Towards Explainable Evaluation Metrics for Machine Translation

Christoph Leiter  
Natural Language Learning Group  
University of Mannheim  
B6 26, 68159 Mannheim, Germany  
christoph.leiter@uni-mannheim.de

Piyawat Lertvittayakumjorn  
Imperial College London  
PL1515@IMPERIAL.AC.UK

Marina Fomicheva  
University of Sheffield  
m.fomicheva@sheffield.ac.uk

Wei Zhao  
Heidelberg Institute for Theoretical Studies  
WEI.ZHAO@H-ITS.ORG

Yang Gao  
Royal Holloway, University of London  
GAOSTAYYANG@GOOGLE.COM

Steffen Eger  
University of Mannheim  
STEFFEN.eger@uni-mannheim.de

Editor: Ivan Titov

Abstract

Unlike classical lexical overlap metrics such as BLEU, most current evaluation metrics for machine translation (for example, COMET or BERTScore) are based on black-box large language models. They often achieve strong correlations with human judgments, but recent research indicates that the lower-quality classical metrics remain dominant, one of the potential reasons being that their decision processes are more transparent. To foster more widespread acceptance of novel high-quality metrics, explainability thus becomes crucial. In this concept paper, we identify key properties as well as key goals of explainable machine translation metrics and provide a comprehensive synthesis of recent techniques, relating them to our established goals and properties. In this context, we also discuss the latest state-of-the-art approaches to explainable metrics based on generative models such as ChatGPT and GPT4. Finally, we contribute a vision of next-generation approaches, including natural language explanations. We hope that our work can help catalyze and guide future research on explainable evaluation metrics and, mediately, also contribute to better and more transparent machine translation systems.

Keywords: evaluation metrics, explainability, interpretability, machine translation, machine translation evaluation

1. Introduction

The field of evaluation metrics for Natural Language Generation (NLG), especially machine translation (MT) is in a crisis (Marie et al., 2021). Despite the development of multiple high-quality evaluation metrics in recent years (e.g., Zhao et al., 2019; Zhang et al., 2020a;...
Rei et al., 2020; Sellam et al., 2020; Yuan et al., 2021; Rei et al., 2023a; Kocmi and Federmann, 2023a), the Natural Language Processing (NLP) community appears hesitant to adopt them for assessing NLG systems (Marie et al., 2021; Gehrmann et al., 2023). Empirical investigations of Marie et al. (2021) indicate that the majority of MT papers relies on surface-level evaluation metrics such as BLEU and METEOR (Papineni et al., 2002; Banerjee and Lavie, 2005), which were created two decades ago, a trend that may allegedly have worsened in recent times. These surface-level metrics cannot (even) measure semantic similarity of their inputs and are thus fundamentally flawed, particularly when it comes to assessing the quality of recent state-of-the-art MT systems (e.g., Peyrard, 2019; Freitag et al., 2022), raising concerns about the credibility of the scientific field. We argue that the potential reasons for this neglect of recent high-quality metrics include: (i) non-enforcement by reviewers; (ii) easier comparison to previous research, for example, by copying BLEU-based results from tables of related work (potentially a pitfall in itself); (iii) computational inefficiency to run expensive new metrics at large scale; (iv) lack of trust in and transparency of high-quality black box metrics. In this work, we concern ourselves with the last-named reason, and address the explainability of such metrics.

In recent years, explainability has become a crucial research area in AI due to its potential benefits for users, designers, and developers of AI systems (Samek et al., 2018; Vaughan and Wallach, 2021). For users of the AI systems, explanations help them make more informed decisions (especially in high-stake domains) (Sachan et al., 2020; Lertvittayakumjorn et al., 2021), better understand and hence gain trust of the AI systems (Pu and Chen, 2006; Toreini et al., 2020), and even learn from the AI systems to accomplish tasks more successfully (Mac Aodha et al., 2018; Lai et al., 2020). For AI system designers and developers, explanations allow them to identify the problems and weaknesses of the systems (Krause et al., 2016; Han et al., 2020), calibrate the confidence of the systems (Zhang et al., 2020b), and improve them accordingly (Kulesza et al., 2015; Lertvittayakumjorn and Toni, 2021).

Explainability is particularly desirable for evaluation metrics. Sai et al. (2022) suggest that explainable NLG metrics should focus on providing more information than just a single score (such as fluency or adequacy). Celikyilmaz et al. (2020) stress the need for explainable evaluation metrics to spot system quality issues and to achieve a higher trust in the evaluation of NLG systems. Explanations indeed play a vital role in building trust for new evaluation metrics. For instance, if explanations of the scores align closely with human reasoning and faithfully reflect a metric’s internal decision process, this metric may be more likely to gain acceptance within the research community (see §2.2 for a definition of faithfulness and related terminology).

1. As of now, there is no universally accepted definition of explainability in the AI community. In this work, we adapt the definition by Barredo Arrieta et al. (2020), which we discuss in §2.2.
2. As an illustrating example, Moosavi et al. (2021) distrust using the metric BERTScore (Zhang et al., 2020a) applied to a novel text generation task, as it assigns a score of 0.74 to a nonsensical output, which could be taken as an unreasonably high value. While BERTScore may indeed be unsuitable for their novel task, the score of 0.74 is meaningless here, as evaluation metrics may have arbitrary ranges. In fact, BERTScore typically has a particularly narrow range, so a score of 0.74 even for bad outputs may not be surprising (BERTScore provides a scaling method that solves these issues to some degree). Explainability techniques would be very helpful in preventing such misunderstandings, e.g., by showing that the features BERTScore attends to align with human judgement.
Towards Explainable Evaluation Metrics For MT

By contrast, if faithfulness is given and the explanations are counter-intuitive, users and developers might lower their trust and be alerted to take additional actions, such as trying to improve the metrics using insights from the explanations or looking for alternative metrics that are more trustworthy. Furthermore, explainable metrics can be used for other purposes: for example, when a metric produces a low score for a given input, highlighted words (a widely used method for explanation, see §4) in the input are natural candidates for manual post-editing (when working with reference-free metrics, see §2.1).

This concept paper aims at providing a systematic overview of the existing efforts in explainable MT evaluation metrics and an outlook for promising future research directions. The main focus of this work lies on explainable evaluation metrics for MT, although many of our observations and discussions can likely be adapted to other NLG tasks.

We make the following contributions, outlined by the structure of our paper:

**§2 Background & Terminology:** We give an overview of concepts important for explainable MT evaluation. We also highlight different goals and the audiences that require them.

**§4 Taxonomy:** We provide a structured overview of previous efforts in explainable MT evaluation. An overview is shown in Figure 3. The process of selecting works for this taxonomy is described in §3 Literature Review & Selection.

**§5 Future Work:** An analysis of underexplored research directions, and a collection of exemplary methods from other NLP domains to illustrate potential future paths. Here, we also discuss the usage of recent large language models (LLMs), like ChatGPT (OpenAI, 2023a) and GPT4 (OpenAI, 2023b).

We provide an overview of Related Work in §6 and a Conclusion in §7.

With this work, we aim to solidify the field of explainable MT metrics and provide guidance for researchers, e.g., metric developers, who aim to better understand and explain MT metrics, as well as users that want to employ explainability techniques to explain metric outputs. In the long term, we envision that our work will aid the development of improved MT metrics and thereby help to improve the quality of machine translations. Potentially, explanations can also aid other use cases such as translation selection and semi-automatic labeling (see §2.3). We also hope that our work will more generally inspire explainable NLG metric design.

2. Background & Terminology

In this section, we first introduce and relate definitions and dimensions of MT metrics and explainability, which we use in the later parts of the paper. Further, we collect goals and target audiences of explainable MT metrics.

2.1 Machine Translation Evaluation Metrics

MT metrics grade machine translated content, the hypothesis, based on ground truth data. Metrics can be categorized along several dimensions (e.g., Sai et al., 2022; Celikyilmaz et al., 2020). Here, we focus on a minimal specification that we will leverage later. A summary is shown in Table 1. Some parts of this section refer to explainability (see definitions in §2.2).
<table>
<thead>
<tr>
<th>Dimension</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Input type</td>
<td>Whether source, reference translation or both are used as ground truth for comparison</td>
</tr>
<tr>
<td>Granularity</td>
<td>At which level a metric operates: Word-level, sentence-level, document-level</td>
</tr>
<tr>
<td>Quality aspect</td>
<td>What a metric measures: Adequacy, fluency, etc.</td>
</tr>
<tr>
<td>Learning objective</td>
<td>How a metric is induced: Regression, ranking, etc.</td>
</tr>
</tbody>
</table>

Table 1: A typology of categorizations for evaluation metrics.

*Input type* We call a metric *reference-based* if it requires one or multiple human reference translations to compare with the hypothesis as ground truth, which can be seen as a form of supervision (e.g., Papineni et al., 2002; Zhao et al., 2019; Rei et al., 2022a). *Reference-free* metrics do not require a reference translation to grade the hypothesis (e.g., Zhao et al., 2020; Ranasinghe et al., 2020a; Belouadi and Eger, 2023). Instead, they directly compare the source to the hypothesis. In the literature, reference-free MT evaluation is sometimes also referred to as “reference-less” (e.g., Mathur et al., 2020) or “quality estimation” (e.g., Zerva et al., 2022). This dimension is important for explainable MT evaluation, as explainability techniques might need to consider which resources are available.

