DEMETR: Diagnosing Evaluation Metrics for Translation

Marzena Karpinska[♢] **Nishant Raj**[♢] **Katherine Thai**[♢]
Yixiao Song[♠] **Ankita Gupta**[♢] **Mohit Iyyer**[♢]

[♢]Manning College of Information and Computer Sciences, UMass Amherst
[♠]Department of Linguistics, UMass Amherst
{mkarpinska,kbthai,ankitagupta,miyyer}@cs.umass.edu
{nishantraj,yixiaosong}@umass.edu

Abstract

While machine translation evaluation metrics based on string overlap (e.g., BLEU) have their limitations, their computations are transparent: the BLEU score assigned to a particular candidate translation can be traced back to the presence or absence of certain words. The operations of newer *learned* metrics (e.g., BLEURT, COMET), which leverage pretrained language models to achieve higher correlations with human quality judgments than BLEU, are opaque in comparison. In this paper, we shed light on the behavior of these learned metrics by creating DEMETR, a diagnostic dataset with *31*K English examples (translated from *10* source languages) for evaluating the sensitivity of MT evaluation metrics to *35* different linguistic perturbations spanning semantic, syntactic, and morphological error categories. All perturbations were carefully designed to form minimal pairs with the actual translation (i.e., differ in only one aspect). We find that learned metrics perform substantially better than string-based metrics on DEMETR. Additionally, learned metrics differ in their sensitivity to various phenomena (e.g., BERTSCORE is sensitive to untranslated words but relatively insensitive to gender manipulation, while COMET is much more sensitive to word repetition than to aspectual changes). We publicly release DEMETR to spur more informed future development of machine translation evaluation metrics[1].

1 Introduction

Automatically evaluating the output quality of machine translation (MT) systems remains a difficult challenge. The BLEU metric (Papineni et al., 2002), which is a function of n-gram overlap between system and reference outputs, is still used widely today despite its obvious limitations in measuring

[1] https://github.com/marzenakrp/demetr

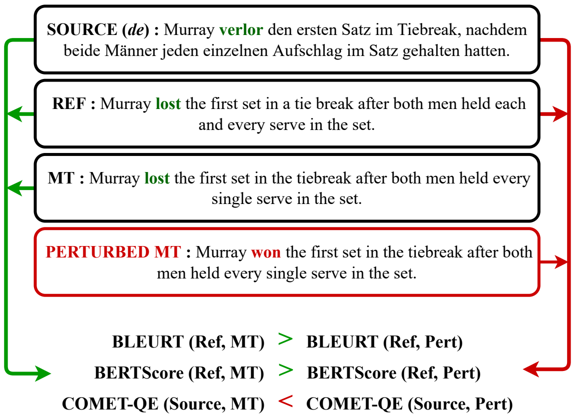

Figure 1: An example perturbation (antonym replacement) from our DEMETR dataset. We measure whether different MT evaluation metrics score the unperturbed translation higher than the perturbed translation; in this case, BLEURT and BERTSCORE accurately identify the perturbation, while COMET-QE fails to do so.

semantic similarity (Fomicheva and Specia, 2019; Marie et al., 2021; Kocmi et al., 2021; Freitag et al., 2021). Recently-developed *learned* evaluation metrics such as BLEURT (Sellam et al., 2020a), COMET (Rei et al., 2020), MOVERSCORE (Zhao et al., 2019), or BARTSCORE (Yuan et al., 2021a) seek to address these limitations by either fine-tuning pretrained language models directly on human judgments of translation quality or by simply utilizing contextualized word embeddings. While learned metrics exhibit higher correlation with human judgments than BLEU (Barrault et al., 2021), their relative lack of interpretability leaves it unclear as to *why* they assign a particular score to a given translation. This is a major reason why some MT researchers are reluctant to employ learned metrics in order to evaluate their MT systems (Marie et al., 2021; Gehrmann et al., 2022; Leiter et al., 2022).

In this paper, we build on previous metric explainability work (Specia et al., 2010; Macketanz

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 9540–9561
December 7-11, 2022 ©2022 Association for Computational Linguistics

et al., 2018; Fomicheva and Specia, 2019; Kaster et al., 2021; Sai et al., 2021a; Barrault et al., 2021; Fomicheva et al., 2021; Leiter et al., 2022) by introducing DEMETR, a dataset for **D**iagnosing **E**valuation **METR**ics for machine translation, that measures the sensitivity of an MT metric to *35* different types of linguistic perturbations spanning common syntactic (e.g., incorrect word order), semantic (e.g., undertranslation), and morphological (e.g., incorrect suffix) translation error categories. Each example in DEMETR is a tuple containing {`source`, `reference`, `machine translation`, `perturbed machine translation`}, as shown in Figure 1. The entire dataset contains of *31*K total examples across *10* different source languages (the target language is always English). The perturbations in DEMETR are produced semi-automatically by manipulating translations produced by commercial MT systems such as Google Translate, and they are manually validated to ensure the only source of variation is associated with the desired perturbation.

We measure the accuracy of a suite of *14* evaluation metrics on DEMETR (as shown in Figure 1), discovering that learned metrics perform far better than string-based ones. We also analyze the relative *sensitivity* of metrics to different grades of perturbation severity. We find that metrics struggle at times to differentiate between minor errors (e.g., punctuation removal or word repetition) with semantics-warping errors such as incorrect gender or numeracy. We also observe that the reference-free[2] COMET-QE learned metric is more sensitive to word repetition and misspelled words than severe errors such as entirely unrelated translations or named entity replacement. We publicly release DEMETR and associated code to facilitate more principled research into MT evaluation.

2 Diagnosing MT evaluation metrics

Most existing MT evaluation metrics compute a score for a candidate translation t against a reference sentence r.[3] These scores can be either a simple function of character or token overlap between t and r (e.g., BLEU), or they can be the result of a complex neural network model that embeds t and r (e.g., BLEURT). While the latter class of *learned* metrics[4] provides more meaningful judgments of translation quality than the former, they are also relatively uninterpretable: the reason for a particular translation t receiving a high or low score is difficult to discern. In this section, we first explain our perturbation-based methodology to better understand MT metrics before describing the collection of DEMETR, a dataset of linguistic perturbations.

2.1 Using translation perturbations to diagnose MT metrics

Inspired by prior work in *minimal pair*-based linguistic evaluation of pretrained language models such as BLiMP (Warstadt et al., 2020), we investigate how sensitive MT evaluation metrics are to various perturbations of the candidate translation t. Consider the following example, which is designed to evaluate the impact of word order in the candidate translation:

> `reference translation` r: Pronunciation is relatively easy in Italian since most words are pronounced exactly how they are written.
>
> `machine translation` t: Pronunciation is relatively easy in Italian, as most words are pronounced exactly as they are spelled.
>
> `perturbed machine translation` t': Spelled pronunciation as Italian, relatively are most is as they pronounced exactly in words easy.

If a particular evaluation metric SCORE is sensitive to this shuffling perturbation, SCORE(r, t'), the score of the perturbed translation, should be lower than SCORE(r, t).[5] Note that while other minor translation errors may be present in t, the perturbed translation t' differs only in a specific, controlled perturbation (in this case, shuffling).

2.2 Creating the DEMETR dataset

To explore the above methodology at scale, we create DEMETR, a dataset that evaluates MT metrics on *35* different linguistic phenomena with 1K perturbations per phenomenon.[6] Each example in DEMETR consists of (1) a sentence in one of *10*

[2]While prior work uses also terms such as "reference-less" and "quality estimation," we employ the term "reference-free" as it is more self-explanatory.

[3]Some metrics, such as COMET, additionally condition the score on the source sentence. "Reference-less" metrics, such as COMET-QE, compare the candidate translation t directly against the source text s.

[4]We define *learned* metrics as any metric which uses a machine learning model (including both pretrained and supervised methods).

[5]For reference-free metrics like COMET-QE, we include the source sentence s as an input to the scoring function instead of the reference.

[6]As some perturbations require presence of specific items (e.g., to omit a named entity, one has to be present) not all perturbations include exactly 1k sentences.

ID	Category	Description	Error severity
1	accuracy	word repetition (twice)	minor
2		word repetition (four times)	minor
3		too general word (undertranslation)	major
4		untranslated word (codemix)	major
5		omitted perpositional phrase	major
6		incorrect word added	critical
7		change to antonym	critical
8		change to negation	critical
9		replaced named entity	critical
10		incorrect numeric	critical
11		incorrect gender pronoun	critical
12	fluency	omitted conjunction	minor
13		part of speech shift	minor
14		switched word order (word swap)	minor
15		incorrect case (pronouns)	minor
16		incorrect preposition or article	minor-major
17		incorrect tense	major
18		incorrect aspect	major
19		change to interrogative	major
20	mixed	omitted adj/adv	minor-major
21		omitted content verb	critical
22		omitted noun	critical
23		omitted subject	critical
24		omitted named entity	critical
25	typography	misspelled word	minor
26		deleted character	minor
27		omitted final punctuation	minor
28		added punctuation	minor
29		tokenized sentence	minor
30		lowercased sentence	minor
31		first word lowercased	minor
32	baseline	empty string	base
33		unrelated translation	base
34		shuffled words	base
35		reference as translation	base

Table 1: List of perturbations included in DEMETR with their corresponding error severity. Details can be found in Appendix A

source languages, (2) an English translation written by a human translator, (3) a machine translation produced by Google Translate,[7] and (4) a perturbed version of the Google Translate output which introduces exactly one mistake (semantic, syntactic, or typographical).

Data sources and filtering: We utilize *X-to-English* translation pairs from two different datasets, WMT (Callison-Burch et al., 2009; Bojar et al., 2013, 2015, 2014; Akhbardeh et al., 2021; Barrault et al., 2020) and FLORES (Guzmán et al., 2019), aiming at a wide coverage of topics from different sources. WMT has been widely used over the years as a popular MT shared task, while FLORES was recently curated to aid MT evaluation. We consider only the test split of each dataset to prevent possible leaks, as both current and future metrics are likely to be trained on these two datasets. We sample *100* sentences (*50* from each of the two datasets) for each of the following *10* languages: French (*fr*), Italian (*it*), Spanish (*es*), German (*de*), Czech (*cs*), Polish (*pl*), Russian (*ru*), Hindi (*hi*), Chinese (*zh*), and Japanese (*ja*).[8] We pay special attention to the language selection, as newer MT evaluation metrics, such as COMET-QE or PRISM-QE, employ only the source text and the candidate translation. We control for sentence length by including only sentences between *15* and *25* words long, measured by the length of the tokenized reference translation. Since we re-use the same sentences across multiple perturbations, we did not include shorter sentences because they are less likely to contain multiple linguistic phenomena of interest.[9] As the quality of sampled sentences varies, we manually check each source sentence and its translation to make sure they are of satisfactory quality.[10]

Translating the data: Given the filtered collection of source sentences, we next translate them into English using the Google Translate API.[11] We manually verify each translation, editing or resampling the instances where the machine translation contains critical errors.[12] Through this process,

[7]We edit the machine translation to assure a satisfactory quality. In cases where the Google Translate output is exceptionally poor, we either replace the sentence or replace the translation with one produced by DeepL (Frahling, 2022) or GPT-3 (Brown et al., 2020).

[8]We choose languages that represent different families (Romance, Germanic, Slavic, Indo-Iranian, Sino-Tibetan, and Japonic) with different morphological traits (fusional, agglutinative, and analytic) and wide range of writing systems (Latin alphabet, Cyrillic alphabet, Devanagari script, Hanzi, and Kanji/Hiragana/Katakana).

[9]Similarly, we do not include sentences over 25 words long in DEMETR as some languages may naturally allow longer sentences than others, and we wanted to control the length distribution.

[10]In the sentences sampled from WMT, we notice multiple translation and grammar errors, such as translating Japanese その最大は本州列島で、世界で7番目に大きい島とされています。 as *(the biggest being Honshu), making Japan the 7th largest island in the world*, which would suggest that Japan is an island, instead of *the largest of which is the Honshu island, considered to be the seventh largest island in the world.* or "kakao" ("cacao") incorrectly declined as "kakaa" in Polish. These sentences were rejected, and new ones were sampled in their place. We also resampled sentences which translations contained artifacts from neighboring sentences due to partial splits and merges, and sentences which exhibit *translationese*, that is sentences with source artifacts (Koppel and Ordan, 2011). Finally, we omit or edit sentences with translation artifacts due to the direction of translation, as both WMT and FLORES contain sentences translated from English to another languages. Since the translation process is *not* always fully reversible, we omit sentences where translation from the given language to English would not be possible in the form included in these datasets (e.g., due to addition or omission of information).

