continued to receive tens of millions of dollars every season in royalties.

To create difficult examples of equivalent translations, we also paraphrase the English references of 10 seed examples with ChatGPT to introduce syntactic changes. We validated all examples, ensuring that edits actually match the intended divergences in the first case (Figures 6 and 7) and that syntactic changes do not alter meaning in the second (Figure 8). The final dataset consists of 35 semantically equivalent translations (split into 25 original trans-

lations and 10 syntactically divergent paraphrases) and 15 fine-grained divergences resembling translation processes.

5.2 Study Design

We ran a controlled evaluation across the two language pairs (English-French and English-Spanish). We describe the task and study design below.

Task Description An annotation instance is a sentence pair in English and, e.g., French. Bilingual speakers are asked to read each sentence pair closely, determine whether the *"sentence pair contains any difference in meaning"*, and finally characterize the difference as "Added" vs. "Changed"."Added" refers to added information, while "Changed" refers to information that is present in both languages but does not precisely match—mirroring prior annotation divergence protocols (Briakou and Carpuat, 2020). We include a screenshot of the task description presented to annotators in Figures 5 and 7 of Appendix A.4.

Conditions We study two conditions: one in which participants are shown highlights (✓ HIGHLIGHTS) and a second in which they are not (✗ HIGHLIGHTS). The only information available to participants about the highlights is that they *"are AI-generated and indicative of meaning differences"*. We emphasize that highlights are not presented as explanations of divergence predictions, as we want annotators to choose how and whether they use them in their assessments based on their intuitions.

Procedures We conduct a between-subjects study where participants are randomly assigned to a condition. Participants are first presented with a tutorial that explains the task and relevant ter-

---

[3]Details of the ChatGPT prompts are in Appendix A.3.

minology. Each participant is presented with 25 instances from one of the two studied conditions. Each batch of annotated instances contained 30% of semantically divergent edits, 20% of syntactically divergent edits, and 50% original semantically equivalent pairs. Instances within each batch are randomized. We include two attention checks where participants are asked to indicate their answers to the previous question. After completing the task, participants were asked to complete a brief survey assessing their perception of the generated highlights (if assigned to this condition) and finally were asked to provide their free-form feedback on the study along with demographic information, such as gender and age. The average time of the study was 20 minutes. In sum, we collect 3 annotations per instance and annotate a total of 100 instances (50 per condition) for each of the two language pairs studied.

Participants   We recruit 12 participants per language pair using Prolific.[4] Each participant is restricted to taking the study only once. None of them failed both attention checks; hence we did not exclude any of the collected annotations from our analysis. All participants identified as bilingual speakers in the languages involved in each study. Participants are compensated at 15 USD per hour.

5.3   Measures

Our main evaluation measures are Precision, Recall, and F1 computed by assuming that semantically edited instances correspond to divergences, while the rest are treated as semantically equivalent pairs. We report those accuracy statistics both at the group level and also by majority voting annotations. Furthermore, we summarize the responses provided as free-form feedback and report subjective measures that reflect participants' perceived understanding of the explanations, i.e., *"the highlights are useful in detecting meaning differences"*, if provided. The latter measures are collected on a 5-point Likert scale. We report statistical significance based on bootstrap resampling. We draw samples randomly from the collected annotations with replacement. The sample size is the same as the one of the collected annotations, while the number of resamples is set to 1,000. For the application-grounded evaluation I, which contains a total of 150 annotations per language, the significance level is $p = 0.1$.

---

[4] https://www.prolific.co/

5.4   Study Results

Reliability of Annotations   As shown in Figure 3, the bilingual *group* that annotated fine-grained meaning differences in the presence of contrastive explanations—✓ HIGHLIGHTS , is significantly ($p = 0.1$) more accurate across Precision, Recall, and F1 scores and both language pairs, compared to ✗ HIGHLIGHTS . Furthermore, those improvements carry over when aggregating annotation results by *majority* voting annotations across instances. As a result, this leads us to accept the hypothesis that contrastive phrasal highlights improve the annotation of fine-grained divergences by bilingual speakers. Finally, detecting divergences with highlights additionally improves the reliability of annotations are measured by Cohen's Kappa inter-annotator agreement statistics:

	EN-FR	EN-ES
✗ HIGHLIGHTS	0.51 (moderate)	0.33 (fair)
✓ HIGHLIGHTS	0.66 (substantial)	0.52 (moderate)

Subjective Measures & User Feedback   Overall, bilingual speakers presented with contrastive phrasal highlights agreed they were useful in helping them spot fine-grained divergences—average self-reported usefulness of 3.8 for EN-FR and 4.2 for EN-ES. Finally, although contrastive phrasal highlights were useful, annotators also note that they cannot entirely rely on them. For instance, some of the participant's feedback is *"The highlights are useful but not 100% reliable, that is because I found other added/changed words that AI did not highlight"* and *"In most cases, the words highlighted by the AI have helped to detect possible differences"*.