*Granularity* Translation quality can be evaluated at different levels of granularity: word-level, sentence-level and document-level. The majority of metrics for MT return a single sentence-level score for each input sentence (e.g., Zhao et al., 2019; Zhang et al., 2020a). Beyond individual sentences, a metric may also score whole documents (multiple sentences) (Jiang et al., 2022; Zhao et al., 2023). In the MT community, metrics that evaluate translations at the word-level (for example, whether individual words are correct or not) are also common (Turchi et al., 2014; Shenoy et al., 2021). Recently, there has been a tendency to compare metrics to human scores derived from fine-grained error analysis as they tend to correspond better to human professional translators (Freitag et al., 2021a). Metrics of higher granularity (e.g., word-level) provide more explanatory value than those of lower granularity, as they provide more details on which parts of the input might be translated incorrectly (e.g., Leiter, 2021; Fomicheva et al., 2021, 2022). Due to their prevalence, most techniques we describe in §4 target the explainability of metrics evaluated on sentence-level.

*Quality aspect* This refers to the properties a metric evaluates and measures, for example, human assigned adequacy (e.g., via direct assessment, see Graham et al. (2016), obtained from crowd-workers), fluency, or other aspects (such as relevance, informativeness, or correctness, mostly in other NLG fields) (e.g., Sai et al., 2022; Celikyilmaz et al., 2020). We use this dimension in §4, as metrics that report multiple quality aspects provide more information on specific translations, i.e., they have more explanatory value for potential users.

*Learning objective* Yuan et al. (2021) describe a further distinction based on the task
a metric is designed to solve (or how a metric is implemented) to achieve high correlation with human judgments. They identify the three tasks below:\[^3\]

- **Unsupervised Matching:** They place all metrics that match hypothesis and reference tokens in this dimension, e.g., BLEU (Papineni et al., 2002) and BERTScore (Zhang et al., 2020a).

- **Supervised Regression/Ranking:** These models use supervised regression to train a metric to predict continuous values that directly correlate with human judgments from hypothesis and ground truth tokens. Alternatively, a metric may be trained to rank hypothesis sentences. For example, Rei et al. (2020) propose a ranking setup in which they minimize a loss that becomes smaller the greater the distance between embedding representations of the two sentences become. Especially in the MT community, metrics trained on human scores (e.g., COMET ;Rei et al. 2022a) have become more and more dominant, while Belouadi and Eger (2023) argue that fully unsupervised metrics, i.e., metrics that were constructed without reference sentences, parallel data and human scores, are more widely applicable.

- **Text Generation:** Some metrics use the probability that a sentence is created by a text generation system with a paraphrasing or translation task as sentence-level score. For example, the probability of translating a source sentence into a hypothesis sentence with an MT model can be used as score. Yuan et al. (2021) place the metric PRISM (Thompson and Post, 2020) and their own metric BARTScore in this category. In contrast, some recent metrics are based on in-context learning and directly generate scores (e.g. Kocmi and Federmann, 2023a; Lu et al., 2023b).

The distinction of learning objectives is important for explainable MT evaluation, as the applicability of some explainers depends on the objective.

Sentence-level MT metrics also have multiple use cases; see Figure 1 (in the following, numbers written in brackets correspond to this figure). (1) One use case is to rate a single translated text, for example, in a scenario where a human user wants to check their own human\[^4\] translation or a translation they received by an MT model before using it in a downstream task (e.g., Murgolo et al., 2022; Zouhar et al., 2023). This setup can also be used to filter parallel corpora for high-quality training data of new MT models (e.g., Ramos et al., 2023; Peter et al., 2023). (2) A second, common use case is to rate a large number of translations of different texts by one MT system to obtain a system-level score (for example by averaging sentence-level scores) (e.g., Freitag et al., 2022). Then, different MT systems can be compared their system-level scores that were computed on the same corpus. (3) Another use case is to employ metrics in the training process of MT models, e.g., using reinforcement learning (e.g., Wu et al., 2018) or optimizing by computing derivatives with Gumbel-softmax (e.g., Jauregi Unanue et al., 2021). (4) The fourth use case is to use

---

\[^3\] Chen and Eger (2023) argue for metrics induced from natural language inference (NLI) as a conceptual framework.

\[^4\] Note, however, that humans might make different types of translation errors (Specia et al., 2018). Therefore, metric performance first needs to be evaluated for this use case.
the metrics in the decoding process of machine translation systems for reranking or with minimum Bayes risk decoding (Fernandes et al., 2022a).

The use cases can have different audiences and explainability might play a different role to them (see 2.2). For example, anyone might be a user that wants to rate a translation, but usually only certain user groups, such as translation experts or MT experts, want to rate a translation model.

<table>
<thead>
<tr>
<th>Use Case</th>
<th>Metric User</th>
<th>Translated Texts</th>
<th>Metric</th>
<th>Score</th>
<th>Accept, decline or rewrite the translated text</th>
<th>Compare, accept, decline or revise the translation model</th>
<th>New translation model</th>
<th>Select the best candidate translation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.) Rate a translation</td>
<td>anyone</td>
<td>Translated Text</td>
<td>Metric</td>
<td>Score</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.) Rate a translation model</td>
<td>translation experts, MT experts</td>
<td>Translated Texts</td>
<td>Metric</td>
<td>Scores</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3.) Train a translation model</td>
<td>MT experts</td>
<td>Translated Texts</td>
<td>Metric</td>
<td>Scores</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4.) Metrics for decoding/reranking</td>
<td>MT experts</td>
<td>Candidate Translations/Generation Probabilities</td>
<td>Metric</td>
<td>Scores</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 1: An overview of use-cases for MT metrics. In number 4, many candidate translations are produced for a single source. One of them is chosen as the main translation, via the metric.

2.2 Explainability

In this section, we (1) define explainability in the scope of this work, (2) discuss the definitions and roles of faithfulness and plausibility for explainable MT metrics and (3) contrast explainability with the terms robustness and fairness. Then, we (4) discuss the properties we use to structure and describe the taxonomy of prior works on explainable MT metrics.
Towards Explainable Evaluation Metrics For MT (see §4), namely explanation level, explanation type, model access and audiences. Based on (1), we decide on the scope of papers we include in the taxonomy in §3.

**Explainability & interpretability**  

Prior work has noted that the terms interpretability and explainability vary in definition and are sometimes used ambiguously throughout the literature (Barredo Arrieta et al., 2020; Jacovi and Goldberg, 2020; Vilone and Longo, 2021; Freiesleben and König, 2023). In this paper, we follow the notion of passive interpretability and active explainability (for example, Barredo Arrieta et al., 2020). In particular, we adapt the definition of Barredo Arrieta et al. (2020) as follows:

An MT metric is explainable if itself or some external instance, an explainer, can actively provide explanations that make its decision process or its output clearer to a certain audience.

This definition varies from Barredo Arrieta et al. (2020) in two aspects: (1) It includes external explainers (also called post-hoc methods) and (2) it includes explanations of the metric’s output. We note that (2) is different from explaining the metric’s decision process, as output explanations may not provide insights on how the metric works. Still, we include (2) in our explainability definition for two reasons. One is that there have already been multiple shared tasks and other works on metric explainability that only evaluate how plausible explanations explain the output of a metric (e.g., Fomicheva et al., 2020, 2021; Zerva et al., 2022; Xu et al., 2023). We want to accommodate these works in our definition. The other reason is that there are use cases for such explanations of the output (specifically for the goals of making metrics more accessible and semi-automatic labeling; see §2.3).

In contrast, a metric is interpretable if it can be passively understood by humans (for example, decision trees and k-nearest neighbors) (Barredo Arrieta et al., 2020). Modern MT metrics are mostly based on embeddings and black box language models, making them non-interpretable. The papers we categorize in this work generate explanations for MT metrics and their outputs, i.e., they consider explainability.

**Faithfulness & plausibility**  

An explanation is faithful if it accurately represents the reasoning process underlying a model’s prediction(s) (Jacovi and Goldberg, 2020). In contrast, an explanation is plausible if it convinces its audience that a model’s prediction is correct (Jacovi and Goldberg, 2021), i.e., if it aligns with human judgment. We illustrate the properties with the following example:

- **Source**: Things became complicated
- **Hypothesis**: Die Dinge bekamen kompliziert
- **Metric(Source, Hypothesis) = 0.8**
- **Explanation 1**: The word *became* is wrongly translated into *bekamen* instead of *wurden*. However, the meaning of the translation can easily be inferred. Therefore, on a scale of 0 to 1, a score of 0.8 is given.
- **Explanation 2**: The words *Die Dinge* have 8 letters. Therefore, a score of 0.8 is assigned.
In this example, the English source sentence was (wrongly) translated into the German hypothesis sentence. A metric grades the translation with a score of 0.8 and two explanations (here, natural language explanations) are provided. Comparing these explanations, explanation 1 is a more plausible description (to translation experts) of why the score of 0.8 was assigned, as it bases the score on translation errors. However, which of the explanations describes the metric more faithfully cannot be inferred from the text. For example, internally the metric could be relying on word lengths, making explanation 2 more faithful, even though it is linguistically non-sensical.

When explainability techniques are used for metrics, faithfulness is always desirable and often, depending on the goal of the explanations, required. Plausibility (to translation experts), on the other hand, is required for some goals and not important for others (see §2.3).