[11]All sentences were translated in May, 2022.

[12]We pay special attention to errors which overlap with our perturbations. For instance, we check all the named entities,

we obtain *1*K curated examples per perturbation (*100* sentences $\times$ *10* languages) that each consist of source and reference sentences along with a machine translation of reasonable quality.

2.3 Perturbations in DEMETR

We perturb the machine translations obtained above in order to create *minimal pairs*, which allow us to investigate the sensitivity of MT evaluation metrics to different types of errors. Our perturbations are loosely based on the Multidimensional Quality Metrics (Burchardt, 2013, MQM) framework developed to identify and categorize MT errors. Most perturbations were performed semi-automatically by utilizing STANZA (Qi et al., 2020), SPACY[13] or GPT-3 (Brown et al., 2020), applying handcrafted rules and then manually correcting any errors. Some of the more elaborate perturbations (e.g., translation by a too general term, where one had to be sure that a better, more precise term exists) were performed manually by the authors or linguistically-savvy freelancers hired on the Upwork platform.[14] Special care was given to the plausibility of perturbations (e.g., numbers for replacement were selected from a probable range, such as *1-12* for months). See Table 2 for descriptions and examples of most perturbations; full list in Appendix A.

We roughly categorize our perturbations into the following four categories:

- **ACCURACY**: Perturbations in the accuracy category modify the semantics of the translation by either incorporating misleading information (e.g., by adding plausible yet inadequate text or changing a word to its antonym) or omitting information (e.g., by leaving a word untranslated).

- **FLUENCY**: Perturbations in the fluency category focus on grammatical accuracy (e.g., word form agreement, tense, or aspect) and on overall cohesion. Compared to the mistakes in the accuracy category, the true meaning of the sentence can be usually recovered from the context more easily.

- **MIXED**: Certain perturbations can be classified as both accuracy and fluency errors. Concretely, this category consists of omission errors that not only obscure the meaning but also affect the grammaticality of the sentence. One such error is *subject removal*, which will result not only in an ungrammatical sentence, leaving a gap where the subject should come, but also in information loss.

- **TYPOGRAPHY**: This category concerns punctuation and minor orthographic errors. Examples of mistakes in this category include punctuation removal, tokenization, lowercasing, and common spelling mistakes.

- **BASELINE**: Finally, we include both upper and lower bounds, since learned metrics such as BLEURT and COMET do not have a specified range that their scores can fall into. Specifically, we provide three baselines: as lower bounds, we either change the translation to an unrelated one or provide an empty string,[15] while as an upper bound, we set the perturbed translation t' equal to the reference translation r, which should return the highest possible score for reference-based metrics.

Error severity: Our perturbations can also be categorized by their *severity* (see Table 1). We use the following categorization scheme for our analysis experiments:

- **MINOR**: In this type of error, which includes perturbations such as dropping punctuation or using the wrong article, the meaning of the source sentence can be easily and correctly interpreted by human readers.

- **MAJOR**: Errors in this category may not affect the overall fluency of the sentence but will result in some missing details. Examples of major errors include undertranslation (e.g., translating "church" as "building"), or leaving a word in the source language untranslated.

- **CRITICAL**: These are catastrophic errors that result in crucial pieces of information going missing or incorrect information being added in a way unrecognizable for the reader, and are also likely to suffer from severe fluency issues. Errors in this category include

as replacing an already incorrect named entity with another incorrect named entity does NOT make the perturbed translation *worse* than the original.

[13] https://spacy.io/usage/linguistic-features

[14] See https://www.upwork.com/. Freelancers were paid an equivalent of $15 per hour.

[15] Since most of the metrics will not accept an empty string, we pass a full stop instead.

Category	Type	Example	Description	Implementation	Error Severity
ACCURACY	**repetition**	I don't know if you realize that most of the goods imported into this country from Central America are duty **free**. I don't know if you realize that most of the goods imported into this country from Central America are duty **free free**.	The last word is being repeated twice. Punctuation is added after the last repeated word.	automatic	minor
	repetition	Gordon Johndroe, Bush's spokesman, referred to the North Korean commitment as "an important advance towards the goal of achieving verifiable denuclearization of the Korean **penisula**." Gordon Johndroe, Bush's spokesman, referred to the North Korean commitment as "an important advance towards the goal of achieving verifiable denuclearization of the Korean **penisula penisula penisula penisula**."	The last word is being repeated four times. Punctuation is added after the last repeated word.	automatic	minor
	hypernym	The language most of the people working in the Vatican City use on a daily basis is Italian, and Latin is often used in religious **ceremonies.** The language most of the people working in the Vatican City use on a daily basis is Italian, and Latin is often used in religious **activities.**	A word translated by a too general term (undertranslation). Special care was given in order to assure the word used in perturbed text is more general, and incorrect, translation of the original word.	manual with suggestions from GPT-3	major
	untranslated	The Polish Air Force will eventually be equipped with 32 F-35 Lightning II fighters **manufactured** by Lockheed Martin. The Polish Air Force will eventually be equipped with 32 F-35 Lightning II fighters **produkowane** by Lockheed Martin.	One word is being left untranslated. We manually assure that each time only one word is left untranslated.	manual	major
	completeness	She is **in custody** pending prosecution and trial; but any witness evidence could be negatively impacted because her image has been widely published. She is _____ pending prosecution and trial; but any witness evidence could be negatively impacted because her image has been widely published.	One prepositional phrase is being removed. Whenever possible, we remove the shortest prepositional phrase in order to assure that the perturbed sentence is not much shorter than the original translation.	automatic (Stanza) with manual check	major
	addition	_____ Plants look their best when they are in a natural environment, so resist the temptation to remove "just one." **Power** plants look their best when they are in a natural environment, so resist the temptation to remove "just one."	One word is being added. We make sure that the added word does not disturb the grammaticality of the sentence but changes the meaning in a significant way.	manual	critical
	antonym	He has been unable to relieve the **pain** with medication, which the competition prohibits competitors from taking. He has been unable to relieve the **pleasure** with medication, which the competition prohibits competitors from taking.	One word (noun, verb, adj., or adv.) is being changed to its antonym.	manual with suggestions from GPT-3	critical
	mistranslation negation	Last month, a presidential committee **recommended** the resignation of the former CEP as part of measures to push the country toward new elections. Last month, a presidential committee **didn't recommend** the resignation of the former CEP as part of measures to push the country toward new elections.	Affirmative sentences are being changed into negations. Rare negations are being changed to affirmative sentences.	manual	critical
	mistranslation named entity	Late night presenter **Stephen Colbert** welcomed 17-year-old Thunberg to his show on Tuesday and conducted a lengthy interview with the Swede. Late night presenter **John Oliver** welcomed 17-year-old Thunberg to his show on Tuesday and conducted a lengthy interview with the Swede.	Named entity is replaced with another named entity from the same category (person, geographic location, and organization).	automatic (Stanza) with manual check	critical
	mistranslation numbers	The Chinese Consulate General in Houston was established in **1979** and is the first Chinese consulate in the United States. The Chinese Consulate General in Houston was established in **1997** and is the first Chinese consulate in the United States.	A number is being replaced with an incorrect one. Special attention was given to keep the numerals with resonable/common range for the given category (e.g., 0-100 for percentages; 1-12 for months). We also assure that the replacement will not create an illogical sentence (e.g., replacing "1920" with "1940" in "from 1920 to 1930")	manual	critical
	mistranslation gender	**He** has been unable to relieve the pain with medication, which the competition prohibits competitors from taking. **She** has been unable to relieve the pain with medication, which the competition prohibits competitors from taking.	Exactly one feminine pronoun in the sentence (such as "she" or "her") is being with a masculine pronouns (such as "he" or "him") or vice-versa. This includes reflexive pronouns (i.e., "him/herself") and possessive adjectives (i.e., "his/her").	automatic with manual check	critical
FLUENCY	**cohesion**	Scientists want to understand how planets have formed **since** a comet collided with Earth long ago, and especially how Earth has formed. Scientists want to understand how planets have formed _____ a comet collided with Earth long ago, and especially how Earth has formed.	A conjunction, such as "thus" or "therefore" is removed. Special attention was given to keep the rest of the sentence unperturbed.	automatic (spaCy) with manual check	minor
	grammar pos shift	The U.S. Supreme Court last year blocked the Trump **administration** from including the citizenship question on the 2020 census form. The U.S. Supreme Court last year blocked the Trump **administrate** from including the citizenship question on the 2020 census form.	Affix of the word is being changed keeping the stem kept constant (e.g., "bad" to "badly") which results in the part-of-speech shift. The degree to which the original meaning is affected varies, however, the intended meaning is easily retrivable from the stem and context.	manual	minor
	grammar swap order	I don't know if you realize that most of the goods imported **into this** country from Central America are duty free. I don't know if you realize that most of the goods imported **this into** country from Central America are duty free.	Two neighboring words are being swapped to mimic word order error.	automatic (spaCy)	minor
	grammar case	**She** announced that after a break of several years, a Rakoczy horse show will take place again in 2021. **Her** announced that after a break of several years, a Rakoczy horse show will take place again in 2021.	One pronoun in the sentence is being changed into a different, incorrect, case (e.g., "he" to "him").	automatic (spaCy) with manual check	minor
	grammar function word	Last month, **a** presidential committee recommended the resignation of the former CEP as part of measures to push the country toward new elections. Last month, **an** presidential committee recommended the resignation of the former CEP as part of measures to push the country toward new elections.	A preposition or article is being changed into an incorrect one to mimic mistake in function words usage. While most perturbations result in minor mistakes (i.e., the original meaning is easily retrivable) some may be more severe.	automatic with manual check	minor-major
	grammar tense	Cyanuric acid and melamine **were** both found in urine samples of pets who died after eating contaminated pet food. Cyanuric acid and melamine **are** both found in urine samples of pets who died after eating contaminated pet food.	A tense is being change into an incorrect one. We consider past, present, as well as the future tense (although this may be classified as modal verb in English)	manual	major
	grammar aspect	He **has been** unable to relieve the pain with medication, which the competition prohibits competitors from taking. He **is being** unable to relieve the pain with medication, which the competition prohibits competitors from taking.	Aspect is being changed to an incorrect one (e.g., perfective to progressive) *without* changing the tense.	manual	major
	grammar interrogative	**This is** the tenth time since the start of the pandemic that Florida's daily death toll has surpassed 100. **Is this** the tenth time since the start of the pandemic that Florida's daily death toll has surpassed 100?	Affirmative mood is being changed to interrogative mood.	manual	major
MIXED	**omission adj/adv**	Rangers **closely** monitor shooters participating in supplemental pest control trials as the trials are monitored and their effectiveness assessed. Rangers _____ monitor shooters participating in supplemental pest control trials as the trials are monitored and their effectiveness assessed.	An adjective or adverb is being removed. While in most cases this leads to	automatic (spaCy) with manual check	minor-major
	omission content verb	Catri **said** that 85% of new coronavirus cases in Belgium last week were under the age of 60. Catri _____ that 85% of new coronavirus cases in Belgium last week were under the age of 60.	Content verb is being removed (this excludes auxilary verbs and copulae).	Automatic with manual check	critical
	omission noun	In 1940 he stood up to other government **aristocrats** who wanted to discuss an "agreement" with the Nazis and he very ably won. In 1940 he stood up to other government _____ who wanted to discuss an "agreement" with the Nazis and he very ably won.	Noun, which is not a named entity or a subject, is being removed. We remove the head of the noun phrase including compound nouns.	automatic (spaCy) with manual check	critical
	omission subject	His **research** shows that the administration of hormones can accelerate the maturation of the baby's fetal lungs. His _____ shows that the administration of hormones can accelerate the maturation of the baby's fetal lungs.	Subject is being removed. We remove the head of the noun phrase including compound nouns.	automatic (spaCy) with manual check	critical
	omission named entry	I don't know if you realize that most of the goods imported into this country from **Central America** are duty free. I don't know if you realize that most of the goods imported into this country from _____ are duty free.	Named entity, which is not a subject, is being removed.	automatic (Stanza) with manual check	critical

Table 2: A subset of perturbations in DEMETR along with examples (detailed changes are highlighted in purple). A full list of perturbations is provided in Table A1 and Table A2 in Appendix A.

subject deletion or replacement of a named entity.