6   Application-Grounded Evaluation II: Critical Error Detection

In this section, we examine the potential of using contrastive phrasal highlights to assist bilingual humans in detecting critical errors in machine translation outputs. Recent work on quality estimation for machine translation proposes an error-based evaluation framework where bilingual *professional* annotators are asked to: highlight translation errors; rate their severity as major or minor; and finally, indicate their type (e.g., mistranslation vs. terminology, etc.) (Freitag et al., 2021a). Drawing on those intuitions, we aim to study whether a simplified evaluation framework that uses contrastive

phrasal highlights yields reliable annotations of critical errors with bilingual *crowd workers*. This is a hard task as these errors are rare in high-quality systems and might be missed when parsing translations quickly. In what follows, we test the hypothesis: *Contrastive phrasal highlights help bilingual crowd-workers detect critical (accuracy) errors in machine translation outputs*.

6.1 Experimental Setup

Explanandum We explain the predictions of Divergent mBERT trained for English-Portuguese following the process detail in Appendix A.1. We emphasize that the explanandum is sensitive to detecting any meaning differences and not only critical errors, and therefore, we expect the contrastive phrasal highlights to surface both minor and major accuracy errors.

Study Data For the purposes of the study, we use the synthetic Portuguese-English dataset from the Critical Error Detection task of WMT (Freitag et al., 2021b). The dataset consists of pairs sampled from the OPUS parallel corpus (Tiedemann, 2012) treated as "not containing a critical error". Then, the WMT organizers artificially corrupted 10% of those pairs to introduce critical errors that reflect hallucinated content, untranslated content, or mistranslated texts. We will refer to those errors as WMT for the rest of the paper. Additionally, to make the task hard for humans, we a) filter out WMT critical errors that concern deviation in numbers, time, units, or dates and b) artificially introduce *negation* errors: we edit the English text locally by changing one or two words in such a way that the meaning of the sentence is entirely flipped (e.g., the word *benefits* in an English reference is replaced by the word *harms*). In total, the study dataset consists of 30 translations that do not contain a critical error (also detected as equivalent under the explanandum), 10 translations reflecting WMT errors, and 10 translations reflecting local *negation* errors.

6.2 Study Design

We test our hypothesis with a user study of critical error detection in English translations of Portuguese texts, as described below.

Task Description An annotation instance is an excerpt in Portuguese and its (machine) translation in English. Bilingual speakers are asked to read the two excerpts, determine whether the translation contains an accuracy error, and then rate its severity

Figure 4: Application-grounded evaluation results comparing ✗ HIGHLIGHTS vs. ✓ HIGHLIGHTS annotation conditions. Contrastive phrasal highlights significantly improve the Recall and F1 scores when detecting *negation* errors ($p = 0.05$) and errors in the ALL set. Precision improvements across sets and any improvements on the WMT set are not significant ($p > 0.05$).

as being *minor* or *major*. Following Freitag et al. (2021a), minor errors are defined as errors that do not lead to loss of meaning and would not confuse or mislead the reader, while major ones may confuse or mislead the reader due to significant changes in meaning. We ask participants to factor out fluency issues in their assessment as much as possible. We include a screenshot of the task description presented to annotators in Figures 9 and 10 of Appendix A.4.

Conditions & Evaluation Constructs We use Precision, Recall, and F1 as our main evaluation measures and follow the same conditions and subjective measures described in §5. The above scores are calculated only against major errors, as those are the ones for which we have gold labels. We report statistical significance based on bootstrap resampling as outlined in §5.3. For this study, which contains a total of 250 annotations, the significance level is $p = 0.05$.

Procedures & Participants We collect 5 assessments per instance and annotate a total of 100 instances (50 per condition). Each participant is presented with 25 instances, among which 20% correspond to WMT critical errors, 20% to negation errors, and the rest to original OPUS parallel texts. We recruit a total of 40 participants, all of whom self-identified as proficient in English and Portuguese. All other participant recruitment details and study procedures are the same as in §5.