As a side note, “complete faithfulness” of an explainability technique may never be established because it would require capturing the full decision process, e.g., of neural networks, in human understandable descriptions. For example, Jacovi and Goldberg (2020) deem it as “a unicorn which will likely never be found”.

Fairness & robustness We also categorize methods that test the robustness and fairness of a model (here, a metric) as explainability techniques. This is because these tests produce explanations that will help certain audiences to understand a model’s behavior better (Barredo Arrieta et al., 2020).

Here, robustness measures how well a model can address execution errors, erroneous and malicious inputs and unseen data (e.g., Li et al., 2023a). In contrast, Li et al. (2023a) define fairness as the goal that a model should not have biases that result in unfair treatment for certain groups of people.5

In the MT metrics literature, this overlap in terminology is apparent for techniques that check how robust metrics are to input perturbations. Here, some works frame these as explainability techniques (e.g., Sai et al., 2021; Karpinska et al., 2022), while others frame them as robustness tests (e.g., He et al., 2023; Chen and Eger, 2023); see §4.

Robustness could be measured coarse-grainedly with a single score, such as the minimal number of adversarial perturbations required to successfully attack a metric, or fine-grainedly, by using predefined perturbation templates (e.g., Sai et al., 2021) or attackers that will attack with perturbations of specific properties (e.g., Chen and Eger, 2023). While a single score can also explain whether metric behavior is robust, susceptibility scores to various properties allow to assess metrics in greater detail. The same is true for fairness (Sun et al., 2022).

Explanation level Explainability techniques can be distinguished into decision-understanding and model-understanding (e.g., Gehrmann et al., 2020; Holzinger et al., 2022). Decision-understanding methods are described as explaining the decision process for a spe-

---

5. There are many related and partly conflicting definitions and terms in explainable AI research (e.g., Vilone and Longo, 2021) and the given definitions are not yet firmly established in the literature. For example, Sharma et al. (2020) define explainability, robustness and fairness as exclusive properties. The main focus of our work is to build a taxonomy of methods for explainable MT metrics and the chosen definitions allow us to reasonably relate existing work.
cific model output with respect to a specific input while model-understanding methods are
described as explaining general properties of a model that steer its decisions. For exam-
ple, a decision-understanding method could show that a metric assigned a score of 0.5,
because word *xyz* was translated incorrectly. In contrast, a model-understanding method
could show that a metric will always return a score of 1, given certain conditions. Often,
decision-understanding is also referred to as *local* and model-understanding as *global* (e.g.,
Lipton, 2016; Guidotti et al., 2018; Doshi-Velez and Kim, 2017; Danilevsky et al., 2020).

*Explanation type*  Further, we distinguish explanation types, i.e., the output format of
the explanations (e.g., Wiegreffe and Marasovic, 2021). One example are natural language
explanations that describe the model output or its internal decision process with free text.
In the case of MT metrics, this could be a sentence like “The metric returned a score of
0.6 for the translation, as it correctly captures all details of the source sentence but has
severe grammatical errors at positions [...]”. There are many possible explanation types
(e.g., Barredo Arrieta et al., 2020; Wiegreffe and Pinter, 2019), such as, feature importance
explanations or counterfactuals. For our taxonomy, we select those that occur in our target
papers (see §3). The definition of each explanation type is paired with the description of
methods in §4.

*Model Access*  Most taxonomies of explainability techniques consider the required level
of model access (e.g., Barredo Arrieta et al., 2020; Linardatos et al., 2021; Madsen et al.,
2022). Here, techniques that require no access (that can work with any *black box model*)
are called *model-agnostic*. Techniques that require access to model parameters are called
*model-specific*, as only specific model architectures satisfy the requirements of the method.
Due to the prevalent role of transformer architectures, many modern evaluation metrics use
*embeddings* and *attention* mechanisms and allow for *gradient* computations. This allows to
apply many techniques that are specific to these architectures to such metrics. A special
type of model-specific explanations are those that are given by a model (here a metric)
itself. We call these metrics “explainable by design”, as they are designed in a way that
provides more details on their explanations or decision process. Examples for MT metrics
in this category are unsupervised matching metrics like BERTScore (Zhang et al., 2020a)
and MoverScore (Zhao et al., 2019), as they compute token-level similarity scores that can
be used as explanations of the sentence-level score (Leiter, 2021; Fomicheva et al., 2021).
Finally, there are also *interpretable models* which are easy to understand for humans such
as Naive Bayes and decision trees (Freitas, 2014; Arya et al., 2019). We can normally
obtain decision-understanding explanations of these models at the same time as predictions
(for example, the corresponding path in the decision tree), while the models can also be
considered explanations of themselves (for example, the trained decision tree itself). We do
not use the dimension of *interpretable models* in our taxonomy (see §3) as none of the state-
of-the-art metrics falls into this category. Model-agnostic techniques require least model
access. Further, model-specific techniques, which use the gradient, embeddings or attention
weights require more access. Finally, models that are explainable by design require access to the full model to generate explanations.\footnote{Note that besides interpretable models, none of the access levels inherently leads to faithful explanations, i.e., explanations that correctly describe the internal processes of a model. Hence, this property needs to be considered separately (e.g., Jacovi and Goldberg, 2020).}

**Audiences** Prior works in explainability point out that different audiences require different types of explanations (e.g., Barredo Arrieta et al., 2020; Sharma et al., 2020). For example, non-experts might have difficulties understanding explanations given in the domain language of experts. Barredo Arrieta et al. (2020) distinguish *model users, affected users, developers, regulatory entities, and managers*. Users of MT metrics are often experts in MT that want to evaluate or improve MT systems, i.e., MT developers. Another group of users are translation experts that want to compare MT systems and rate translations for usage in daily tasks (for example for post-editing). A third group of users might be non-experts that want to use the models to check their own translations or machine generated translations. Affected users might be the ones that receive translations that were graded by a metric or generated by a model that was supervised or tested with the metric. For example, Fernandes et al. (2022a) employ MT metrics in the decoding process of MT models. We consider users of those MT models as affected users. Developers are the designers of MT metrics. Regulatory entities and managers have an interest in enforcing guidelines. We describe the relation of the audiences towards goals of explainable MT metrics in §2.3 and associate audiences with explanation types in §4.

### 2.3 Goals of Explainable MT Metrics

In this section, we discuss goals that require specific focus for explainable MT evaluation. We orient ourselves on common goals proposed by Lipton (2016), Barredo Arrieta et al. (2020) and Jacovi and Goldberg (2021), and identify four specific ones that match the literature we identify in §3 and use cases that might commonly occur in our field of research (see Figure 2). Note that goal 1 and 4 address metric internals, while goal 2 and 3 mostly consider explanations of metric outputs (and with this, also the quality of the input translations). We follow with a short description of each of them below.

- **Goal 1: Diagnose and improve metrics.** By explaining why a metric predicted a certain score for a machine translation, where humans would assign a different score, developers might understand its weaknesses (e.g., Kaster et al., 2021; Sai et al., 2021). This can enable architectural changes or the selection of different data sets, leading to metric improvement. Likewise, explaining the whole metric can unveil whether it follows desired general properties or might otherwise be led astray by carefully crafted adversarial inputs. This goal requires explanations to be *faithful* to some extent, such that developers can understand the actual faults of the explained metric. *Plausibility* on the other hand may be harmful and trick developers into believing that their metric is working, therefore, it should be evaluated separately (Jacovi and Goldberg, 2020). Goal 1 is especially important for the audience *metric developers*, as they need to understand the shortcomings to improve on them.
Towards Explainable Evaluation Metrics For MT

**Goal 2:** Make metrics more expressive and accessible. A metric that assigns a score on a sentence-level is difficult to understand, especially for users who are not experienced with MT and MT metrics. If further information about a metric’s output is provided, such as which words it considers incorrect, the metric may become more accessible and thus foster trust (which may lead to more widespread adoption of the metric). Exemplary works in this direction are Fomicheva et al. (2021) and Bao et al. (2023). Accessible metrics might also help with use cases like translation error correction in the downstream tasks of language learning (non-experts) or human translation services (translation experts) (e.g., Zouhar et al., 2023). In these latter two use cases, a high average explanation plausibility (to translation experts) and a strong metric performance are important. This is because, in cases where (1) the metric output is correct and (2) the explanation of the output is plausible to an expert, the explanation very likely correctly describes the shortcomings and strengths of the translation (on the level of an expert). Faithfulness is not necessary in this scenario, because the quality of the explanation in describing the translation already follows from (1) and (2). However, faithfulness is still desirable, as the internal decision process of a strong metric might support giving correct explanations (see §2.2). Goal 2 is important for the audience of metric users. These could, for example, be non-

---

7. To foster trust, it might be sufficient, but unethical, if explanations of any metric (also weak ones) are plausible to its users (not necessarily to experts) (e.g., Jacovi and Goldberg, 2020; Jin et al., 2023). The problem is that users could be tricked to believe into the performance of weak metrics or adapt the mindset of wrong explanations that have believable arguments. Hence, benchmarks of metrics and explanations in their correlation to human expert quality scores and explanations are important. Generally, users should be informed that hallucinations, where an explanation is untrustworthy, might occur.
experts that want to understand specific metric outputs or developers of MT systems that want to understand if a metric is a good fit for their tasks (see Figure 1).