3 Performance of MT evaluation metrics on DEMETR

We test the accuracy and sensitivity of *14* popular MT evaluation metrics on the perturbations in DEMETR. We include both traditional string-based metrics, such as BLEU or CHRF, as well as newer learned metrics, such as BLEURT and COMET. Within the latter category, we also include two reference-free metrics, which rely only on the source sentence and translation and open possibilities for a more robust MT evaluation. The rest of this section provides an overview of the evaluation metrics before analyzing our findings. Detailed results of each metric on every perturbation are found in Table A3.

3.1 Evaluation metrics

String-based metrics can be used to evaluate *any* language, provided the availability of a reference translation (see Table 3). Their score is a function of string overlap or edit-distance, though it may not be always easily interpretable (Müller, 2020). Only BLEU[16] allows for multiple references in order to account for many possible translations of a sentence; however, it is rarely used with more than one reference due to the lack of multi-reference datasets (Mathur et al., 2020). Learned metrics, on the other hand, are much less transparent. BERTSCORE relies on contextualized embeddings, while PRISM employs zero-shot paraphrasing. COMET and BLEURT directly fine-tune pretrained language models on human judgments provided as Direct Assessments or MQM annotations.[17]

3.2 Perturbation accuracy

First, we measure the *accuracy* of each metric on DEMETR. For each perturbation, we define the accuracy as the percentage of the time that $\texttt{SCORE}(r, t)$

[16]For all string-based metrics we use the HuggingFace implementations available at `https://huggingface.co/evaluate-metric`. In the case of BLEU, we use the SacreBLEU version 2.1.0 (Post, 2018).

[17]We use the HuggingFace implementation of BERTSCORE, BLEURT-20, COMET, and COMET-QE. For BLEURT-20, we use BLEURT-20, the most recent and recommended checkpoints, for COMET and COMET-QE we use the SOTA models from WMT21 shared task, wmt21-comet-mqm and wmt21-comet-qe-mqm checkpoints, and for BERTScore we use roberta-large. For PRISM, we use the implementation available at `https://github.com/thompsonb/prism`

Metric	# Params	Language
string-based metrics		
BLEU (Papineni et al., 2002)	–	any
CER (Morris et al., 2004)	–	any
CHRF (Popović, 2015)	–	any
CHRF2 (Popović, 2017)	–	any
METEOR (Banerjee and Lavie, 2005)	–	any
ROUGE-2 (Lin, 2004)	–	any
TER (Snover et al., 2006)	–	any
pre-trained metrics		
BARTSCORE (Yuan et al., 2021b)	406M	50
BERTSCORE (Zhang* et al., 2020)	355M	104
BLEURT-20 (Sellam et al., 2020b)	579M	104
COMET (Rei et al., 2021)	580M	100
PRISM (Thompson and Post, 2020)	745M	39
pre-trained reference-free metrics		
COMET-QE (Rei et al., 2021)	569M	100
PRISM-QE (Thompson and Post, 2020)	745M	39

Table 3: Details of metrics tested on DEMETR. We report the parameter count for the largest available checkpoint of each learned metric. For learned metrics, we report the maximum number of languages that each can accept as input. While most of the learned metrics leverage pretrained multilingual language models (e.g., mBERT), it is important to note that they have not been validated against human judgments of MT quality on all of these languages (e.g., BLEURT-20 is only validated on 13 languages).

Metric	Base	Crit.	Maj.	Min.	All
string-based metrics					
BLEU	100.00	80.29	83.43	72.49	78.70
CER	99.15	80.37	83.59	80.20	81.88
CHRF	100.00	91.13	90.89	81.23	87.54
CHRF2	100.00	91.27	92.21	83.68	**88.80**
METEOR	100.00	82.95	79.69	58.97	73.60
ROUGE-2	99.90	76.91	80.99	47.10	66.58
TER	99.20	72.57	77.93	59.13	69.39
learned metrics					
BARTSCORE	100.00	95.11	89.68	79.48	88.16
BERTSCORE	100.00	98.11	96.22	98.50	98.11
BLEURT-20	100.00	98.78	95.63	97.98	98.06
COMET	100.00	96.24	92.96	93.46	94.83
PRISM	100.00	98.74	97.51	99.44	**98.92**
COMET-QE	77.80	84.49	76.73	89.85	85.16
PRISM-QE	97.40	96.70	95.68	99.21	97.63

Table 4: Accuracy on DEMETR perturbations for both string-based and learned metrics, shown bucketed by error severity (baseline, critical, major, and minor errors) as well as averaged across all perturbations. Baseline accuracies were computed excluding the *reference as translation* identity perturbation. Detailed accuracies for all perturbations along with the significance testing are shown in Table A3 in the Appendix A.

is greater than $\texttt{SCORE}(r, t')$.[18] Since all perturbed

[18]We do not give metrics credit for giving an equal score to both perturbed and unperturbed sentences.

sentences are *less correct* versions of the original machine translation, we expect all metrics to perform well on this task. Table 4 contains the accuracies averaged across both error severity as well as overall. Interesting results include:

Learned metrics achieve higher accuracy than string-based ones: All but two learned metrics (BARTSCORE and COMET-QE) achieve around or over *95%* accuracy,[19] which is to be expected, as each perturbation *clearly* affects the quality of the translation, though to varying degrees. PRISM is the most accurate metric on DEMETR, reaching an accuracy of *98.92%*. Performance of string-based metrics, on the other hand, is alarmingly bad. BLEU, often the *only* metric employed to evaluate the MT output (Marie et al., 2021), achieves an overall accuracy of only *78.70%*. To illustrate their struggles, the accuracy of string-based metrics ranges from *54%* to *84%* on the adjective/adverb removal perturbation, where a single adjective or adverb is omitted.

The best performing string-based metric is CHRF2, which corroborates results reported in Kocmi et al. (2021).

PRISM-QE achieves better accuracy than COMET-QE for reference-free metrics: Of the two reference-free metrics we evaluate, we notice that COMET-QE struggles with some perturbations. Most notably, its accuracy when given a *random* translation (i.e., a translation that does not match the source sentence) oscillates around *50%* (chance level) across all languages. Furthermore, COMET-QE shows low accuracy on gender (i.e., masculine pronouns replaced with feminine pronouns or vice-versa), number (i.e., a number replaced for another, reasonable number), and interrogatives (i.e., change of affirmative mood into interrogative mood). COMET-QE also strongly prefers (*88%*) the translation stripped of final punctuation over the complete sentence, in comparison to *0%* for PRISM-QE. In terms of accuracy, PRISM-QE performs exceptionally well on all perturbations, achieving lower accuracies (yet still around *80%*) only for Hindi—a language it was not trained on.

[19]This is true even for PRISM-QE, whose base neural MT model does not support Hindi but still manages to perform decently without the source.

4 Sensitivity analysis

While the accuracy of a metric on DEMETR is useful to know, it also obscures the *sensitivity* of a metric to a particular perturbation. Are metrics more sensitive to CRITICAL errors than MINOR ones? Are different *learned* metrics comparatively more or less sensitive to a particular perturbation? In this section, we explore these questions and highlight interesting observations, focusing primarily on the behavior of learned metrics.

Measuring sensitivity: Since each of our metrics has a different score range, we cannot naïvely just compare their score differences to analyze sensitivity. Instead, we compute a ratio that intuitively answers the following question: how much does SCORE drop on this perturbation compared to the catastrophic error of producing an empty string? We choose the empty string as a control since it is the perturbation that results in the largest SCORE drop for most metrics. Concretely, for a given reference translation r_i, machine translation t_i, and perturbed translation t'_i, we compute a ratio z_i as:

$$z_i = \frac{\texttt{SCORE}(r_i, t_i) - \texttt{SCORE}(r_i, t'_i)}{\texttt{SCORE}(r_i, t_i) - \texttt{SCORE}(r_i, \textit{empty string})} \tag{1}$$

Then, for each perturbation category, we aggregate the example-level ratios to obtain z by simply taking a mean, $z = \sum_i \frac{z_i}{N}$, where N is the number of examples for that perturbation (in most cases, *1*K).[20] Figure 2 contains a heatmap plotting this z ratio for each perturbation and learned metric, and forms the core of the following analysis.

BERTSCORE is relatively more sensitive to some minor errors than it is to critical errors: Although we observe that BERTSCORE drops only by a small absolute number for most perturbations, it is actually quite sensitive to many perturbations, especially when passing an unrelated translation and a shuffled version of the existing translation – two of the most drastic perturbations. It also shows higher sensitivity to untranslated words (i.e., codemixing) than to the remaining perturbations, which is to be expected as BERTSCORE uses a multilingual model. However, its sensitivity

[20]The ratio is a reasonable but also a rough estimate of metric sensitivity. Since it depends highly on the scores given by the metric to an empty string, we also make sure that all tested metrics achieve an accuracy close to 100% and can significantly distinguish between an empty string and the actual translation.

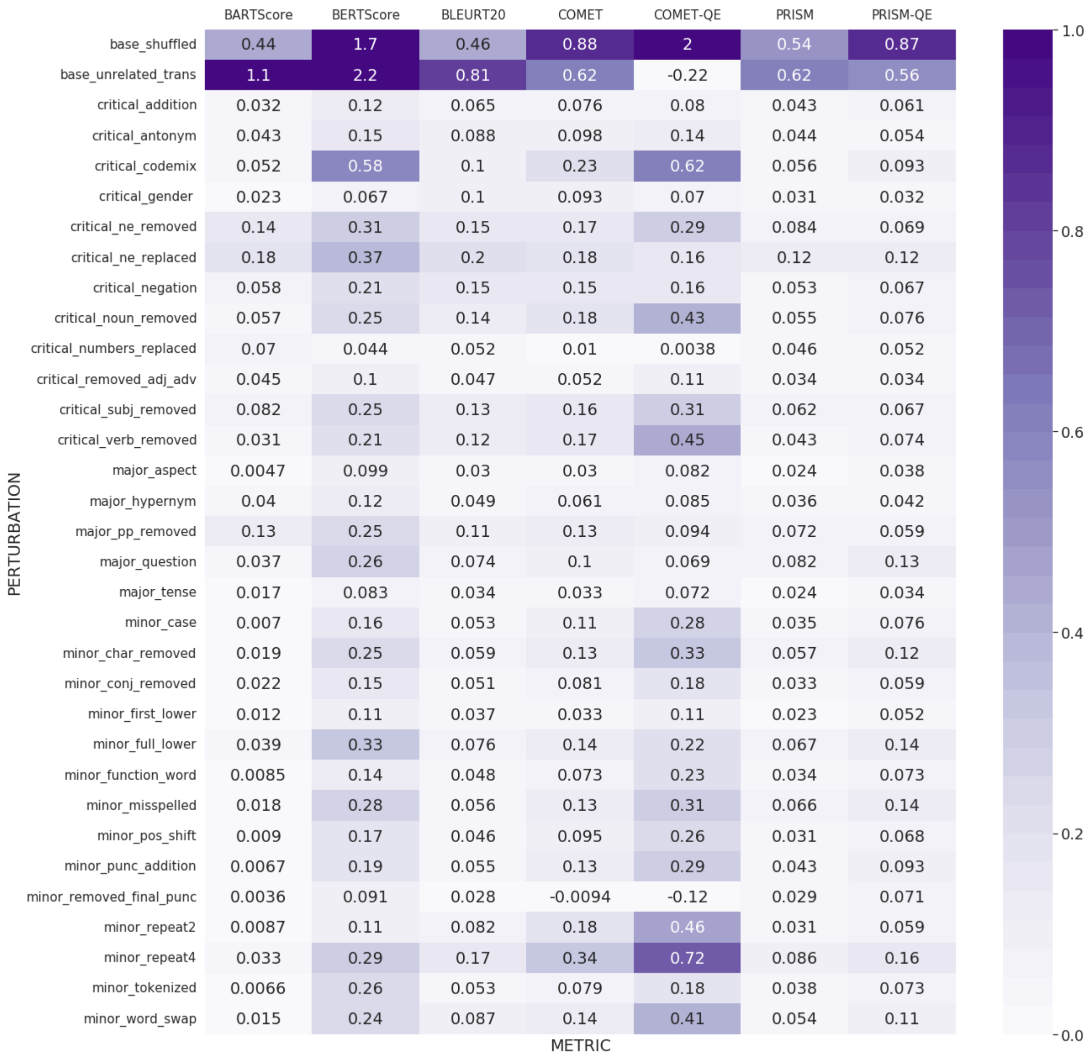

Figure 2: A heatmap of the sensitivity of learned metrics to different perturbations in DEMETR. The numbers are the ratios z computed as described in Section 4. Higher values denote higher relative sensitivity to the perturbation and are marked by a darker color. The error severity categories are arranged from ***minor*** (bottom part) through ***major*** (middle part) to ***critical*** (upper part). The last two errors are baselines.

to incorrect numbers (*0.044*), gender information (*0.067*), or aspect change (*0.099*) is lower than sensitivity to less severe errors, such as tokenized sentence (*0.26*) or lower-cased sentence (*0.33*) – a trend visible in other metrics, though not to such an extent.