6.3 Study Results

Main Findings As shown in Figure 4, the user group presented with contrastive phrasal highlights

Table 3: Critical Error Detection user study results. Each row represents an annotator, and each cell is an annotation, where ✔ indicates no accuracy error detected, ⚠ indicates a minor error, and ⊘ a critical error.

exhibits a higher recall in detecting critical errors compared to the one that does not access them. Crucially, we report significant improvements ($p = 0.05$) in detecting *negation* errors, which are harder to spot. A closer look at the differences between the two error classes reveals that detecting negation errors is more challenging as annotators have to pay close attention to the compared texts and the relationships between them to spot local mistranslations that cause a major shift in meaning. On the other hand, WMT errors are more easily spotted as they mostly represent detached hallucination phenomena (i.e., an entire phrase is added to the English translation), which we hypothesize can be implied by length differences of the compared texts. Our full annotation results, presented in Table 3, further validate our hypothesis: WMT errors are not genuinely missed out but rather rated as minor by some annotators. On the other hand, usually, negation errors are not only misperceived as minor but are entirely overlooked. We include examples of annotations in Figure 10.

Subjective Measures & User Feedback   Overall, bilingual speakers found the highlights helpful, with an average self-reported usefulness of 3.9. Additionally, they reported they would like to use them as a tool to assist them in detecting critical errors, with an average score of 3.8. We include details on the distribution of ratings in Table 4. Last, a

closer look at the users' feedback sheds some light on our approach's current strengths and limitations. Although highlights were in principle useful and a *"good feature to have especially when proofreading a great amount of translated texts"*, some annotators spotted *"a high percentage of false positives"* (see Figures 10i and 10j in Appendix A.4) and raised concerns that *"relying solely on them might make them less aware of errors hidden in non-highlighted texts"* (see Figure 10b in Appendix A.4).

7   Conclusion

We introduce an approach to extracting contrastive phrasal highlights that explain NLP models that take as input *two* texts. Unlike existing techniques that highlight input *tokens* independently, we consider the input's structure and explicitly model the relationships between contrasting inputs.

We study the effectiveness of contrastive phrasal highlights by explaining the predictions of a divergence ranker that compares and contrasts meaning differences in bilingual texts. Our proxy evaluation confirms our approach outperforms standard explainability approaches, that highlight tokens, by matching human-provided rationales of divergences in English-French Wikipedia texts. Finally, through a series of human-centered evaluations, we explore the usefulness of contrastive phrasal highlights in application-grounded contexts for three language pairs. Our results suggest that contrastive phrasal highlights assist bilingual speakers in detecting fine-grained meaning differences in (human) translated texts and critical errors due to local mistranslations in machine-translated texts.

Our findings create opportunities for designing better machine-in-the-loop pipelines to identify critical machine translation errors grounded in high-stake settings, study translation data at scale, and facilitate the creation of multilingual content through crowd-based efforts.

Acknowledgements

Limitations

Our work contributed evidence that contrastive phrasal highlights can provide a framework that assists humans in detecting meaning differences of a specific nature (i.e., fine-grained translation processes or local critical errors). However, detecting translation differences in the wild requires covering the entire distribution of meaning differences (Freitag et al., 2021a), which we leave for future work.

Moreover, although our method builds on unsupervised modules (i.e., Divergent mBERT and SimAlign) that do not rely on supervised data and can, in principle, be applied to any language pair, we have only evaluated our approach in high-resource settings where we expect both the explanandum and the word alignment models to be reasonably accurate. Therefore, further work should be conducted to explore how our findings generalize to other settings, such as in low-resource regimes where we might expect the alignment to be of poorer quality, introducing errors that may impact humans' perceived understanding of the contrastive highlights differently.

Additionally, future work should explore how the alignment module should be operationalized for other NLP tasks that take as input two texts, potentially exploring monolingual aligners (Lan et al., 2021) or additional structured information such as abstract meaning representations (Banarescu et al., 2013).

Finally, the significance of our results could be strengthened by increasing the sample sizes of future user studies. We view our current findings as solid starting points for an in-depth exploration of the usefulness of highlights for human-centered applications that future studies can build upon and extend with broad explorations.