- **Goal 3: Support Semi-Automatic labeling.** Fine-grained (human) annotations like word-level translation error labels (e.g., Fomicheva et al., 2021) are especially difficult to obtain. Obtaining automatic explanations to aid human annotators could boost their efficiency tremendously (e.g., Desmond et al., 2021). To our best knowledge, none of the current works has applied explainers in these settings for metrics. Semi-automatic labeling could also help with post-editing of translations (e.g., Specia and Farzindar, 2010) to guide how sentences should be adapted. Goal 3 is important for metric developers that want to create such word-level translation error data sets and metric users who want to correct translations. The requirements for *faithfulness* and *plausibility* are the same as for goal 2.

- **Goal 4: Checking for social biases.** Learned metrics might be biased towards certain types of texts used during training (e.g., Sun et al., 2022). Biases could be detected by observing explanations where sensitive attributes (for example, gender, political or racial aspects) are flagged to be important (Lipton, 2016). For example, explanations could show that a metric considers male names more important for a translation than female names in the same scenario. Goal 4 can be important for all audiences. Affected users might be treated unfairly if biases occur in a metric. Metric users, metric developers, regulatory entities and managers might want to prevent this from happening to conform with given regulations. The requirements for *faithfulness* and *plausibility* are the same as for goal 1.

### 3. Literature Review & Selection

To build the taxonomy in §4, we survey literature which either explicitly state that they address the explainability of MT metrics or propose techniques that make MT metrics or their outputs clearer to understand by certain audiences (techniques that provide explanations). We restrict our search and taxonomy to works that consider neural network based MT metrics. We conducted an initial literature search by selecting 11 “seed papers”. Then, we leveraged the API of Semantic Scholar\(^8\) to query new papers citing the work of the seed papers and papers citing those papers (citation tree at depth 2). This amounted to 383 papers. Next, we filtered the papers by keywords, leaving 177 papers. Then, we manually annotated whether their contents fit our definitions. Thereby, we identified 12 papers that consider the explainability of MT metrics. In later stages, we progressively added further work released after the initial search and related works that we identified from other sources. In total, we describe 37 papers, some of which are not shown in Figure 3 due to space constraints, but discussed throughout the chapter.

Following §2.1, we consider higher granularity scores (e.g., word-level) as explanations for lower granularity metrics (e.g., sentence-level), i.e., *feature importance* explanations. Due to the high number of existing word-level approaches, for this explanation type we only consider works that explicitly employ word-level scores to explain sentence-level scores.

\(^8\) https://www.semanticscholar.org/
Likewise, following §2.1, metrics that report multiple quality aspects such as adequacy or fluency provide more explanatory value on why a translation is good or bad than metrics that return a single score. For the quality aspect explanation type, we make an exemplary selection of techniques derived from recent metrics that leverage generation probabilities (e.g., Thompson and Post, 2020; Yuan et al., 2021).

We also include a few non-MT works into our taxonomy, which show how some aspects that are currently not addressed by the MT related works might be tackled.

4. Taxonomy

Based on the background described in the previous sections and the literature search we conducted, we create a taxonomy of prior works that consider the explainability of MT metrics. This taxonomy is shown in Figure 3. The taxonomy contains five explanation types based on the selected papers' contents: quality aspects, feature importance, fine-grained errors, perturbation robustness and linguistic properties. We structure this section based on these explanation types. Figure 4 shows an example for each explanation type. For each type, we describe the techniques applied by the previous works and how they relate to each other. We begin with explanation types that explain specific samples and move to techniques that explain metric behavior in general.

Quality aspects (Row 1 in Figure 3) This category comprises works that present explanations of the metric output by producing separate scores for different aspects of the translation. Most methods that we list develop generation-based metrics that are explainable by design (“model specific” in Figure 3). This means, they do not use an external explainer that could be applied to other metrics. BARTScore (Yuan et al., 2021) uses the autoregressive generation model BART (Lewis et al., 2020) to predict the average word generation probability of a sentence $B = (b_1, ..., b_m)$, given a sentence $A = (a_1, ..., a_n)$:

$$BARTScore(A \rightarrow B) = \sum_{t=1}^{m} \log p(b_t|b_{<t}, A, \theta)$$

This formula sums over the log-probability of token $b_t$ being generated by BART with parameters $\theta$ given all previous tokens of $B$ and all tokens of $A$. Depending on whether $A$ is the ground truth sentence and $B$ the hypothesis (or vice-versa), Yuan et al. (2021) state that different aspects of the translation are evaluated. For example, they associate the generation direction of $A$:source $\rightarrow$ B:hypothesis with the properties coherence and fluency. This idea of using the generation probabilities in multiple directions was first introduced in PRISM (Thompson and Post, 2020), but Yuan et al. (2021) formalized the connection of the directions to known quality aspects. DATScore (Kamal Eddine et al., 2023) extends BARTScore by using a multilingual model and further generation directions (additionally using the source sentence). The metric GPTScore (Fu et al., 2023) provides a suite of prompts for language models that emphasize specific aspects when requesting a translation or paraphrase. Then, similar to BARTScore, the generation probabilities are averaged over all tokens. For example, they use the prompt “Rewrite the following text to make it more grammatical and well-written: [reference or hypothesis] In other words,” to generate
Figure 3: Overview of papers addressing the explainability of machine translation evaluation. The graphic structure is adapted from a survey on post-hoc methods for explainable NLP by Madsen et al. (2022). The rows show the types of explanations returned by the respective methods. They are ordered from decision-understanding to model-understanding. Decision-understanding techniques explain specific outputs of a metric, while model-understanding techniques describe general properties. The columns show the required level of model access that each explainability technique needs. Fields, where boxes overlap are colored in a darker grey. Note that most work listed under feature importance was developed as part of the Eval4NLP21 (Fomicheva et al., 2021) and the WMT22 QE (Zerva et al., 2022) shared tasks. \(^N\) indicates that a technique is not directly from the MT domain. \(^R\) indicates that a technique was proposed to tackle robustness. \(^F\) indicates that the technique was proposed to explore fairness aspects. \(^E\) is set when the papers’ authors themselves perceived that their method tackles explainability.

<table>
<thead>
<tr>
<th></th>
<th>Less Information</th>
<th>More Information</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Post-hoc</td>
<td>Intrinsic</td>
</tr>
<tr>
<td></td>
<td>Gradient</td>
<td>Model Specific</td>
</tr>
<tr>
<td></td>
<td>Attention</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Weights</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Embeddings</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Black-box</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Quality Aspects</th>
<th>(Kaster et al., 2021 (^E))</th>
<th>(Golovneva et al., 2023 (^N))</th>
<th>(Thompson and Post, 2020)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Feature Importance</td>
<td>(Rei et al., 2022b (^E))</td>
<td>(Leiter et al., 2022) (^E)</td>
<td>(Lu et al., 2023a,b)</td>
</tr>
<tr>
<td>Fine-Grained Errors</td>
<td>(Kaster et al., 2021 (^E))</td>
<td>(Fernandes et al., 2022b (^E))</td>
<td>(Kapir and Carpuat, 2021 (^E))</td>
</tr>
<tr>
<td>Perturbation Robustness</td>
<td>(Rei et al., 2022b (^E))</td>
<td>(Kabir and Carpuat, 2021 (^E))</td>
<td>(Rubino et al., 2021 (^E))</td>
</tr>
<tr>
<td>Linguistic Properties</td>
<td>(He et al., 2023 (^E))</td>
<td>(Chen and Eger, 2023 (^E))</td>
<td>(Wang et al., 2022 (^E))</td>
</tr>
</tbody>
</table>

\(^E\) indicates that the papers’ authors themselves perceived that their method tackles explainability.
Towards Explainable Evaluation Metrics For MT

<table>
<thead>
<tr>
<th>Example Data:</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Ground Truth:</strong> “Things became complicated”</td>
</tr>
<tr>
<td><strong>Hypothesis:</strong> “Die Dinge bekamen kompliziert”</td>
</tr>
<tr>
<td><strong>Score:</strong> 0.8</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Quality Aspects</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adequacy: 0.7</td>
</tr>
<tr>
<td>Fluency: 1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Feature Importance</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.5 0.1 0.5 → “Things became complicated”</td>
</tr>
<tr>
<td>0.5 0.4 0.1 0.5 → “Die Dinge bekamen kompliziert”</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Fine Grained Errors</th>
</tr>
</thead>
<tbody>
<tr>
<td>“Things &lt;sense&gt;became&lt;/sense&gt; complicated”</td>
</tr>
<tr>
<td>“Die Dinge &lt;sense&gt;becamen&lt;/sense&gt; kompliziert”</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Perturbation Robustness</th>
</tr>
</thead>
<tbody>
<tr>
<td>Numeric: “3 Dinge bekamen kompliziert” → lower score</td>
</tr>
<tr>
<td>Order: “kompliziert die Dinge bekamen” → lower score</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Linguistic Properties</th>
</tr>
</thead>
<tbody>
<tr>
<td>Metric score = syntax + semantics + lexical</td>
</tr>
</tbody>
</table>

Figure 4: Explanation types shown in §3. For each type we provide a hypothetical example. For this we assume the exemplary translation that is shown in the top-left box. Here, the polysemic word ”became” is translated with the wrong word sense ”to receive” instead of ”to get/to develop into”. The correct translation is ”wurden”. The example for perturbation robustness, perturbs the hypothesis sentence.

a paraphrase whose probabilities they use for a fluency score. The example specifies reference/hypothesis to describe that the metric checks either the generation probability of the hypothesis given the reference, or vice-versa.