COMET-QE, a metric adapted to MQM scoring, does not perform well on DEMETR: COMET-QE trained on MQM ratings (i.e., on the identification of mistakes similar to those included in DEMETR) varies in its sensitivity to perturbations. While it is sensitive to a sentence with shuffled words, it is not sensitive to a different, unrelated translation (an observation in line with its accuracy). COMET-QE also seems to be insensitive to minor errors such as the removal of the final punctuation, but also to some major or critical errors such as gender and number replacement.[21] Furthermore, COMET-QE is much more sensitive to word repetition (*0.46-0.72*) and word swap (*0.41*) than to some critical or major errors, such as named entity replacement (*0.16*) or sentence negation (*0.16*). Overall, COMET-QE behaves very differently from most of the other metrics, and in ways that are difficult to explain.

Overall, all metrics struggle to differentiate between minor and critical errors: While all metrics other than COMET-QE are very sensitive to the two baselines (different translation and shuf-

[21] Welsch t-test also reveals that the difference between the scores for the original MT and perturbed text is not significant (p-val>*.05*)

fled words) when compared to other perturbations (*0.44-2.20*), they struggle to differentiate the severity of some critical errors, such as an addition of a plausible but meaning-changing word (*0.032-0.12*) or incorrect number (*0.0038-0.07*). These ratios are lower than of some minor errors such as a word repeated four times (*0.086-0.72*). In fact, BERTSCORE, COMET, and COMET-QE are *more* sensitive to word repetition than to an addition of a word which ultimately critically changes the meaning.

5 Related Work

Our work builds on the previous efforts to analyze the performance of MT evaluation metrics, as well as efforts to curate diagnostic datasets for NLP.

Analysis of MT evaluation metrics: Fomicheva and Specia (2019) show that metric performance varies significantly across different levels of MT quality. Freitag et al. (2020) demonstrate the importance of reference quality during evaluation. Kocmi et al. (2021) investigate the performance of pretrained and string-based metrics, and conclude that learned metrics outperform string-based metrics, with COMET being the best-performing metric at the time. However, Amrhein and Sennrich (2022) explore COMET models in more depth finding, just as in the current study, that the models are *not* sensitive to number and named entity errors. Hanna and Bojar (2021), on the other hand, find that BERTSCORE is more robust to errors in major content words, and less so to small errors. Finally, Kasai et al. (2021) introduce a leaderboard for generation tasks that ensembles many of the metrics used here.

Diagnostic datasets: A number of previous studies employed diagnostic tests to explore the performance of NLP models. Marvin and Linzen (2018) evaluate abilities of LSTM based language models to rate grammatical sentence higher than ungrammatical ones by curating a dataset of minimal pairs in English. Warstadt et al. (2020) also utilize the concept of linguistic minimal pairs to evaluate the sensitivity of language models to various linguistic errors. Ribeiro et al. (2020) curate a checklist of perturbations to test the robustness of general NLP models.

Specia et al. (2010) introduce a simplified dataset of translations by four MT systems annotated for their quality in order to evaluate MT evaluation metrics. Sai et al. (2021b) also propose a checklist-style method to test the robustness of evaluation metrics for MT; however, they limit themselves to Chinese-to-English translation. Furthermore, many of the perturbations introduced in Sai et al. (2021b) does not control for a single aspect, as DEMETR does, and are not manually verified. Macketanz et al. (2018), on the other hand, design a linguistic test suite to evaluate the quality of MT from German to English, which WMT21 (Barrault et al., 2021) utilizes as a challenge dataset for MT evaluation metrics. Finally, Barrault et al. (2021) create a nine-category challenge set from a Chinese to English corpus, in order to test MT evaluation metrics, that are being submitted to the shared task.

6 Conclusion

We present DEMETR, a dataset designed to diagnose MT evaluation metrics. DEMETR consists of *31*K semi-automatically generated perturbations that cover *35* different linguistic phenomena. Our experiments showed that learned metrics are notably better than *any* string-based metrics at distinguishing perturbed from unperturbed translations, which confirms results reported in other studies (Kocmi et al., 2021; Fomicheva and Specia, 2019). We further explore the sensitivity of learned metrics, showing that even the best-performing metrics struggle to distinguish between minor errors such as word repetition and critical errors such as incorrect number, aspect, and gender. We will publicly release DEMETR to spur more informed future development of machine translation evaluation metrics.

Limitations

While DEMETR incorporates a wide range of linguistic phenomena, including various semantic, pragmatic, and morphological errors, all examples included in DEMETR are of translations *into*-English. It is likely that other translation directions may introduce other errors or metrics may be more/less sensitive to them. Furthermore, we decided to utilize sentence level translation as most metrics evaluate the translation on the sentence level and to highlight specific errors, which could be less apparent in the paragraph level setup. However, sentence level data cannot model discourse level errors, which remain an open problem in both machine translation and its evaluation. Furthermore, as DEMETR was constructed using WMT and FLORES the domains incorporated in DEMETR are restricted to the ones present in these two datasets (i.e., mostly news and informational materials). Finally, even though in most cases multiple correct translations of the source sentence exist, we provide only one reference. We decided not to include multiple reference due to the time restrictions as well as the fact that the only metric currently supporting multiple references is BLEU.

Ethical Considerations

Some perturbations were conducted manually with a help of freelancers hired on Upwork. The freelancers were informed of the purpose of this experiment. They were paid an equivalent of $15 per hour. We also adjusted this hourly rate to cover the 20% Upwork charge, which the platform charges the freelancers.

Acknowledgements

We would like to show our appreciation for the annotators hired on Upwork who took time and effort to fully understand the project and who helped us to assure good quality of the data. We would also like to thank Sergiusz Rzepkowski, Paula Kurzawska, and Kalpesh Krishna for their help in verifying the quality of translation and/or the perturbations. Finally, we would like to thank the reviewers for their constructive comments which helped to shape this paper, as well as UMass NLP community for their insights and discussions during this project. This project was partially supported by awards IIS-1955567 and IIS-2046248 from the National Science Foundation (NSF).

References

Farhad Akhbardeh, Arkady Arkhangorodsky, Magdalena Biesialska, Ondřej Bojar, Rajen Chatterjee, Vishrav Chaudhary, Marta R. Costa-jussa, Cristina España-Bonet, Angela Fan, Christian Federmann, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Leonie Harter, Kenneth Heafield, Christopher Homan, Matthias Huck, Kwabena Amponsah-Kaakyire, Jungo Kasai, Daniel Khashabi, Kevin Knight, Tom Kocmi, Philipp Koehn, Nicholas Lourie, Christof Monz, Makoto Morishita, Masaaki Nagata, Ajay Nagesh, Toshiaki Nakazawa, Matteo Negri, Santanu Pal, Allahsera Auguste Tapo, Marco Turchi, Valentin Vydrin, and Marcos Zampieri. 2021. Findings of the 2021 conference on machine translation (WMT21). In *Proceedings of the Sixth Conference on Machine Translation*, pages 1–88, Online. Association for Computational Linguistics.

Chantal Amrhein and Rico Sennrich. 2022. Identifying weaknesses in machine translation metrics through minimum bayes risk decoding: A case study for comet.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Loïc Barrault, Magdalena Biesialska, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Matthias Huck, Eric Joanis, Tom Kocmi, Philipp Koehn, Chi-kiu Lo, Nikola Ljubešić, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Santanu Pal, Matt Post, and Marcos Zampieri. 2020. Findings of the 2020 conference on machine translation (WMT20). In *Proceedings of the Fifth Conference on Machine Translation*, pages 1–55, Online. Association for Computational Linguistics.

Loic Barrault, Ondrej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussa, Christian Federmann, Mark Fishel, Alexander Fraser, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Tom Kocmi, Andre Martins, Makoto Morishita, and Christof Monz, editors. 2021. *Proceedings of the Sixth Conference on Machine Translation*. Association for Computational Linguistics, Online.

Ondřej Bojar, Christian Buck, Chris Callison-Burch, Christian Federmann, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2013. Findings of the 2013 Workshop on Statistical Machine Translation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 1–44, Sofia, Bulgaria. Association for Computational Linguistics.

Ondřej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, Radu Soricut, Lucia Specia, and Aleš Tamchyna. 2014. Findings of the 2014 workshop on statistical machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 12–58, Baltimore, Maryland, USA. Association for Computational Linguistics.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Barry Haddow, Matthias Huck, Chris Hokamp, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Carolina Scarton, Lucia Specia, and Marco Turchi. 2015. Findings of the

2015 workshop on statistical machine translation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 1–46, Lisbon, Portugal. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Aljoscha Burchardt. 2013. Multidimensional quality metrics: a flexible system for assessing translation quality. In *Proceedings of Translating and the Computer 35*, London, UK. Aslib.

Chris Callison-Burch, Philipp Koehn, Christof Monz, and Josh Schroeder. 2009. Findings of the 2009 Workshop on Statistical Machine Translation. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 1–28, Athens, Greece. Association for Computational Linguistics.

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

Marina Fomicheva and Lucia Specia. 2019. Taking MT Evaluation Metrics to Extremes: Beyond Correlation with Human Judgments. *Computational Linguistics*, 45(3):515–558.

Gereon Frahling. 2022. Deepl translate: The world's most accurate translator.

Markus Freitag, David Grangier, and Isaac Caswell. 2020. BLEU might be guilty but references are not innocent. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 61–71, Online. Association for Computational Linguistics.

Markus Freitag, David Grangier, Qijun Tan, and Bowen Liang. 2021. Minimum bayes risk decoding with neural metrics of translation quality. *CoRR*, abs/2111.09388.

Sebastian Gehrmann, Elizabeth Clark, and Thibault Sellam. 2022. Repairing the cracked foundation: A survey of obstacles in evaluation practices for generated text.

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

Michael Hanna and Ondřej Bojar. 2021. A fine-grained analysis of BERTScore. In *Proceedings of the Sixth Conference on Machine Translation*, pages 507–517, Online. Association for Computational Linguistics.

Jungo Kasai, Keisuke Sakaguchi, Ronan Le Bras, Lavinia Dunagan, Jacob Morrison, Alexander R. Fabbri, Yejin Choi, and Noah A. Smith. 2021. Bidimensional leaderboards: Generate and evaluate language hand in hand. *CoRR*, abs/2112.04139.

Marvin Kaster, Wei Zhao, and Steffen Eger. 2021. Global explainability of BERT-based evaluation metrics by disentangling along linguistic factors. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8912–8925, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tom Kocmi, Christian Federmann, Roman Grundkiewicz, Marcin Junczys-Dowmunt, Hitokazu Matsushita, and Arul Menezes. 2021. To ship or not to ship: An extensive evaluation of automatic metrics for machine translation. *CoRR*, abs/2107.10821.

Moshe Koppel and Noam Ordan. 2011. Translationese and its dialects. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, page 1318–1326, USA. Association for Computational Linguistics.

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

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Vivien Macketanz, Renlong Ai, Aljoscha Burchardt, and Hans Uszkoreit. 2018. TQ-AutoTest – an automated test suite for (machine) translation quality. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Benjamin Marie, Atsushi Fujita, and Raphael Rubino. 2021. Scientific credibility of machine translation research: A meta-evaluation of 769 papers. 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 7297–7306, Online. Association for Computational Linguistics.

Rebecca Marvin and Tal Linzen. 2018. Targeted syntactic evaluation of language models. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202, Brussels, Belgium. Association for Computational Linguistics.