Ethics Statement

Both studies are approved by the University of Maryland Institutional Review Board (IRB number $2018458 - 1$). As discussed in the paper, all participants involved in our studies gave their consent and were compensated at a rate of 15 USD per hour. We collected minimal demographic information, and participants could opt out of answering the demographic questions. We did not collect any further potentially identifiable information. Participation in our study was completely voluntary, and participants had the right to withdraw at any time.

References

Eneko Agirre, Daniel Cer, Mona Diab, and Aitor Gonzalez-Agirre. 2012. SemEval-2012 task 6: A pilot on semantic textual similarity. In *SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 385–393, Montréal, Canada. Association for Computational Linguistics.

Leila Arras, F. Horn, Grégoire Montavon, Klaus-Robert Müller, and Wojciech Samek. 2016. "what is relevant in a text document?": An interpretable machine learning approach. *PLoS ONE*, 12.

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract Meaning Representation for sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 178–186, Sofia, Bulgaria. Association for Computational Linguistics.

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

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Jordan Boyd-Graber, Samuel Carton, Shi Feng, Q. Vera Liao, Tania Lombrozo, Alison Smith-Renner, and Chenhao Tan. 2022. Human-centered evaluation of explanations. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Tutorial Abstracts*, pages 26–32, Seattle, United States. Association for Computational Linguistics.

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.

Zana Buçinca, Phoebe Lin, Krzysztof Z Gajos, and Elena L. Glassman. 2020. Proxy tasks and subjective measures can be misleading in evaluating explainable ai systems. *Proceedings of the 25th International Conference on Intelligent User Interfaces*.

Saneem Chemmengath, Amar Prakash Azad, Ronny Luss, and Amit Dhurandhar. 2022. Let the CAT out of the bag: Contrastive attributed explanations for text. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7190–7206, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Qianglong Chen, Feng Ji, Xiangji Zeng, Feng-Lin Li, Ji Zhang, Haiqing Chen, and Yin Zhang. 2021. KACE: Generating knowledge aware contrastive explanations for natural language inference. 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 2516–2527, Online. Association for Computational Linguistics.

E. Choi, Mohammad Taha Bahadori, Jimeng Sun, Joshua A. Kulas, Andy Schuetz, and Walter F. Stewart. 2016. Retain: An interpretable predictive model for healthcare using reverse time attention mechanism. In *NIPS*.

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*.

Shi Feng, Eric Wallace, Alvin Grissom II, Mohit Iyyer, Pedro Rodriguez, and Jordan Boyd-Graber. 2018. Pathologies of neural models make interpretations difficult. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3719–3728, Brussels, Belgium. Association for Computational Linguistics.

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

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André F. T. Martins. 2022. Results of WMT22 metrics shared task: Stop using BLEU – neural metrics are better and more robust. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 46–68, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

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

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Peter Hase and Mohit Bansal. 2020. Evaluating explainable AI: Which algorithmic explanations help users predict model behavior? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5540–5552, Online. Association for Computational Linguistics.

Denis Hilton. 2017. Social attribution and explanation. *Oxford Handbook of Causal Reasoning*, page 645–676.

Alon Jacovi and Yoav Goldberg. 2021. Aligning faithful interpretations with their social attribution. *Transactions of the Association for Computational Linguistics*, 9:294–310.

Sarthak Jain and Byron C. Wallace. 2019. Attention is not explanation. In *North American Chapter of the Association for Computational Linguistics*.

Masoud Jalili Sabet, Philipp Dufter, François Yvon, and Hinrich Schütze. 2020. SimAlign: High quality word alignments without parallel training data using static and contextualized embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1627–1643, Online. Association for Computational Linguistics.

Yuncheng Jiang, Wen Bai, Xiaopei Zhang, and Jiaojiao Hu. 2017. Wikipedia-based information content and semantic similarity computation. *Information Processing & Management*, 53(1):248–265.

Siwon Kim, Jihun Yi, Eunji Kim, and Sungroh Yoon. 2020. Interpretation of NLP models through input marginalization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3154–3167, Online. Association for Computational Linguistics.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran,

Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

Wuwei Lan, Chao Jiang, and Wei Xu. 2021. Neural semi-Markov CRF for monolingual word alignment. 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 6815–6828, Online. Association for Computational Linguistics.