Another way to gather fine-grained scores is to apply suites of metrics (each of which considers different aspects), like ROSCOE (Golovneva et al., 2023), an embedding based metric suite, which has been proposed to grade step-by-step reasoning processes (column “embeddings” in Figure 3).

Besides adding explanatory value per se, quality aspect explanations have two more potential use cases for explainable metrics: (1) separate quality scores could be aggregated, to compute single, decomposable sentence-level scores and (2) scores could be aggregated over data sets to check which aspects a metric favors the most. For example, Kaster et al. (2021) and Opitz and Frank (2022) also assign quality aspect scores, but mostly use them for model understanding; see linguistic properties explanations for more details.

In the example presented in Figure 4, we report fluency and adequacy as quality aspects. We assign a fluency score of 1 on a scale from 0 to 1, as the translation is fluent and an adequacy of 0.7 as the translation of “became” is incorrect.
Explanation Type: Quality Aspects

Goals: Quality aspect approaches could be used for metric accessibility and semi-automatic labeling. The usage for metric diagnosis is limited, as the approaches are often based on models that are explainable by design and might not be faithful (the models generate their own explanations). The current approaches cannot be used to detect social biases.

Audiences: Quality aspects like fluency and adequacy follow simple definitions, such that these metrics could be used by non-expert and expert users.

Limitations: The generation directions might not fully match the quality aspects. Also, the usage of prompting leaves it to another black box setting whether the model picks up the desired properties or not. Lastly, the explanatory value of quality aspect approaches per se is rather small, as they only allow to reason, “the rated translation is good because of aspect x” and to guess about the model’s reasoning. As none of the current works specifically targets explainability, they do not consider faithfulness. Plausibility is automatically considered in standard evaluation schemes with human judgments.

\[a.\] Kaster et al. (2021) and Opitz and Frank (2022) are considered at a later point.

Feature importance (Row 2 in Figure 3) Feature importance techniques assign scores to all features in the input of a metric (here the input tokens) that express their importance to the output of the metric.\(^9\) Fomicheva et al. (2022) note that high feature importance scores should correspond to translation errors, as humans pay most attention to errors when rating a translation. They build on the idea that when humans evaluate translations, they often focus on the errors that they can identify on a word- or phrase-level (Freitag et al., 2021a). Then, Fomicheva et al. (2022) evaluate how well feature importance is correlated with human word-level error annotations, i.e., the plausibility of the explanations.

We note that, depending on the explainability method, these scores might be inverted, following the interpretation “errors are not important for a metric to achieve a high score”.\(^{10}\) Two shared tasks for the explainability of reference-free metrics have originated from the setting in Fomicheva et al. (2022) so far (Fomicheva et al., 2021; Zerva et al., 2022). As the respective papers already give an overview of the participating methods, we will only briefly describe the methods and set our focus to describe the model access they require. We describe the model access levels of Figure 3 from left to right. Approaches that work directly on black box models are explored in Fomicheva et al. (2022), Eksi et al. (2021) and Treviso et al. (2021). They leverage model agnostic explainers such as LIME (Ribeiro et al., 2016) and SHAP (Lundberg and Lee, 2017). These methods check how a metric’s sentence-level score changes when the input sentence is perturbed at selected positions. While the results

\(^9\) Other names include relevance scores or attribution scores.

\(^{10}\) Most gradient and attention based approaches return high scores for errors (e.g., Treviso et al., 2021; Rubino et al., 2021), for embedding-based approaches it depends on whether cosine-similarity or -distance is used (e.g., Leiter, 2021; Tao et al., 2022), for attribution based techniques using LIME (Ribeiro et al., 2016) or SHAP (Lundberg and Lee, 2017), low scores are returned for errors (e.g., as part of the baseline implementations in Eval4NLP21: https://github.com/eval4nlp/SharedTask2021/tree/main/baselines).
correlate with the human annotations to some degree, techniques that use model access performed better in the shared tasks. Gradient based methods are explored by Fomicheva et al. (2022), Eksi et al. (2021), Treviso et al. (2021) and Rei et al. (2022a). These papers use methods like Integrated Gradients (Sundararajan et al., 2017) to determine the effect each input token has on the output. Attention values perform especially good for Treviso et al. (2021) and Rei et al. (2022b), leading to winning techniques for most language pairs in both shared tasks, i.e., producing the most plausible explanations. In specific, Treviso et al. (2021) succeed by scaling attention weights by the norm of value vectors and filtering for attention heads that have the highest scores in detecting errors. Rei et al. (2022b) build on their approach and (1) instead scale attention weights with the l2-norm of the gradient of value vectors and (2) designing an ensemble method for attention heads. In their final submission, they ensemble the best performing attention heads over all layers. Word-level contextualized embeddings are used by Leiter (2021), Tao et al. (2022) and Azadi et al. (2022). Specifically, they leverage the word-level embedding distances that are inherently computed in token-matching metrics (see §2.1) and use them as importance scores. For model-specific works, Rubino et al. (2021) jointly fine-tune their metrics on the word-level and sentence-level using synthetic data, to directly output the computed word-level scores. Kabir and Carpuat (2021) use LIME with Transquest (Ranasinghe et al., 2020b) and merge the word-level importance scores with additional word-level scores obtained from a pure word-level model. After the shared tasks, Fernandes et al. (2022b) enhance explainability techniques (in case of MT they use a gradient technique) by adding learnable parameters and employing a student teacher mechanism to improve the correlation to human word-level scores. Rei et al. (2023b) extend the work of Rei et al. (2022b) by using reference-based metrics and importance scores of the references. They construct a data set from the WMT21 MQM annotations (Freitag et al., 2021b) by assuming that feature attribution scores link to annotated error spans. Evaluations on this data set show that the winning explainability techniques of the shared task perform best and that references can improve the plausibility. Notably, they also show that explanations can help to detect critical errors in translations. Based on this, they employ the explanations for metric diagnosis, to test whether their metrics are more prone to ignoring certain types of critical errors. In their limitations, they note that the explainability techniques they use have shown to be faithful in some scenarios, but also note that they did not specifically test whether they are faithful for their architecture. Finally, we note that many more metrics can provide feature-importance like explanations by design. Similar to the base metrics used by Leiter (2021), Tao et al. (2022) and Azadi et al. (2022), most unsupervised matching metrics could be used to determine importance scores based on word- or embedding-similarities. Further, the per-word generation probabilities in text generation metrics like PRISM or BARTScore can also be interpreted as feature importance (they can also be interpreted as the confidence of the generation models; Fomicheva et al. 2020).

In the example presented in Figure 4, high feature importance scores are assigned to words that were translated correctly, i.e., “Things/complicated” and lower scores indicate words that are wrong. Leiter et al. (2022b) take a different approach on using feature importance techniques for the

---

11. Rubino et al. (2021) won the constrained track of Eval4NLP21, which did not allow for word-level finetuning.
goal of metric improvement. Instead of using the feature importance scores in a task that might involve human explainees, the authors aggregate and directly integrate the feature importance scores into metric scores. This follows the intuition that the explanations carry additional information that can be beneficial to the sentence-level correlation of a metric with human judgments.

**Explanation Type:** Feature Importance

**Goals:** The discussed feature importance approaches are tested for plausibility, so strong approaches could be used for metric accessibility/improvement and semi-automatic labeling. The usage for metric diagnosis is currently limited, as only samples are explained, and approaches have not been tested for faithfulness. The current approaches have not been used to detect social biases.

**Audiences:** The relation of low/high scores and translation errors is intuitive, such that these techniques could be used by non-expert and expert users. Metric developers could use the methods to create new data sets (by supporting the fine-grained human annotation of errors) or to directly incorporate them into metrics (Leiter et al., 2022b).

**Limitations:** For all approaches, it is not specifically tested whether the explanations are faithful. Notably, Rei et al. (2023b) consider that faithfulness is probably given in their case but note the missing tests as limitation. Second, certain translation errors cannot be easily captured by highlighting specific words. For example, the word-level scores cannot represent cases where the MT fails to explicitly express a grammatical feature that is omitted in the source language but is required to be explicit in the target language (for example, the use of articles when translating from Russian into German). Third, the explanation type does not provide correspondence between highlighted error words in the source and target language.