Nitika Mathur, Timothy Baldwin, and Trevor Cohn. 2020. Tangled up in BLEU: Reevaluating the evaluation of automatic machine translation evaluation metrics. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4984–4997, Online. Association for Computational Linguistics.

Andrew Morris, Viktoria Maier, and Phil Green. 2004. From wer and ril to mer and wil: improved evaluation measures for connected speech recognition.

Mathias Müller. 2020. Single training runs and estimates of variance.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Ricardo Rei, Ana C Farinha, Chrysoula Zerva, Daan van Stigt, Craig Stewart, Pedro Ramos, Taisiya Glushkova, André F. T. Martins, and Alon Lavie. 2021. Are references really needed? unbabel-IST 2021 submission for the metrics shared task. In *Proceedings of the Sixth Conference on Machine Translation*, pages 1030–1040, Online. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 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.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online. Association for Computational Linguistics.

Ananya B. Sai, Tanay Dixit, Dev Yashpal Sheth, Sreyas Mohan, and Mitesh M. Khapra. 2021a. Perturbation CheckLists for evaluating NLG evaluation metrics. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7219–7234, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ananya B Sai, Tanay Dixit, Dev Yashpal Sheth, Sreyas Mohan, and Mitesh M Khapra. 2021b. Perturbation checklists for evaluating nlg evaluation metrics. *arXiv preprint arXiv:2109.05771*.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020a. 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.

Thibault Sellam, Dipanjan Das, and Ankur P. Parikh. 2020b. Bleurt: Learning robust metrics for text generation. In *ACL*.

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

Lucia Specia, Nicola Cancedda, and Marc Dymetman. 2010. A dataset for assessing machine translation evaluation metrics. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Brian Thompson and Matt Post. 2020. Automatic machine translation evaluation in many languages via zero-shot paraphrasing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 90–121, Online. Association for Computational Linguistics.

Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. 2020. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020. BLiMP: The benchmark of linguistic minimal pairs for English. *Transactions of the Association for Computational Linguistics*, 8:377–392.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021a. Bartscore: Evaluating generated text as text generation.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021b. Bartscore: Evaluating generated text as text generation. In *Advances in Neural Information Processing Systems*, volume 34, pages 27263–27277. Curran Associates, Inc.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. 2019. MoverScore: Text generation evaluating with contextualized embeddings and earth mover distance. 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 563–578, Hong Kong, China. Association for Computational Linguistics.

A Appendix

ID	Category	Type	Example	Description	Application	Error Severity
1	ACCURACY	repetition	I don't know if you realize that most of the goods imported into this country from Central America are duty **free**. I don't know if you realize that most of the goods imported into this country from Central America are duty **free free**.	The last word is being repeated twice. Punctuation is added after the last repeated word.	automatic	minor
2		repetition	Gordon Johndroe, Bush's spokesman, referred to the North Korean commitment as "an important advance towards the goal of achieving verifiable denuclearization of the Korean **penisula**." Gordon Johndroe, Bush's spokesman, referred to the North Korean commitment as "an important advance towards the goal of achieving verifiable denuclearization of the Korean **penisula penisula penisula penisula**."	The last word is being repeated four times. Punctuation is added after the last repeated word.	automatic	minor
3		hypernym	The language most of the people working in the Vatican City use on a daily basis is Italian, and Latin is often used in religious **ceremonies.** The language most of the people working in the Vatican City use on a daily basis is Italian, and Latin is often used in religious **activities.**	A word translated by a too general term (undertranslation). Special care was given in order to assure the word used in perturbed text is more general, and incorrect, translation of the original word.	manual	major
4		untranslated	The Polish Air Force will eventually be equipped with 32 F-35 Lightning II fighters **manufactured** by Lockheed Martin. The Polish Air Force will eventually be equipped with 32 F-35 Lightning II fighters **produkowane** by Lockheed Martin.	One word is being left untranslated. We manually assure that each time only one word is left untranslated.	manual	major
5		completeness	She is **in custody** pending prosecution and trial; but any witness evidence could be negatively impacted because her image has been widely published. She is _____ pending prosecution and trial; but any witness evidence could be negatively impacted because her image has been widely published.	One prepositional phrase is being removed. Whenever possible, we remove the shortest prepositional phrase in order to assure that the perturbed sentence is not much shorter than the original translation.	automatic (Stanza) with manual check	major
6		addition	_____ Plants look their best when they are in a natural environment, so resist the temptation to remove "just one." **Power** plants look their best when they are in a natural environment, so resist the temptation to remove "just one."	One word is being added. We make sure that the added word does not disturb the grammaticality of the sentence but changes the meaning in a significant way.	manual	critical
7		antonym	He has been unable to relieve the **pain** with medication, which the competition prohibits competitors from taking. He has been unable to relieve the **pleasure** with medication, which the competition prohibits competitors from taking.	One word (noun, verb, adj., or adv.) is being changed to its antonym.	manual	critical
8		mistranslation - negation	Last month, a presidential committee **recommended** the resignation of the former CEP as part of measures to push the country toward new elections. Last month, a presidential committee **didn't recommend** the resignation of the former CEP as part of measures to push the country toward new elections.	Affirmative sentences are being changed into negations. Rare negations are being changed to affirmative sentences.	manual	critical
9		mistranslation - named entity	Late night presenter **Stephen Colbert** welcomed 17-year-old Thunberg to his show on Tuesday and conducted a lengthy interview with the Swede. Late night presenter **John Oliver** welcomed 17-year-old Thunberg to his show on Tuesday and conducted a lengthy interview with the Swede.	Named entity is replaced with another named entity from the same category (person, geographic location, and organization).	automatic (Stanza) with manual check	critical
10		mistranslation - numbers	The Chinese Consulate General in Houston was established in **1979** and is the first Chinese consulate in the United States. The Chinese Consulate General in Houston was established in **1997** and is the first Chinese consulate in the United States.	A number is being replaced with an incorrect one. Special attention was given to keep the numerals with reasonable/common range for the given category (e.g., 0-100 for percentages; 1-12 for months). We also assure that the replacement will not creat illogical sentence (e.g., replacing "1920" with "1940" in "from 1920 to 1930")	manual	critical
11		mistranslation - gender	**He** has been unable to relieve the pain with medication, which the competition prohibits competitors from taking. **She** has been unable to relieve the pain with medication, which the competition prohibits competitors from taking.	Exactly one feminine pronoun in the sentence (such as "she" or "her") is being with a masculine pronouns (such as "he" or "him") or vice-versa. This includes reflexive pronouns (i.e., "him/herself") and possessive adjectives (i.e., "his/her").	automatic with manual check	critical
12	FLUENCY	cohesion	Scientists want to understand how planets have formed **since** a comet collided with Earth long ago, and especially how Earth has formed. Scientists want to understand how planets have formed _____ a comet collided with Earth long ago, and especially how Earth has formed.	A conjunction, such as "thus" or "therefore" is removed. Special attention was given to keep the rest of the sentence unperturbed.	automatic (spaCy) with manual check	minor
13		grammar - pos shift	The U.S. Supreme Court last year blocked the Trump **administration** from including the citizenship question on the 2020 census form. The U.S. Supreme Court last year blocked the Trump **administrate** from including the citizenship question on the 2020 census form.	Suffix of the word is being changed keeping the root constant (e.g., "bad" to "badly") which results in the part-of-speech shift. The degree to which the original meaning is affected varies, however, the intended meaning is easily retrivable from the perturbed word.	manual	minor
14		grammar - order swap	I don't know if you realize that most of the goods imported **into this** country from Central America are duty free. I don't know if you realize that most of the goods imported **this into** country from Central America are duty free.	Two neighboring words are being swapped to mimic word order error.	automatic (spaCy)	minor
15		grammar - case	**She** announced that after a break of several years, a Rakoczy horse show will take place again in 2021. **Her** announced that after a break of several years, a Rakoczy horse show will take place again in 2021.	One pronoun in the sentence is being changed into a different, incorrect, case (e.g., "he" to "him").	automatic (spaCy) with manual check	minor
16		grammar - function word	Last month, **a** presidential committee recommended the resignation of the former CEP as part of measures to push the country toward new elections. Last month, **an** presidential committee recommended the resignation of the former CEP as part of measures to push the country toward new elections.	A preposition or article is being changed into an incorrect one to mimic mistake in function words usage. While most perturbations result in minor mistakes (i.e., the original meaning is easily retrivable) some may be more severe.	automatic with manual check	minor-major
17		grammar - tense	Cyanuric acid and melamine **were** both found in urine samples of pets who died after eating contaminated pet food. Cyanuric acid and melamine **are** both found in urine samples of pets who died after eating contaminated pet food.	A tense is being change into an incorrect one. We consider past, present, as well as the future tense (although this may be classified as modal verb in English)	manual	major
18		grammar - aspect	He **has been** unable to relieve the pain with medication, which the competition prohibits competitors from taking. He **is being** unable to relieve the pain with medication, which the competition prohibits competitors from taking.	Aspect is being changed to an incorrect one (e.g., perfective to progressive) *without* changing the tense.	manual	major
19		grammar - interrogative	This is the tenth time since the start of the pandemic that Florida's daily death toll has surpassed 100. Is this the tenth time since the start of the pandemic that Florida's daily death toll has surpassed 100?	Affirmative mood is being changed to interrogative mood.	manual	major
20	MIXED	omission - adj/adv	Rangers **closely** monitor shooters participating in supplemental pest control trials as the trials are monitored and their effectiveness assessed. Rangers _____ monitor shooters participating in supplemental pest control trials as the trials are monitored and their effectiveness assessed.	An adjective or adverb is being removed. While in most cases this leads to	automatic with manual check	minor-major
21		omission - content verb	Catri **said** that 85% of new coronavirus cases in Belgium last week were under the age of 60. Catri _____ that 85% of new coronavirus cases in Belgium last week were under the age of 60.	Content verb is being removed (this excludes auxilary verbs and copulae).	Automatic with manual check	critical
22		omission - noun	In 1940 he stood up to other government **aristocrats** who wanted to discuss an "agreement" with the Nazis and he very ably won. In 1940 he stood up to other government _____ who wanted to discuss an "agreement" with the Nazis and he very ably won.	Noun, which is not a named entity or a subject, is being removed. We remove the head of the noun phrase including compound nouns.	automatic (spaCy) with manual check	critical
23		omission - subject	His **research** shows that the administration of hormones can accelerate the maturation of the baby's fetal lungs. His _____ shows that the administration of hormones can accelerate the maturation of the baby's fetal lungs.	Subject is being removed. We remove the head of the noun phrase including compound nouns.	automatic (spaCy) with manual check	critical
24		omission - named entry	I don't know if you realize that most of the goods imported into this country from **Central America** are duty free. I don't know if you realize that most of the goods imported into this country from _____ are duty free.	Named entity, which is not a subject, is being removed.	automatic (Stanza) with manual check	critical

Table A1: A full list of perturbations included in DEMETR .

ID	Category	Type	Example	Description	Application	Error Severity
25	TYPOGRAPHY	spelling - misspelled	Scientists want to understand how planets have formed since a comet collided with Earth long ago, and **especially** how Earth has formed. Scientists want to understand how planets have formed since a comet collided with Earth long ago, and **expecially** how Earth has formed.	One word is being misspelled based on the list of most common misspelled words.[22] A word is considered a candidate for misspelling only up to 10 times.	automatic	minor
26		spelling - char removed	I don't know if you **realize** that most of the goods imported into this country from Central America are duty free. I don't know if you **realie** that most of the goods imported into this country from Central America are duty free.	A character in a word is being deleted. We consider only nouns, adverbs, adjectives, and verbs as candidates.	automatic	minor
27		punctuation - removed	When a satellite in space receives a call, it reflects it back almost immediately**.** When a satellite in space receives a call, it reflects it back almost immediately_	Final punctuation is being removed.	Automatic	minor
28		punctuation - added	Comets may have been the source of Earth's water and organic matter that can form proteins and sustain life. Comets may have been the source of Earth's**,** water and organic matter that can form proteins and sustain life.	A punctuation is being added.	Automatic	minor
29		tokenized	At 9:30 a.m. on July 26, the reporter saw at the scene of Jiangkouhe Lianxu that the local area had made various preparations before flood distribution. At 9:30 a.m. on July 2**6 , t**he reporter saw at the scene of Jiangkouhe Lianxu that the local area had made various preparations before flood distribution .	The sentence is tokenized.	Automatic	minor
30		lowercases - whole	For example, U.S. citizens in the Middle East may face different situations than Europeans or Arabs. for example, u.s. citizens in the middle east may face different situations than europeans or arabs.	The entire sentence is lowercased.	Automatic	minor
31		lowercases - first word	For example, U.S. citizens in the Middle East may face different situations than Europeans or Arabs. for example, U.S. citizens in the Middle East may face different situations than Europeans or Arabs.	The first word in a sentence is lowercased.	Automatic	minor
32	BASELINE	empty	In the next two instances they have proved Freudenberg the right, but the opposite part continues to fight today.	Empty string (since most automatic metrics will not allow an empty string we pass a full stop instead).	Automatic	base
33		different	I don't know if you realize that most of the goods imported into this country from Central America are duty free. **It was the last game for the All Blacks, who had won the trophy two weeks earlier.**	Unrelated translation.	Automatic	base
34		unintelligible	Cyanuric acid and melamine were both found in urine samples of pets who died after eating contaminated pet food. Pets urine in of and acid were both died melamine found pet after who eating food contaminated cyanuric samples.	Shuffled words.	Automatic	base
35		reference	Last month, a presidential committee recommended the resignation of the former CEP as part of measures to push the country toward new elections. Last month a presidential commission recommended the prior CEP's resignation as part of a package of measures to move the country towards new elections.	Reference passed as the translation.	Automatic	base

Table A2: Table A1 continued.