Chuanrong Li, Lin Shengshuo, Zeyu Liu, Xinyi Wu, Xuhui Zhou, and Shane Steinert-Threlkeld. 2020. Linguistically-informed transformations (LIT): A method for automatically generating contrast sets. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 126–135, Online. Association for Computational Linguistics.

Jiwei Li, Will Monroe, and Dan Jurafsky. 2016. Understanding neural networks through representation erasure. *ArXiv*, abs/1612.08220.

Qingzi Vera Liao, Dan Gruen, and Sarah Miller. 2020. Questioning the ai: Informing design practices for explainable ai user experiences. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*.

Peter Lipton. 1990. Contrastive explanation. *Royal Institute of Philosophy Supplements*, 27:247–266.

Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Qingsong Ma, Ondřej Bojar, and Yvette Graham. 2018. Results of the WMT18 metrics shared task: Both characters and embeddings achieve good performance. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 671–688, Belgium, Brussels. Association for Computational Linguistics.

Qingsong Ma, Johnny Wei, Ondřej Bojar, and Yvette Graham. 2019. Results of the WMT19 metrics shared task: Segment-level and strong MT systems pose big challenges. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 62–90, Florence, Italy. Association for Computational Linguistics.

Nitika Mathur, Johnny Wei, Markus Freitag, Qingsong Ma, and Ondřej Bojar. 2020. Results of the WMT20 metrics shared task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 688–725, Online. Association for Computational Linguistics.

Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sc iences. *Artificial Intelligence*, 267:1–38.

Marc Miquel-Ribé and David Laniado. 2018. Wikipedia culture gap: Quantifying content imbalances across 40 language editions. *Frontiers in Physics*, 6.

Pooya Moradi, Nishant Kambhatla, and Anoop Sarkar. 2019. Interrogating the explanatory power of attention in neural machine translation. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pages 221–230, Hong Kong. Association for Computational Linguistics.

Pooya Moradi, Nishant Kambhatla, and Anoop Sarkar. 2021. Measuring and improving faithfulness of attention in neural machine translation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2791–2802, Online. Association for Computational Linguistics.

W. James Murdoch, Peter J. Liu, and Bin Yu. 2018. Beyond word importance: Contextual decomposition to extract interactions from lstms. Cite arxiv:1801.05453Comment: Under review at ICLR 2018.

Dong Nguyen. 2018. Comparing automatic and human evaluation of local explanations for text classification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1069–1078, New Orleans, Louisiana. Association for Computational Linguistics.

Franz Josef Och, Christoph Tillmann, and Hermann Ney. 1999. Improved alignment models for statistical machine translation. In *1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*.

Bhargavi Paranjape, Julian Michael, Marjan Ghazvininejad, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2021. Prompting contrastive explanations for commonsense reasoning tasks. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4179–4192, Online. 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. 2022. CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid). 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.

Alexis Ross, Ana Marasović, and Matthew Peters. 2021. Explaining NLP models via minimal contrastive editing (MiCE). In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3840–3852, Online. Association for Computational Linguistics.

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

Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda B. Viégas, and Martin Wattenberg. 2017. Smoothgrad: removing noise by adding noise. *ArXiv*, abs/1706.03825.

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

Lucia Specia, Frédéric Blain, Varvara Logacheva, Ramón F. Astudillo, and André F. T. Martins. 2018. 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.

Harini Suresh, Steven R. Gomez, Kevin K. Nam, and Arvind Satyanarayan. 2021. Beyond expertise and roles: A framework to characterize the stakeholders of interpretable machine learning and their needs. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA. Association for Computing Machinery.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

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.

Martin Tutek and Jan Snajder. 2020. Staying true to your word: (how) can attention become explanation? In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 131–142, Online. Association for Computational Linguistics.

Yogarshi Vyas, Xing Niu, and Marine Carpuat. 2018. Identifying semantic divergences in parallel text without annotations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1503–1515, New Orleans, Louisiana. Association for Computational Linguistics.

Junlin Wang, Jens Tuyls, Eric Wallace, and Sameer Singh. 2020. Gradient-based analysis of NLP models is manipulable. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 247–258, Online. Association for Computational Linguistics.

Shira Wein, Lucia Donatelli, Ethan Ricker, Calvin Engstrom, Alex Nelson, Leonie Harter, and Nathan Schneider. 2022. Spanish Abstract Meaning Representation: Annotation of a general corpus. In *Northern European Journal of Language Technology, Volume 8*, Copenhagen, Denmark. Northern European Association of Language Technology.