---

*Fine-grained errors (Row 3 in Figure 3)* Some recent works detect errors on a word-level and leverage their severity, number and type to compute a sentence-level score (Lu et al., 2023a,b; Bao et al., 2023; Xu et al., 2023; Fernandes et al., 2023; Kocmi and Federmann, 2023a). This follows the ideas of the MQM annotation scheme (Lommel et al., 2014; Freitag et al., 2021a). This follows the ideas of the MQM annotation scheme (Lommel et al., 2014; Freitag et al., 2021a) in which human translation experts evaluate translations in a similar manner. Freitag et al. (2021a) have shown that this annotation scheme improves over previous annotation methods such as direct assessments (Graham et al., 2015). While the main goal of these works is to provide sentence-level scores, they are also explainable-by-design and the generated error-labels carry explanatory value (model-specific in Figure 3). In specific, BARTScore++ (Lu et al., 2023a) distinguishes explicit errors that are easy to detect and implicit errors. They detect the explicit errors based on BART generation probabilities. Then they generate a sentence-level score based on an additional improved hypothesis sentence in multiple BARTScore calculations. Lu et al. (2023b) also combine fine-grained errors into a score. They first request ChatGPT to generate explanations based on MQM annotations (erroneous words, error type and error majority). In a second step, they ask a model to calculate a score from these. This is a form of chain-of-thought prompting (Wei
et al., 2022), a prompting technique where a language model is instructed to first generate an explanation/reasoning before returning a result. Lu et al. (2023b) find that their approach is competitive to other state-of-the-art metrics, especially with newer versions of ChatGPT. Xu et al. (2023) fine-tune InstructScore, a LLaMA model (Touvron et al., 2023) that generates a text containing error majority, error spans, error types (MQM) and an explanation for every error in an input sentence. The generated texts are similar to Lu et al. (2023b) but add more details on each error. Similarly, Xu et al. (2023) construct a score from the count and severity of errors. They obtain synthetic training data from GPT4 (OpenAI, 2023b), analyze failure modes of InstructScore during training and later optimize the metric on data sets with real machine translation outputs and annotated ground truth scores. Their model shows state-of-the-art results for reference-free sentence-level evaluations. Further, the detailed error annotations can be used as fine-grained explanations. Contrary to Instructscore, instead of fine-tuning, Fernandes et al. (2023) propose AutoMQM a prompting technique that leverages few-shot learning (Brown et al., 2020) and chain-of-thought prompting to generate MQM error annotations. They use it to prompt PALM 2 (Anil et al., 2023) and achieve competitive results for system- and sentence-level evaluation. In a similar way, Kocmi and Federmann (2023a) design the GEMBA-MQM prompting technique and prompt GPT4 in a few-shot and chain-of-thought setting to produce MQM annotations. A main distinction is that AutoMQM dynamically selects few-shot examples from pools of example translations and ratings based on the input. In contrast, GEMBA-MQM uses only a few, pre-selected examples and is therefore easier to apply to further language pairs. Both, AutoMQM and GEMBA-MQM calculate their final scores based on the detected errors with MQM heuristics (Freitag et al., 2021a).

The WMT 2023 shared task on quality estimation (Blain et al., 2023) also includes a new sub-track where the submitted metrics should detect fine-grained errors and their severity on MQM data, to facilitate the development of more explainable metrics. The participants explore two main directions: (1) detection with GPT4 (Rei et al., 2023a)12 and (2) fine-tuning models for error span detection, either based on word-level supervision (Geng et al., 2023; Li et al., 2023b) or with multi-task objectives (Rei et al., 2023a). Notably the two winning submissions from Geng et al. (2023) and Rei et al. (2023a) are ensemble approaches, where Rei et al. (2023a) also use pseudo-references from DeepL and Google Translate (if their quality is better than a threshold, as measured by another metric) (Blain et al., 2023). Similarly, but not related to the shared task, Bao et al. (2023) fine-tune a seq2seq model to predict addition and omission errors on a word-level.

In the example presented in Figure 4, a tag “sense” flags the wrong word sense of “became”. Notably, all fine-grained error approaches we describe are very recent, which is potentially caused by the adaptation of MQM in the WMT metrics shared tasks and the recent successes of LLMs.

12. Another participating team KUNMT did not submit a paper.
Explanation Type: Fine-grained errors

Goals: These fine-grained error approaches can be used for increased accessibility and semi-supervised labeling, as the methods model plausible explanations (some of them by testing the alignment of scores aggregated from explanations with human MQM scores). The usage for metric diagnosis is limited, as only samples are explained, and explanations have not been tested for faithfulness. The current approaches cannot be used to detect social biases.

Audiences: Flagged translation errors are simple to understand, such that these techniques could be used by non-expert and expert users. Metric developers might use them to create new data sets to train new metrics.

Limitations: The current approaches (besides Xu et al. 2023) do not have the direct goal to provide explanations that make the sentence-level scores more descriptive. Instead, they aim to build sentence-level metrics from LLM outputs. As such, there might be open opportunities in leveraging the explanations provided by these models to explain to a human explainee. Another limitation are missing evaluations of faithfulness.

Perturbation robustness (Row 4 in Figure 3) The methods described up to now addressed decision-understanding techniques (Gehrmann et al., 2020). This paragraph and the next describe model-understanding techniques (we describe these terms in §2.2). Many recent works evaluate the robustness of MT metrics (Sai et al., 2021; Karpinska et al., 2022; He et al., 2023; Sun et al., 2022; Chen and Eger, 2023; Vu et al., 2022). Robustness evaluation qualifies as model-understanding, as it usually infers properties that describe a model (see §2.2). All works in this category are structured into “black box” in Figure 3.

Sai et al. (2021) propose perturbation checklists, following Ribeiro et al. (2020), that evaluate how susceptible metrics are towards predefined types of perturbations as evaluation criterion. This allows developers to check whether all invariances that are required for a specific task are fulfilled. In particular, they compare the change in the metric’s score with the change in score a human would assign after the perturbation. These templates encompass dropping or adding of context as well as negations. They show that perturbation checklists allow to pick metrics that are strong with respect to specific properties. For example, their experiments indicate MoverScore (Zhao et al., 2019) would capture fluency better than BERTScore (Zhang et al., 2020a) due to its ability to cope with the jumbling of words. The checklists allow for more fine-grained assessment of automatic evaluation metrics, which exposes their limitations and supports the design of better metrics. Chen and Eger (2023) create a similar, but fully automatic variant of this evaluation. They design several perturbation types (framed as adversarial attacks) such as addition, omission and negation. Then they measure whether a tested metric prefers a meaning preserving, but lexically dissimilar paraphrase of a source/reference sentence over a minimally attacked but lexically similar version of the source/reference sentence. This means they check whether the metric score of the attacked example is successfully lower than the metric score of a paraphrase. One of their findings is that standard metrics often have problems in detecting wrong names or numbers in translations. Karpinska et al. (2022) create DEMETR, a data
set of machine translations that were perturbed using several perturbation types. They propose to test a metric’s preference between these perturbations and the original translation. Their task is easier for a metric to compute than Chen and Eger (2023), as the preference is checked directly between attacked and non-attacked version, while Chen and Eger (2023) use a dissimilar paraphrase instead of the non-attacked version. Also, they define more perturbation types than Chen and Eger (2023), but require manual annotation for some of them. He et al. (2023) also compare the original translation with certain perturbation types. They differ from the other two papers in terms of domains they are applying their method to and in the perturbation types that they propose. Vu et al. (2022) use character-level adversarial attacks to evaluate the robustness of several BERTScore variants.

The WMT21 (Freitag et al., 2021b), WMT22 (Freitag et al., 2022) and WMT23 (Freitag et al., 2023) metrics shared tasks introduce a similar setting for robustness tests. In WMT21, the organizers provided the participants with a challenge set composed of translations that were perturbed in a specific way, to test whether metrics reduce their score as expected when presented with respective phenomena. In WMT22 and WMT23, participants of a challenge set subtask (Alves et al., 2022; Avramidis and Macketanz, 2022; Chen et al., 2022; Amrhein et al., 2022; Lo et al., 2023; Amrhein et al., 2023; Avramidis et al., 2023) designed data sets posing challenges in a similar manner. In the example presented in Figure 4, we show tests like those performed by techniques of this category. Here, a model-understanding explanation could be that the metric generally assigns lower scores when input sentences are subjected to numeric or word order perturbations.

The work by Guerreiro et al. (2023) is also related to these approaches. One of their contributions is an expert annotated data set of machine translation hallucinations. They use this data set to test how well selected metrics can identify each hallucination type. Further, they give the recommendation to simply use the sequence log-probability for hallucination detection. While the previous techniques describe the susceptibility of metrics towards various perturbation types, Sun et al. (2022) describe methods to analyze the metrics’ social biases. They also base their evaluation on perturbed input sentences; however, they evaluate whether a metric grades two versions of a sentence that contain opposite social stereotypes the same with respect to a neutral reference sentence (that does not contain stereotypes). They find that generation-based metrics show the least biases.
**Explanation Type:** Perturbation Robustness

**Goals:** The presented *perturbation robustness* approaches give detailed insights into metric weaknesses. Therefore, they are most suited to be used for metric diagnosis and improvement.

**Audiences:** Metric developers can use these methods for diagnosis. They can also be used by metric users to check if a metric fits their use case. Regulatory entities might test if metrics fulfill their guidelines, e.g., exhibit no social biases. Lastly, affected users might gather evidence when treated unfairly.

**Limitations:** A drawback of the perturbation-based approaches is that they require predefined error types. Therefore, errors lacking a definition in the evaluation scheme that would fool the machine translation metrics could be missed. These approaches are faithful to some degree, namely in the domains they have been evaluated in. For small perturbation sets it could be questionable if the model generally behaves as predicted. Therefore we would suggest evaluating each property on a large, general data sample. Still, it could be advisable to employ faithfulness tests, if possible, to achieve guarantees.

---

**Linguistic properties (Row 5 in Figure 3)**

Here, we summarize model-understanding techniques that directly summarize various linguistic properties a metric fulfills. So far, the only method that falls into this category for MT is Kaster et al. (2021) (see linguistic properties/black box in Figure 3). They propose a model-understanding technique that decomposes the score of sentence-level BERT-based metrics into linguistic factors. In particular, they explore the degree to which metrics consider the properties *syntax*, *semantics*, *morphology* and *lexical overlap*. To do so, they measure these properties in a given data set and learn regressors that explain how much each distinct property contributes to the original metric’s score. In their experiments, they show that each metric captures semantic similarity and lexical overlap to some degree. Syntactic and morphological similarity are either captured to a smaller extent or are even negatively correlated with the metric score. Especially, MoverScore (Zhao et al., 2019) and BERTScore (Zhang et al., 2020a) have a comparatively high coefficient for the lexical score.