ID	perturbation	metric	type	Welsch *t*-test		df	accuracy
				t	*p*-val		
1	addition (repetition)	BLEU	string	2.98	0.003	1,992.12	93.2%
		METEOR	string	0.44	0.662	1,997.34	85.4%
		CHRF	string	1.00	0.316	1,997.04	89.9%
		CHRF2	string	0.96	0.337	1,997.23	92.7%
		TER	string	-3.88	<0.001	1,996.94	77.7%
		CER	string	-5.61	<0.001	1,995.16	88.8%
		ROUGE2	string	1.69	0.092	1,996.37	99.7%
		BERTSCORE	learned	8.43	<0.001	1,997.55	97.5%
		COMET-QE	learned	21.33	<0.001	1,991.48	99.1%
		COMET	learned	22.46	<0.001	1,973.93	99.1%
		BLEURT20	learned	24.43	<0.001	1,996.98	99.0%
		PRISM-QE	learned	6.86	<0.001	1,989.71	98.9%
		PRISM	learned	9.61	<0.001	1,996.17	99.9%
		BARTScore	learned	1.49	0.137	1,997.92	78.0%
2	addition (repetition)	BLEU	string	6.47	<0.001	1,960.90	95.7%
		METEOR	string	1.63	0.104	1,997.00	85.8%
		CHRF	string	3.88	<0.001	1,992.23	97.6%
		CHRF2	string	3.53	<0.001	1,993.49	98.5%
		TER	string	-13.93	<0.001	1,996.08	95.2%
		CER	string	-18.65	<0.001	1,992.80	96.1%
		ROUGE2	string	4.92	<0.001	1,984.62	99.7%
		BERTSCORE	learned	22.91	<0.001	1,990.44	99.9%
		COMET-QE	learned	34.03	<0.001	1,982.14	100.0%
		COMET	learned	42.55	<0.001	1,955.19	100.0%
		BLEURT20	learned	50.88	<0.001	1,991.41	99.9%
		PRISM-QE	learned	19.02	<0.001	1,945.31	98.7%
		PRISM	learned	27.18	<0.001	1,994.44	100.0%
		BARTScore	learned	5.76	<0.001	1,997.69	95.0%
3	hypernym (undertranslation)	BLEU	string	5.11	<0.001	1,767.45	69.5%
		METEOR	string	5.12	<0.001	1,785.42	66.3%
		CHRF	string	8.67	<0.001	1,777.75	89.7%
		CHRF2	string	7.93	<0.001	1,777.56	89.3%
		TER	string	-3.30	<0.001	1,784.05	53.2%
		CER	string	-4.05	<0.001	1,780.32	77.9%
		ROUGE2	string	5.73	<0.001	1,776.50	63.3%
		BERTSCORE	learned	8.86	<0.001	1,784.44	93.6%
		COMET-QE	learned	4.24	<0.001	1,785.74	78.1%
		COMET	learned	7.10	<0.001	1,784.74	91.2%
		BLEURT20	learned	13.80	<0.001	1,786.00	92.7%
		PRISM-QE	learned	4.46	<0.001	1,781.82	94.4%
		PRISM	learned	10.40	<0.001	1,785.87	95.7%
		BARTScore	learned	6.42	<0.001	1,781.16	90.5%
4	untranslated	BLEU	string	5.70	<0.001	1,973.31	73.1%
		METEOR	string	6.18	<0.001	1,995.59	72.5%
		CHRF	string	9.22	<0.001	1,989.48	95.2%
		CHRF2	string	8.56	<0.001	1,988.24	95.3%
		TER	string	-3.82	<0.001	1,994.43	58.3%
		CER	string	-4.53	<0.001	1,992.88	83.5%
		ROUGE2	string	6.35	<0.001	1,984.97	68.4%
		BERTSCORE	learned	36.69	<0.001	1,824.17	99.8%
		COMET-QE	learned	27.31	<0.001	1,994.98	98.3%
		COMET	learned	26.78	<0.001	1,997.69	99.2%
		BLEURT20	learned	24.84	<0.001	1,822.75	99.1%
		PRISM-QE	learned	10.38	<0.001	1,993.90	97.6%
		PRISM	learned	16.62	<0.001	1,988.33	99.8%
		BARTScore	learned	8.91	<0.001	1,991.19	90.8%
5	completeness (omitted pp)	BLEU	string	9.94	<0.001	1,748.59	79.2%
		METEOR	string	17.83	<0.001	1,777.96	89.4%
		CHRF	string	15.88	<0.001	1,777.31	93.1%
		CHRF2	string	15.77	<0.001	1,778.00	94.0%
		TER	string	-9.79	<0.001	1,715.36	76.1%
		CER	string	-8.48	<0.001	1,715.24	73.6%
		ROUGE2	string	8.42	<0.001	1,775.53	78.7%
		BERTSCORE	learned	18.07	<0.001	1,770.58	94.2%
		COMET-QE	learned	6.98	<0.001	1,777.40	80.2%
		COMET	learned	13.84	<0.001	1,777.28	95.5%
		BLEURT20	learned	25.70	<0.001	1,579.21	97.1%
		PRISM-QE	learned	6.14	<0.001	1,776.54	92.6%
		PRISM	learned	19.76	<0.001	1,743.67	96.1%
		BARTScore	learned	19.56	<0.001	1,670.43	96.2%
6	addition	BLEU	string	4.84	<0.001	1,979.16	93.0%
		METEOR	string	1.53	0.127	1,997.87	99.0%
		CHRF	string	3.12	0.002	1,992.92	89.4%
		CHRF2	string	3.06	0.002	1,993.62	91.8%
		TER	string	-4.74	<0.001	1,997.95	80.3%
		CER	string	-6.24	<0.001	1,996.99	92.9%
		ROUGE2	string	4.90	<0.001	1,987.76	99.8%
		BERTSCORE	learned	9.54	<0.001	1,994.95	98.5%
		COMET-QE	learned	4.78	<0.001	1,997.76	69.7%
		COMET	learned	9.25	<0.001	1,996.06	93.3%
		BLEURT20	learned	19.09	<0.001	1,996.53	97.1%
		PRISM-QE	learned	7.13	<0.001	1,991.72	98.3%
		PRISM	learned	13.58	<0.001	1,995.57	99.9%
		BARTScore	learned	5.56	<0.001	1,997.62	94.1%

ID	perturbation	metric	type	Welsch *t*-test *t*	*p*-val	df	accuracy
7	antonym	BLEU	string	5.47	<0.001	1,949.90	65.4%
		METEOR	string	5.67	<0.001	1,963.08	68.9%
		CHRF	string	7.11	<0.001	1,953.68	83.8%
		CHRF2	string	6.81	<0.001	1,954.25	84.0%
		TER	string	-3.49	<0.001	1,961.27	64.9%
		CER	string	-3.34	<0.001	1,959.91	76.8%
		ROUGE2	string	6.10	<0.001	1,952.61	57.7%
		BERTSCORE	learned	11.52	<0.001	1,962.23	98.3%
		COMET-QE	learned	6.61	<0.001	1,963.91	83.7%
		COMET	learned	12.11	<0.001	1,955.91	97.0%
		BLEURT20	learned	24.45	<0.001	1,939.39	98.7%
		PRISM-QE	learned	6.39	<0.001	1,954.01	96.7%
		PRISM	learned	13.85	<0.001	1,960.97	99.2%
		BARTScore	learned	7.34	<0.001	1,963.98	96.9%
8	mistranslation - negation	BLEU	string	7.32	<0.001	1,961.08	91.3%
		METEOR	string	3.36	<0.001	1,995.95	94.7%
		CHRF	string	4.88	<0.001	1,990.62	90.8%
		CHRF2	string	5.14	<0.001	1,990.35	94.6%
		TER	string	-8.26	<0.001	1,995.94	89.6%
		CER	string	-5.29	<0.001	1,995.38	91.0%
		ROUGE2	string	7.67	<0.001	1,979.28	96.3%
		BERTSCORE	learned	15.60	<0.001	1,995.96	99.6%
		COMET-QE	learned	6.86	<0.001	1,995.73	83.7%
		COMET	learned	18.35	<0.001	1,991.67	99.4%
		BLEURT20	learned	41.51	<0.001	1,987.24	99.8%
		PRISM-QE	learned	8.43	<0.001	1,977.03	96.1%
		PRISM	learned	16.17	<0.001	1,994.10	99.8%
		BARTScore	learned	9.44	<0.001	1,988.26	98.5%
9	mistranslation - named entry	BLEU	string	7.00	<0.001	1,339.02	90.5%
		METEOR	string	8.29	<0.001	1,364.57	89.3%
		CHRF	string	12.47	<0.001	1,362.50	98.7%
		CHRF2	string	11.54	<0.001	1,361.48	98.7%
		TER	string	-7.14	<0.001	1,361.50	83.8%
		CER	string	-7.58	<0.001	1,358.56	89.9%
		ROUGE2	string	8.73	<0.001	1,350.27	87.7%
		BERTSCORE	learned	25.35	<0.001	1,358.26	99.1%
		COMET-QE	learned	7.14	<0.001	1,365.67	85.4%
		COMET	learned	18.20	<0.001	1,363.20	98.8%
		BLEURT20	learned	43.02	<0.001	1,279.18	100.0%
		PRISM-QE	learned	12.00	<0.001	1,331.69	95.3%
		PRISM	learned	30.02	<0.001	1,348.23	99.7%
		BARTScore	learned	24.24	<0.001	1,336.08	100.0%
10	mistranslation - numbers	BLEU	string	4.44	<0.001	734.86	89.0%
		METEOR	string	3.97	<0.001	741.65	79.6%
		CHRF	string	2.99	0.003	740.92	96.5%
		CHRF2	string	3.47	<0.001	740.45	96.5%
		TER	string	-2.61	0.009	741.63	79.8%
		CER	string	-0.82	0.415	741.83	80.4%
		ROUGE2	string	4.90	<0.001	738.46	82.5%
		BERTSCORE	learned	2.05	0.041	742.00	98.9%
		COMET-QE	learned	0.16	0.871	741.99	53.2%
		COMET	learned	0.73	0.463	741.82	80.4%
		BLEURT20	learned	9.18	<0.001	739.98	98.7%
		PRISM-QE	learned	4.38	<0.001	741.10	99.5%
		PRISM	learned	8.46	<0.001	741.89	100.0%
		BARTScore	learned	7.21	<0.001	741.99	99.5%
11	mistranslation - gender	BLEU	string	2.17	0.031	221.29	84.1%
		METEOR	string	2.13	0.035	223.97	83.2%
		CHRF	string	1.08	0.283	223.80	87.6%
		CHRF2	string	1.38	0.169	223.72	90.3%
		TER	string	-1.56	0.120	223.87	83.2%
		CER	string	-0.49	0.628	223.98	85.0%
		ROUGE2	string	2.27	0.024	221.81	72.6%
		BERTSCORE	learned	1.95	0.052	223.87	99.1%
		COMET-QE	learned	1.35	0.178	223.51	61.9%
		COMET	learned	4.05	<0.001	222.69	96.5%
		BLEURT20	learned	8.41	<0.001	223.05	99.1%
		PRISM-QE	learned	1.09	0.277	223.93	97.3%
		PRISM	learned	3.16	0.002	223.97	100.0%
		BARTScore	learned	1.45	0.148	223.97	96.5%
12	cohesion	BLEU	string	4.43	<0.001	1,441.98	75.8%
		METEOR	string	5.13	<0.001	1,449.88	78.1%
		CHRF	string	4.50	<0.001	1,448.17	85.3%
		CHRF2	string	4.69	<0.001	1,448.04	84.6%
		TER	string	-2.33	0.020	1,444.72	57.9%
		CER	string	-1.88	0.060	1,443.24	70.4%
		ROUGE2	string	4.26	<0.001	1,447.85	62.0%
		BERTSCORE	learned	9.89	<0.001	1,448.55	93.9%
		COMET-QE	learned	7.03	<0.001	1,448.46	89.7%
		COMET	learned	8.33	<0.001	1,448.86	93.8%
		BLEURT20	learned	12.68	<0.001	1,448.58	95.0%
		PRISM-QE	learned	5.04	<0.001	1,449.60	97.8%
		PRISM	learned	8.57	<0.001	1,449.92	96.6%
		BARTScore	learned	3.25	0.001	1,449.46	83.1%