Shira Wein and Nathan Schneider. 2021. Classifying divergences in cross-lingual AMR pairs. In *Proceedings of the Joint 15th Linguistic Annotation Workshop (LAW) and 3rd Designing Meaning Representations (DMR) Workshop*, pages 56–65, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Sarah Wiegreffe and Ana Marasovic. 2021. Teach me to explain: A review of datasets for explainable natural language processing. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1. Curran.

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.

Kayo Yin and Graham Neubig. 2022. Interpreting language models with contrastive explanations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 184–198, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yuming Zhai, Gabriel Illouz, and Anne Vilnat. 2020. Detecting non-literal translations by fine-tuning cross-lingual pre-trained language models. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5944–5956, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Yuming Zhai, Aurélien Max, and Anne Vilnat. 2018. Construction of a multilingual corpus annotated with translation relations. In *Proceedings of the First Workshop on Linguistic Resources for Natural Language Processing*, pages 102–111, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Yuming Zhai, Pooyan Safari, Gabriel Illouz, A. Allauzen, and Anne Vilnat. 2019. Towards recognizing phrase translation processes: Experiments on english-french. *POLIBITS*, 60:27–33.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *ArXiv*, abs/1904.09675.

A   Appendices

A.1   Explanandum Training Details

The divergence ranking model is trained using the public implementation of Divergent mBERT (Briakou and Carpuat, 2020). Synthetic divergences are generated starting from the 5,000 top-scoring WikiMatrix sentences based on LASER score (i.e., seed equivalents). We fine-tune the "BERT-Base Multilingual Cased" model (Devlin et al., 2019) and set the margin equal to 5 as per the original implementation.

A.2   Subjective Measures
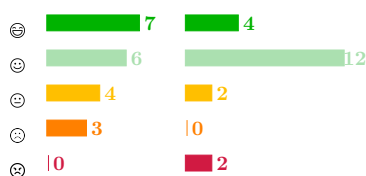
Table 4: Bilingual speakers' agreement ratings with the statement "Highlights were useful in helping me detect meaning differences in the compared texts" (left) and "I would like to use the highlights to help me detect critical errors" (right).

A.3   ChatGPT Details

Starting from a given English sentence, we asked ChatGPT to edit the sentence in a way that introduces a small meaning difference. Below, we include the different prompts we used:

- *"Can you edit a small phrase in sentence X to introduce a small meaning difference?"*

- *"Can you delete/add a phrase in sentence X to introduce a small difference in meaning?"*

- *"Can you edit a small phrase that makes sentence X more general/explicit?"*

We then reviewed the edits introduced by ChatGPT to make sure it indeed introduced small meaning differences. If we are not satisfied by a current edit, we ask again and potentially specify what parts of the sentence we would like to be edited.

A.4   User Study Interfaces

Below we include screenshots of our user study interfaces along with annotation examples for each of our evaluation tasks.

Figure 5: Instructions for application-grounded evaluation I: Annotation of Semantic Divergences.

Figure 6: Annotations of fine-grained semantic divergences (English-Spanish on the left and English-French on the righ) reflective of *explicitation* and *reduction* translation processes. Contrastive highlights subsume added content that is present in one language but missing from the other.

(a)

(b)

(c)

(d)

(e)

(f)

(g)

(h)

(i)

(j)

(k)

(l)

(m)

(n)

Figure 7: Annotations of fine-grained semantic divergences (English-Spanish on the left and English-French on the right) reflective of ***modulation*** translation processes. Contrastive highlights frequently subsume content that does not convey the exact same meaning across languages.

(a)

(b)

(c)

(d)

(e)

(f)

(g)

(h)

Figure 8: Annotations of semantically equivalent pairs (English-Spanish on the left and English-French on the right) that reflect syntactic divergences. Contrastive highlights surface *false positive* segments.

Figure 9: Instructions for application-grounded evaluation II: Critical Error Detection.

(a) **Major error** (negation)

(b) **Major error** (negation)

(c) **Major error** (negation)

(d) **Major error** (hallucination)

(e) **Major error** (hallucination)

(f) **Major error** (hallucination)

(g) **Minor error**

(h) **Minor error**

(i) **No error**

(j) **No error**

Figure 10: Annotations of accuracy errors in Portuguese translations of English texts. Contrastive highlights frequently surface meaning differences across the compared texts.