The example in Figure 4 shows that these techniques can give explanations that highlight which properties a metric generally focuses on. In the specific example, the metric puts a high focus on lexical properties (green bar) and a low focus on semantics (red bar).

Note that the explanation of linguistic properties is close to the decision-understanding of quality aspects. If these are collected over many samples and a regression is performed, similar explanations might be achieved. A related approach used for another NLP task is by Opitz and Frank (2022), who retrain sentence embeddings to decompose into subcomponents that match the aspects of abstract meaning representation, a format to represent the meaning of sentences. This allows them to attribute parts of the score to different aspects of a sentence. They evaluate their approach on the model-understanding level. When this would be applied to MT, it would also count towards this explanation type.
We also list Kaster et al. (2021) and Opitz and Frank (2022) with quality aspect explanations, as the metrics that measure the linguistic factors can also be computed on a sentence-level.

**Explanation Type:** Linguistic Properties  
**Goals:** The discussed linguistic properties approaches evaluate general metric properties and can be used for metric debugging. As they generate sentence/metric-level property scores, the techniques are not usable for word-level semi-automatic labeling. When used on a sentence-level, the techniques can aid the goal of making metrics more accessible. While not employed here, the techniques might be usable to detect social biases by introducing a set of bias detection scores from which to learn the regressors.  
**Audiences:** The linguistic properties can be easily explained to non-experts, however metric selection should not solely be performed based on linguistic properties and is advised to be handled by domain experts for MT and MT metrics. The methods are especially useful to debug metrics, i.e., for metric developers.  
**Limitations:** The approach by Kaster et al. (2021) could not explain reference-free metrics well, so plausibly requires alternate explanatory variables. The search for regressors may be inspired by quality aspect approaches (described earlier this section) where not a global metric score is reported but several scores (such as coherence, fluency, etc.). These could then compose a global MT score. Further, some of the “property metrics” could be considered black box variables themselves and future work might replace them by more transparent factors. One might also explore the collinearity of the different regressors. For Kaster et al. (2021), faithfulness is likely given to some degree, around the samples that was trained on. Still, it could be advisable to employ faithfulness tests, if possible, to achieve guarantees.

5. Future Work

In this section, we describe future research directions for explainable MT metrics. Specifically, we (1) consider directions for explanation types that have been used in the papers described in our taxonomy, (2) describe potential explanation types from other domains, (3) point out needs for future evaluations and concepts and (4) urge for an explanation of domain boundaries.

*Exploring present explanation types* Based on our taxonomy (see Figure 3), we identify unexplored areas of explainable MT evaluation (empty table cells). We find that only *feature importance techniques* are thoroughly explored on all levels of model access. The other methods mostly explain in black box settings or are based on metrics that are explainable by design. We suppose that gradients, attention weights, and embeddings have, so far, mostly been applied in the field of feature importance, as there is often already a one-to-one relationship between their values and input features. Also, due to the recent shared tasks, feature importance has likely received the most attention.
Future work could consider enhancing quality aspect explanations — Row 1 in Figure 3 — by incorporating more model access. For example, a separate fluency score might be realizable by employing measurements on target side embeddings in unsupervised matching metrics. Current model-understanding techniques for MT metrics — Row 4-5 in Figure 3 — are mostly applied in a black box manner. For these, using more model access might also provide further insights. For example, analyses of the contextualized embedding space and how it affects a metric score could explain the metric on a more general level. Many works consider the probing of transformer-based architectures (e.g., Tenney et al., 2019; Liu et al., 2019; Sajjad et al., 2022b) and it would be interesting to see if their results can inform the metric development.

Also, current perturbation robustness techniques often use manually defined perturbation strategies. It would be interesting to use adversarial attacks instead that automatically infer these strategies.

Exploring other explanation types Besides the works in our taxonomy, other fields of NLP and explainable AI in general have proposed further explanation types, e.g., counterfactuals, natural language explanations and rule extraction (e.g., Madsen et al., 2022). We illustrate these options with an example in Figure 5. We already introduced counterfactuals in §2.2, stating their role in evaluations of robustness (as adversarial attacks) and fairness. It remains to explore if single sentence explanations also carry explanatory value. In the examples for counterfactuals in Figure 5, the original word “bekommen” is switched with the words “wurden” (to get) resp. “erhalten” (to receive) in two counterfactuals. Based
Towards Explainable Evaluation Metrics For MT

on the scores achieved by these counterfactuals, a user could infer metric weaknesses. One technique to generate counterfactuals is for example described by Wu et al. (2021).

**Decision rule** explanations describe rules which a model follows when computing its output (Ribeiro et al., 2018). Often, these techniques were designed to explain classification models and are based on perturbations that change the originally predicted class (Ribeiro et al., 2018). For metrics, the scores could be discretized beforehand. An example from other NLP domains are Anchors (Ribeiro et al., 2018). In our example (see Figure 5), the rule describes that if “became” is present in the source and “wurden” is present in the translation, the metric will assign a score higher than 0.5. Obviously in other contexts, such a rule would fail.

Another type of explanations are **hierarchical feature importance** explanations, which will assign importance scores to groups of input tokens in single (Chen et al., 2020) or paired input sentences (Chen et al., 2021). In the given hypothetical example (see Figure 5), the groups of **became/bekommen** and **Things/Die Dinge** were formed and get assigned separate importance scores. A related approach was tested in the SemEval-2015 and -2016 shared tasks on semantic similarity, where the task was to assign labels that capture the similarity between token groups from two sentences (Agirre et al., 2015, 2016).

**Natural language explanations** are explanations in free text form (e.g., Wiegreffe and Marasovic, 2021). This is an explainability type where the generative abilities of recent large language models (LLM), like ChatGPT and GPT4, can be leveraged. The works by Fu et al. (2023), Lu et al. (2023b), Kocmi and Federmann (2023b), Kocmi and Federmann (2023a), Fernandes et al. (2023) and Xu et al. (2023) already explore the use of LLMs as metrics and extract the metric scores from LLM outputs (or generation probabilities) that contain (or could easily be extended to provide) natural language descriptions of why a certain score is assigned. In our taxonomy, we grouped them under fine-grained error explanations, as this is their main goal. In these works, the explanations are mostly used as a byproduct, while a thorough evaluation of the explanation quality and further use cases remains for future work.

The example in Figure 5 was generated with the help of ChatGPT, which shows that such an approach might yield plausible explanations. An example from other NLP domains is Rajani et al. (2019), who generate explanations for commonsense reasoning. In NLG in general, many other approaches for LLM-based metrics have been constructed since the release of ChatGPT. These could potentially be extended to MT metrics and explainability (e.g., Chiang and Lee, 2023; Wang et al., 2023; Ji et al., 2023; Chen et al., 2023; Liu et al., 2023). As for example Kocmi and Federmann (2023a) point out, it is dangerous for academia to rely on proprietary models, especially for evaluation purposes. This includes the following issues: Metrics based on proprietary LLMs (1) could be influenced by the LLM providers, (2) could change or become unavailable at any time, (3) could carry more

---

13. We again note that these techniques will not necessarily describe a metric’s internal workings, but might be helpful due to providing plausible explanations, see §2.2.

14. We used ChatGPT in June 2023 with following prompts and asked to rephrase the output once:

Ground Truth: “Things got complicate”
Hypothesis: “Die Dinge bekamen kompliziert”
Is this translation of the ground truth correct? Why or why not?
unexpected biases due to unknown training data and (4) could be contaminated in their evaluation by being trained on common test sets.

Recent works, describe prompting strategies like chain-of-thought (Wei et al., 2022) and tree-of-thought (Yao et al., 2023) where language models show performance improvements if they are giving explanations of their output. While some of the works that are described in our taxonomy already employ such techniques to improve metrics a systematic exploration of the effect on metric scores, or whether these techniques can provide reasonable explanations (that are at least plausible), remains. A first work that considers this systematic exploration is the EVAL4NLP 2023 shared task (Leiter et al., 2023) where participants should improve metrics scores only by prompting. Their results on a small scale experiment indicate that the participants’ generated natural language explanations are not helpful to humans, in most cases.

Neuron Explanations are another potential use case of LLMs to separately explain each Neuron in the underlying networks. Bills et al. (2023) recently showed that GPT4 can (to some extent) summarize highlighted neuron activation in textual input of GPT2 to the same extent humans can. This allows to successfully simulate the behavior and effects of some neurons. Interestingly, they also find error related neurons, like a “post-typo-neuron”. For MT metrics, this would allow to determine the role certain neurons play to determine scores when specific features like “became is used” are present in the metric input. At the moment, the success of neuron explanation methods is still limited (Bills et al., 2023).

Interactive Explanations: Due to their context size, instruction fine-tuning and emergent abilities of recent LLMs allow to interactively ask questions and adapt based on the answers (e.g., OpenAI, 2023a). Jacovi et al. (2023) state that interactivity is a key point of successfully communicating the explained subject to an explainee. LLM-based explanations (like natural language/neuron explanations ... ) might achieve this. For example, a user could ask for more detailed information for a mistranslation that was noted in a first explanation.

New approaches could further consider displaying multiple explanation types together, like Xu et al. (2023) who present fine-grained error annotations with natural language explanations.