ID	perturbation	metric	type	Welsch t-test		df	accuracy
				t	p-val		
13	grammar - pos shift	BLEU	string	5.03	<0.001	1,972.02	63.3%
		METEOR	string	2.01	0.045	1,989.94	29.3%
		CHRF	string	3.82	<0.001	1,984.67	85.5%
		CHRF2	string	4.29	<0.001	1,983.23	85.9%
		TER	string	-3.12	0.002	1,987.02	56.0%
		CER	string	-1.84	0.067	1,989.04	79.4%
		ROUGE2	string	5.75	<0.001	1,974.55	60.3%
		BERTSCORE	learned	14.08	<0.001	1,983.09	96.3%
		COMET-QE	learned	12.20	<0.001	1,987.47	95.0%
		COMET	learned	11.36	<0.001	1,988.92	97.0%
		BLEURT20	learned	13.27	<0.001	1,987.72	96.6%
		PRISM-QE	learned	7.02	<0.001	1,989.26	98.8%
		PRISM	learned	9.68	<0.001	1,990.00	98.8%
		BARTScore	learned	1.60	0.110	1,989.99	72.9%
14	grammar - order swap	BLEU	string	7.51	<0.001	1,968.29	72.9%
		METEOR	string	2.99	0.003	1,997.13	69.9%
		CHRF	string	6.05	<0.001	1,984.74	82.2%
		CHRF2	string	5.92	<0.001	1,984.84	82.3%
		TER	string	-3.65	<0.001	1,993.55	62.9%
		CER	string	-4.35	<0.001	1,985.78	74.2%
		ROUGE2	string	9.80	<0.001	1,969.34	76.8%
		BERTSCORE	learned	18.87	<0.001	1,994.91	98.6%
		COMET-QE	learned	18.47	<0.001	1,996.51	97.8%
		COMET	learned	16.44	<0.001	1,996.53	98.4%
		BLEURT20	learned	24.50	<0.001	1,983.00	98.5%
		PRISM-QE	learned	11.91	<0.001	1,995.77	99.3%
		PRISM	learned	16.57	<0.001	1,997.94	99.3%
		BARTScore	learned	2.74	0.006	1,997.81	78.6%
15	grammar - case	BLEU	string	3.05	0.002	677.34	68.5%
		METEOR	string	2.80	0.005	683.76	63.0%
		CHRF	string	1.91	0.056	683.25	86.9%
		CHRF2	string	2.30	0.022	682.89	88.6%
		TER	string	-2.14	0.032	683.12	67.3%
		CER	string	-1.40	0.161	683.88	83.4%
		ROUGE2	string	3.44	<0.001	679.73	63.8%
		BERTSCORE	learned	7.35	<0.001	682.78	99.7%
		COMET-QE	learned	8.25	<0.001	683.95	97.7%
		COMET	learned	8.12	<0.001	683.13	98.8%
		BLEURT20	learned	9.00	<0.001	684.00	99.1%
		PRISM-QE	learned	5.32	<0.001	683.76	99.7%
		PRISM	learned	6.31	<0.001	684.00	99.7%
		BARTScore	learned	0.75	0.453	683.98	73.8%
16	grammar - function word	BLEU	string	6.07	<0.001	1,943.76	70.1%
		METEOR	string	5.22	<0.001	1,963.26	67.3%
		CHRF	string	3.74	<0.001	1,959.17	83.2%
		CHRF2	string	4.39	<0.001	1,958.09	85.2%
		TER	string	-3.67	<0.001	1,963.43	76.4%
		CER	string	-2.19	0.028	1,965.20	81.1%
		ROUGE2	string	6.47	<0.001	1,949.80	69.9%
		BERTSCORE	learned	11.34	<0.001	1,961.19	97.8%
		COMET-QE	learned	9.19	<0.001	1,962.86	88.6%
		COMET	learned	8.70	<0.001	1,965.91	91.8%
		BLEURT20	learned	13.79	<0.001	1,960.54	93.8%
		PRISM-QE	learned	7.26	<0.001	1,965.92	99.8%
		PRISM	learned	10.49	<0.001	1,965.89	99.3%
		BARTScore	learned	1.53	0.126	1,965.88	78.8%
17	grammar - tense	BLEU	string	6.19	<0.001	1,946.80	78.7%
		METEOR	string	3.66	<0.001	1,973.48	66.7%
		CHRF	string	4.56	<0.001	1,965.71	89.0%
		CHRF2	string	5.08	<0.001	1,964.10	89.7%
		TER	string	-5.37	<0.001	1,971.28	82.5%
		CER	string	-2.67	0.008	1,973.18	82.8%
		ROUGE2	string	7.06	<0.001	1,949.72	81.2%
		BERTSCORE	learned	6.82	<0.001	1,971.00	96.5%
		COMET-QE	learned	3.32	<0.001	1,973.76	82.8%
		COMET	learned	4.03	<0.001	1,970.80	92.7%
		BLEURT20	learned	10.30	<0.001	1,965.99	94.5%
		PRISM-QE	learned	4.00	<0.001	1,970.72	96.2%
		PRISM	learned	7.55	<0.001	1,972.27	98.5%
		BARTScore	learned	2.92	0.004	1,973.99	91.0%
18	grammar - aspect	BLEU	string	6.30	<0.001	1,945.00	92.3%
		METEOR	string	2.25	0.025	1,972.00	84.6%
		CHRF	string	5.14	<0.001	1,959.98	88.6%
		CHRF2	string	5.37	<0.001	1,960.69	90.6%
		TER	string	-6.95	<0.001	1,968.84	84.7%
		CER	string	-5.03	<0.001	1,969.92	91.4%
		ROUGE2	string	7.04	<0.001	1,952.27	96.6%
		BERTSCORE	learned	7.86	<0.001	1,969.14	97.1%
		COMET-QE	learned	3.12	0.002	1,972.00	78.0%
		COMET	learned	3.60	<0.001	1,969.97	89.4%
		BLEURT20	learned	9.15	<0.001	1,953.54	94.2%
		PRISM-QE	learned	4.49	<0.001	1,967.57	95.3%
		PRISM	learned	7.57	<0.001	1,969.38	97.3%
		BARTScore	learned	0.85	0.394	1,971.98	74.8%

ID	perturbation	metric	type	Welsch t-test t	p-val	df	accuracy
19	grammar - interrogative	BLEU	string	11.11	<0.001	1,851.84	97.4%
		METEOR	string	6.80	<0.001	1,917.78	91.4%
		CHRF	string	6.51	<0.001	1,911.45	94.1%
		CHRF2	string	8.64	<0.001	1,905.30	97.5%
		TER	string	-9.91	<0.001	1,911.39	93.2%
		CER	string	-7.11	<0.001	1,925.33	92.3%
		ROUGE2	string	8.22	<0.001	1,892.51	85.3%
		BERTSCORE	learned	21.07	<0.001	1,913.23	99.8%
		COMET-QE	learned	3.58	<0.001	1,924.70	64.5%
		COMET	learned	11.96	<0.001	1,918.07	96.1%
		BLEURT20	learned	22.11	<0.001	1,919.63	99.6%
		PRISM-QE	learned	13.43	<0.001	1,916.54	99.9%
		PRISM	learned	25.55	<0.001	1,924.74	100.0%
		BARTScore	learned	6.37	<0.001	1,925.93	96.0%
20	omission - adj/adv	BLEU	string	4.60	<0.001	1,833.15	70.1%
		METEOR	string	5.52	<0.001	1,842.82	76.1%
		CHRF	string	7.55	<0.001	1,837.42	84.2%
		CHRF2	string	6.93	<0.001	1,837.52	82.6%
		TER	string	-2.26	0.024	1,835.00	53.7%
		CER	string	-3.00	0.003	1,827.77	68.7%
		ROUGE2	string	4.48	<0.001	1,838.99	65.4%
		BERTSCORE	learned	7.73	<0.001	1,843.71	91.2%
		COMET-QE	learned	4.09	<0.001	1,842.82	80.7%
		COMET	learned	6.04	<0.001	1,843.56	91.5%
		BLEURT20	learned	13.07	<0.001	1,836.98	95.2%
		PRISM-QE	learned	3.86	<0.001	1,841.41	90.4%
		PRISM	learned	10.66	<0.001	1,841.29	93.7%
		BARTScore	learned	7.68	<0.001	1,844.00	93.4%
21	omission - content verb	BLEU	string	4.13	<0.001	1,917.46	61.7%
		METEOR	string	4.70	<0.001	1,927.87	63.8%
		CHRF	string	6.85	<0.001	1,923.46	80.7%
		CHRF2	string	6.21	<0.001	1,922.47	78.2%
		TER	string	-1.64	0.100	1,919.74	46.8%
		CER	string	-2.56	0.010	1,911.00	64.3%
		ROUGE2	string	3.62	<0.001	1,921.73	55.3%
		BERTSCORE	learned	16.81	<0.001	1,926.62	96.5%
		COMET-QE	learned	20.93	<0.001	1,922.05	98.4%
		COMET	learned	19.69	<0.001	1,929.97	98.8%
		BLEURT20	learned	30.30	<0.001	1,830.05	98.8%
		PRISM-QE	learned	7.61	<0.001	1,928.22	97.8%
		PRISM	learned	13.16	<0.001	1,929.99	96.5%
		BARTScore	learned	5.36	<0.001	1,928.31	85.9%
22	omission - noun	BLEU	string	5.42	<0.001	1,926.09	71.2%
		METEOR	string	7.52	<0.001	1,942.42	77.5%
		CHRF	string	9.81	<0.001	1,938.83	88.9%
		CHRF2	string	8.89	<0.001	1,938.37	87.1%
		TER	string	-3.66	<0.001	1,929.91	63.9%
		CER	string	-3.88	<0.001	1,911.86	68.7%
		ROUGE2	string	4.86	<0.001	1,938.91	64.1%
		BERTSCORE	learned	20.07	<0.001	1,941.62	97.9%
		COMET-QE	learned	20.86	<0.001	1,943.77	97.3%
		COMET	learned	21.39	<0.001	1,940.48	99.2%
		BLEURT20	learned	34.88	<0.001	1,811.22	99.2%
		PRISM-QE	learned	8.02	<0.001	1,941.14	98.2%
		PRISM	learned	16.77	<0.001	1,943.42	97.8%
		BARTScore	learned	9.61	<0.001	1,933.99	90.2%
23	omission - subject	BLEU	string	5.49	<0.001	1,932.96	74.1%
		METEOR	string	7.47	<0.001	1,954.60	80.1%
		CHRF	string	10.01	<0.001	1,951.54	91.3%
		CHRF2	string	9.47	<0.001	1,949.00	90.5%
		TER	string	-3.84	<0.001	1,942.84	67.3%
		CER	string	-4.87	<0.001	1,926.02	72.6%
		ROUGE2	string	5.39	<0.001	1,948.97	70.3%
		BERTSCORE	learned	19.94	<0.001	1,955.90	98.0%
		COMET-QE	learned	16.41	<0.001	1,955.97	94.7%
		COMET	learned	18.65	<0.001	1,951.01	98.5%
		BLEURT20	learned	32.39	<0.001	1,795.97	99.1%
		PRISM-QE	learned	8.17	<0.001	1,945.91	96.0%
		PRISM	learned	18.64	<0.001	1,949.93	98.3%
		BARTScore	learned	13.39	<0.001	1,901.80	91.8%
24	omission - named entry	BLEU	string	6.06	<0.001	1,336.10	80.4%
		METEOR	string	9.32	<0.001	1,351.99	94.4%
		CHRF	string	11.63	<0.001	1,351.88	97.5%
		CHRF2	string	10.83	<0.001	1,351.34	97.0%
		TER	string	-4.68	<0.001	1,340.39	72.7%
		CER	string	-5.05	<0.001	1,327.37	74.2%
		ROUGE2	string	6.20	<0.001	1,348.43	79.8%
		BERTSCORE	learned	21.78	<0.001	1,351.75	98.5%
		COMET-QE	learned	12.22	<0.001	1,351.92	91.3%
		COMET	learned	16.39	<0.001	1,349.64	98.5%
		BLEURT20	learned	32.33	<0.001	1,288.20	99.4%
		PRISM-QE	learned	6.69	<0.001	1,337.24	93.9%
		PRISM	learned	21.29	<0.001	1,345.33	99.0%
		BARTScore	learned	20.08	<0.001	1,321.04	98.8%