Future evaluation and concepts Current work scarcely performs human evaluation of explanations. This is partly because the focus is often not on the explanations but other aspects, like the sentence-level scores. One form of human evaluation (for plausibility) is often to use pre-computed human explanations and check the correlation to them (e.g., Zerva et al., 2022). Current works on MT metrics do, however, not evaluate whether the explanations are actually helpful to explain the concepts they describe to their audience. This means, whether the explanations manage to help the explainees to adapt their own theory of mind to be a coherent abstraction of the real problem (Jacovi et al., 2023). Hence, we encourage future work to address this, e.g., by considering simulatability (Hase and Bansal, 2020). Human evaluation may also be performed in downstream tasks, by checking performance with and without explanations. Regarding plausibility, it could also be interesting to evaluate whether plausible explanations of incorrect metric scores can actually deceive non-experts or even translation experts, as warned by Jacovi and Goldberg (2021) and Jin et al. (2023) (for general explainability use cases).
As noted throughout the paper, the degree of faithfulness an explanation has to the metric’s internal decision process is an important property (Jacovi and Goldberg, 2020), especially for metric debugging and bias detection. Current works on decision-understanding of MT metrics, however, mostly dismiss faithfulness. Future work might consider the evaluation of metric faithfulness, to enable these use-cases. We note that this evaluation will be complex, due to faithfulness being multi-faceted (Jacovi and Goldberg, 2020). For example, lower-granularity scores can faithfully explain sentence-level scores that are aggregated from them, however this does not explain how the actual black box component produced the lower-granularity scores. Further, there are evaluation pitfalls that exist in faithfulness evaluation (Ju et al., 2022), ruling out common methods like area under the perturbation curve scores (AOPC). Therefore, future work should first consider what constitutes faithful explanations in terms of metrics and design evaluation approaches.

Further, we note that current works mostly consider explaining sentence-level scores. Future work, might also consider other metric granularity types more, such as word-level and document-level metrics. Additionally, future work could explore the influence of language resource availability on the explanations (although it may largely depend on metric quality) and whether some explainable metric setups are not applicable in low-resource settings.

**Exploring domain boundaries** Future work might also consider under-explored use cases of explainability for MT metrics. Currently, only one work considers the social biases of MT metrics (Sun et al., 2022) and none considers the usage of explainability for semi-automatic labeling of training data. Besides examining just the metrics, future work could also inspect the relation of explainable MT metrics, especially plausible-only explanations, to explainable MT (e.g., Stahlberg et al., 2018), as explanations of both will probably be similar. Lastly, future work could explore providing similar taxonomies as ours that are comprising the whole field of explainable natural language generation metrics.

6. Related Work

A large number of conceptual works and surveys have been conducted in the area of explainability (e.g., Lipton, 2016; Biran and Cotton, 2017; Adadi and Berrada, 2018; Doshi-Velez and Kim, 2017; Došilović et al., 2018; Guidotti et al., 2018; Arya et al., 2019; Carvalho et al., 2019; Gilpin et al., 2018; Miller, 2019; Barredo Arrieta et al., 2020; Bodria et al., 2023; Linardatos et al., 2021).

There are few surveys that touch upon the need for explainable NLG (here MT), thereby motivating our work. In their survey of NLG metrics, Celikyilmaz et al. (2020) see the need for explainable evaluation metrics to spot system quality issues and to achieve higher trust in the evaluation of NLG systems. They consider that such quality issues might be unintended biases and factual inconsistencies. Sai et al. (2022) instead propose that explainable NLG metrics should focus on providing more information than just a single score (such as fluency or adequacy). Gehrmann et al. (2023) also notice the need for interpreting NLG metrics and give a brief overview of selected related approaches. The main contribution of their work is an overview of common NLG metric shortcomings and best practices to prevent them. Our focus on concepts for explainable MT metrics is a different one and our taxonomy of approaches is much more extensive.
General surveys on explainable NLP are related to our work, as they highlight how explainability techniques have been used in NLP and might further guide the development of future approaches for explainable MT metrics (e.g., Vijayakumar, 2023; Sajjad et al., 2022a; Gurrapu et al., 2023; Lertvittayakumjorn and Toni, 2021; Saha et al., 2022; Sun et al., 2021; Zini and Awad, 2022; Qian et al., 2021; Madsen et al., 2022). Besides these works, there are also surveys of robustness (e.g., Goyal et al., 2023) and fairness (e.g., Blodgett et al., 2020), which are related, as discussed in §2.2.

The taxonomy graphic in Figure 3 was adapted from the survey of post-hoc explainability methods in NLP by Madsen et al. (2022). In specific, we follow their approach to display model access on the x-axis and explanation types on the y-axis. Further, we have customized their TIKZ code. We specifically focus on the explainability of MT metrics; therefore, our taxonomy is completely different content-wise. Our definitions of explainability, its goals and audiences are in some parts adapted from the survey of Barredo Arrieta et al. (2020) (see §2.2). Our work is different from previous surveys, as we set our focus on the explainability of MT metrics.

7. Conclusion

In this work, we discuss audiences, goals and properties for explainable machine translation metrics, a nascent field that may help further overcome the dominance of classical low-quality evaluation metrics. We also survey and categorize recent approaches on explainable MT metrics into a taxonomy, highlighting their results and limitations. Currently, two dominant approaches to explainability for MT metrics are (1) feature importance explanations that highlight erroneous words in source and hypothesis to explain sentence-level scores and (2) perturbation robustness approaches that check a metric’s robustness to manually devised types of input perturbations. We also identify a current trend towards (3) fine-grained error explanations likely caused by the recent improvements of LLMs like ChatGPT and by the adaptation of MQM annotations in the WMT metrics shared tasks. A major weakness of the current realization of (1) is that error highlights do not consider the correspondence between words in source and target. Further, a weakness of (2) is that perturbation types have to be manually defined, an issue that might be tackled by using more general approaches of adversarial attacks in the future (we discuss this issue in more detail in an earlier version of this paper; Leiter et al., 2022a). Generally, most current evaluations of explainability approaches for MT metrics do not consider the faithfulness of explanations, limiting their use in metric debugging and high-risk scenarios. We also present a vision of future approaches to explainable evaluation metrics, which should help fix the problems of the above paradigms and provide guidance to explore unexplored areas. Here, we also urge future work to consider the desiderata and implications of faithfulness for explainable MT metrics.

Our broader vision is that explainability is now a ‘desirable but optional’ feature, but we argue that in the future it will become essential, even compulsory, especially for evaluation metrics as a highly sensitive task assessing the quality (and veracity) of translated information content. Explainability builds transparency and trust for users, eases bug-fixing and shortens improvement cycles for metric designers and will be required by law/regulations for AI systems to be applied to large-scale, high-stake domains. In this context, we hope
our work will catalyze efforts on the topic of explainable evaluation metrics for machine translation.

Acknowledgments

Since November 2022 Christoph Leiter is financed by the BMBF project “Metrics4NLG”. Piyawat Lertvitayakumjorn had been financially supported by Anandamahidol Foundation, Thailand, from 2015-2021. He mainly contributed to this work only until September 2022 while affiliated with Imperial College London (before joining Google as a research scientist after that). Marina Fomicheva mainly contributed to this work until April 2022. Wei Zhao is supported by the Klaus Tschira Foundation, Heidelberg, Germany. Yang Gao mainly contributed to this work before he joined Google Research in December 2021. Stefan Eger is financed by DFG Heisenberg grant EG 375/5–1 and by the BMBF project “Metrics4NLG”.

References


Keqin Bao, Yu Wan, Dayiheng Liu, Baosong Yang, Wenqiang Lei, Xiangnan He, Derek F.Wong, and Jun Xie. Towards fine-grained information: Identifying the type and location of translation errors. ArXiv, abs/2302.08975, 2023.


Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. Language (technology) is power: A critical survey of “bias” in NLP. In Proceedings of the 58th Annual Meeting
Towards Explainable Evaluation Metrics For MT


Towards Explainable Evaluation Metrics For MT


Yunjie Ji, Yan Gong, Yiping Peng, Chao Ni, Peiyan Sun, Dongyu Pan, Baochang Ma, and Xiangang Li. Exploring chatgpt’s ability to rank content: A preliminary study on consistency with human preferences, 2023.


Tom Kocmi and Christian Federmann. Large language models are state-of-the-art evaluators of translation quality. In Mary Nurminen, Judith Brenner, Maarit Koponen, Sirkku
Towards Explainable Evaluation Metrics For MT


Qingyu Lu, Liang Ding, Liping Xie, Kanjian Zhang, Derek F. Wong, and Dacheng Tao. Toward human-like evaluation for natural language generation with error analysis. In Anna
Towards Explainable Evaluation Metrics For MT


Elena Murgolo, Javad Pourmostafa Roshan Sharami, and Dimitar Shterionov. A quality estimation and quality evaluation tool for the translation industry. In Proceedings of


Towards Explainable Evaluation Metrics For MT


Towards Explainable Evaluation Metrics For MT


Xiaofei Sun, Diyi Yang, Xiaoya Li, Tianwei Zhang, Yuxian Meng, Han Qiu, Guoyin Wang, Eduard H. Hovy, and Jiwei Li. Interpreting deep learning models in natural language processing: A review. ArXiv, abs/2110.10470, 2021.


Towards Explainable Evaluation Metrics For MT