ID	perturbation	metric	type	Welsch *t*-test *t*	*p*-val	df	accuracy
25	spelling - misspelled	BLEU	string	5.04	<0.001	1,744.49	67.7%
		METEOR	string	5.27	<0.001	1,754.73	69.6%
		CHRF	string	3.77	<0.001	1,753.94	84.1%
		CHRF2	string	4.27	<0.001	1,752.53	84.2%
		TER	string	-3.26	0.001	1,754.63	64.6%
		CER	string	-0.93	0.353	1,755.79	70.6%
		ROUGE2	string	5.78	<0.001	1,744.41	64.1%
		BERTSCORE	learned	21.29	<0.001	1,748.10	99.8%
		COMET-QE	learned	14.06	<0.001	1,747.45	92.6%
		COMET	learned	14.60	<0.001	1,755.96	97.4%
		BLEURT20	learned	14.96	<0.001	1,750.64	97.6%
		PRISM-QE	learned	14.35	<0.001	1,750.99	99.8%
		PRISM	learned	19.00	<0.001	1,755.60	100.0%
		BARTScore	learned	2.86	0.004	1,755.76	79.7%
26	spelling - char removed	BLEU	string	5.69	<0.001	1,976.50	68.1%
		METEOR	string	5.52	<0.001	1,996.82	66.4%
		CHRF	string	3.48	<0.001	1,995.51	85.8%
		CHRF2	string	4.20	<0.001	1,993.62	86.0%
		TER	string	-3.45	<0.001	1,995.20	61.0%
		CER	string	-0.50	0.620	1,997.61	65.4%
		ROUGE2	string	6.55	<0.001	1,980.84	61.8%
		BERTSCORE	learned	19.73	<0.001	1,993.08	99.5%
		COMET-QE	learned	14.38	<0.001	1,994.47	95.4%
		COMET	learned	15.28	<0.001	1,997.31	98.2%
		BLEURT20	learned	16.73	<0.001	1,987.94	98.7%
		PRISM-QE	learned	12.91	<0.001	1,994.01	99.7%
		PRISM	learned	17.66	<0.001	1,997.73	99.7%
		BARTScore	learned	3.38	<0.001	1,997.57	80.1%
27	punctuation - removed	BLEU	string	2.07	0.038	1,996.48	76.2%
		METEOR	string	3.30	<0.001	1,989.43	57.3%
		CHRF	string	0.95	0.343	1,997.89	96.4%
		CHRF2	string	1.96	0.050	1,997.73	98.3%
		TER	string	-2.64	0.008	1,993.32	55.8%
		CER	string	-0.81	0.420	1,997.92	80.3%
		ROUGE2	string	0.00	1.000	1,998.00	0.0%
		BERTSCORE	learned	6.83	<0.001	1,997.83	98.5%
		COMET-QE	learned	-6.33	<0.001	1,987.30	12.0%
		COMET	learned	-1.10	0.273	1,997.74	39.8%
		BLEURT20	learned	8.32	<0.001	1,993.90	96.8%
		PRISM-QE	learned	6.29	<0.001	1,995.83	99.9%
		PRISM	learned	9.01	<0.001	1,997.79	100.0%
		BARTScore	learned	0.61	0.544	1,997.99	63.5%
28	punctuation - added	BLEU	string	5.01	<0.001	1,977.93	90.3%
		METEOR	string	5.67	<0.001	1,996.17	69.4%
		CHRF	string	2.61	0.009	1,995.62	97.9%
		CHRF2	string	2.78	0.006	1,995.28	99.0%
		TER	string	-3.68	<0.001	1,995.37	64.7%
		CER	string	-0.88	0.380	1,997.99	85.3%
		ROUGE2	string	0.00	1.000	1,998.00	0.0%
		BERTSCORE	learned	15.85	<0.001	1,983.20	99.3%
		COMET-QE	learned	14.16	<0.001	1,994.92	99.5%
		COMET	learned	15.63	<0.001	1,979.45	99.9%
		BLEURT20	learned	16.35	<0.001	1,997.90	99.3%
		PRISM-QE	learned	9.78	<0.001	1,995.95	99.7%
		PRISM	learned	13.43	<0.001	1,998.00	100.0%
		BARTScore	learned	1.21	0.225	1,997.97	76.9%
29	tokenized	BLEU	string	1.44	0.149	1,997.07	18.6%
		METEOR	string	9.42	<0.001	1,978.96	84.0%
		CHRF	string	0.00	1.000	1,998.00	0.0%
		CHRF2	string	0.45	0.651	1,997.72	23.7%
		TER	string	-17.87	<0.001	1,991.95	88.3%
		CER	string	-2.16	0.031	1,997.79	89.0%
		ROUGE2	string	0.14	0.888	1,998.00	1.2%
		BERTSCORE	learned	19.16	<0.001	1,995.29	99.8%
		COMET-QE	learned	8.88	<0.001	1,997.76	98.5%
		COMET	learned	9.42	<0.001	1,994.27	99.8%
		BLEURT20	learned	14.75	<0.001	1,985.71	99.5%
		PRISM-QE	learned	8.42	<0.001	1,991.80	98.4%
		PRISM	learned	11.73	<0.001	1,997.74	100.0%
		BARTScore	learned	1.17	0.241	1,997.86	69.6%
30	lowercase - whole	BLEU	string	14.04	<0.001	1,955.69	87.8%
		METEOR	string	0.00	1.000	1,998.00	0.0%
		CHRF	string	11.23	<0.001	1,990.65	90.6%
		CHRF2	string	13.39	<0.001	1,990.65	90.5%
		TER	string	0.00	1.000	1,998.00	0.0%
		CER	string	-2.67	0.008	1,996.14	87.4%
		ROUGE2	string	0.00	1.000	1,998.00	0.0%
		BERTSCORE	learned	25.36	<0.001	1,957.60	99.3%
		COMET-QE	learned	10.10	<0.001	1,984.92	97.1%
		COMET	learned	16.13	<0.001	1,990.72	98.1%
		BLEURT20	learned	22.51	<0.001	1,997.97	99.3%
		PRISM-QE	learned	14.11	<0.001	1,995.58	99.6%
		PRISM	learned	20.37	<0.001	1,993.17	99.9%
		BARTScore	learned	7.04	<0.001	1,997.80	93.6%

ID	perturbation	metric	type	Welsch t-test			accuracy
				t	p-val	df	
31	lowercase - first word	BLEU	string	2.25	0.024	1,994.50	66.7%
		METEOR	string	0.01	0.988	1,998.00	0.1%
		CHRF	string	0.96	0.336	1,997.79	71.9%
		CHRF2	string	1.74	0.082	1,997.24	72.0%
		TER	string	-0.01	0.992	1,998.00	0.0%
		CER	string	-0.70	0.482	1,997.86	71.4%
		ROUGE2	string	0.00	1.000	1,998.00	0.0%
		BERTSCORE	learned	8.48	<0.001	1,995.57	99.2%
		COMET-QE	learned	4.82	<0.001	1,998.00	95.0%
		COMET	learned	3.96	<0.001	1,997.61	96.3%
		BLEURT20	learned	11.19	<0.001	1,995.81	98.6%
		PRISM-QE	learned	5.05	<0.001	1,997.98	98.9%
		PRISM	learned	7.03	<0.001	1,997.96	99.0%
		BARTScore	learned	2.14	0.032	1,997.92	89.2%
32	empty	BLEU	string	66.14	<0.001	999.00	100.0%
		METEOR	string	135.10	<0.001	999.08	100.0%
		CHRF	string	166.71	<0.001	1,000.01	100.0%
		CHRF2	string	152.46	<0.001	1,005.75	100.0%
		TER	string	-85.53	<0.001	999.15	99.1%
		CER	string	-110.99	<0.001	999.20	100.0%
		ROUGE2	string	89.55	<0.001	999.00	99.9%
		BERTSCORE	learned	103.38	<0.001	1,643.56	100.0%
		COMET-QE	learned	68.56	<0.001	1,536.15	100.0%
		COMET	learned	153.74	<0.001	1,314.70	100.0%
		BLEURT20	learned	386.70	<0.001	1,296.35	100.0%
		PRISM-QE	learned	139.31	<0.001	1,606.26	100.0%
		PRISM	learned	303.20	<0.001	1,993.05	100.0%
		BARTScore	learned	142.38	<0.001	1,766.71	100.0%
33	different	BLEU	string	62.67	<0.001	1,001.22	100.0%
		METEOR	string	119.40	<0.001	1,113.17	100.0%
		CHRF	string	122.77	<0.001	1,087.45	100.0%
		CHRF2	string	121.52	<0.001	1,071.83	100.0%
		TER	string	-71.77	<0.001	1,988.46	99.3%
		CER	string	-67.59	<0.001	1,955.71	98.3%
		ROUGE2	string	88.50	<0.001	1,011.66	99.9%
		BERTSCORE	learned	184.64	<0.001	1,947.30	100.0%
		COMET-QE	learned	3.26	0.001	1,975.84	55.6%
		COMET	learned	81.20	<0.001	1,904.40	100.0%
		BLEURT20	learned	251.23	<0.001	1,978.21	100.0%
		PRISM-QE	learned	87.09	<0.001	1,556.55	94.8%
		PRISM	learned	191.18	<0.001	1,997.16	100.0%
		BARTScore	learned	147.07	<0.001	1,732.24	100.0%
34	unintelligible (shuffled)	BLEU	string	55.60	<0.001	1,047.83	100.0%
		METEOR	string	55.33	<0.001	1,519.93	99.4%
		CHRF	string	43.90	<0.001	1,592.64	100.0%
		CHRF2	string	45.40	<0.001	1,515.48	100.0%
		TER	string	-57.36	<0.001	1,595.93	99.1%
		CER	string	-63.66	<0.001	1,288.57	96.9%
		ROUGE2	string	75.75	<0.001	1,198.84	99.9%
		BERTSCORE	learned	134.81	<0.001	1,995.55	100.0%
		COMET-QE	learned	98.03	<0.001	1,962.18	100.0%
		COMET	learned	111.85	<0.001	1,910.48	100.0%
		BLEURT20	learned	128.11	<0.001	1,982.63	100.0%
		PRISM-QE	learned	106.85	<0.001	1,838.50	100.0%
		PRISM	learned	140.83	<0.001	1,847.24	100.0%
		BARTScore	learned	65.65	<0.001	1,828.42	100.0%
35	reference	BLEU	string	-90.34	<0.001	999.00	100.0%
		METEOR	string	-60.37	<0.001	999.00	100.0%
		CHRF	string	-76.50	<0.001	999.00	100.0%
		CHRF2	string	-78.24	<0.001	999.00	100.0%
		TER	string	63.71	<0.001	999.00	100.0%
		CER	string	56.77	<0.001	999.00	100.0%
		ROUGE2	string	-77.52	<0.001	999.00	100.0%
		BERTSCORE	learned	-61.70	<0.001	999.00	100.0%
		COMET-QE	learned	1.43	0.152	1,997.61	44.4%
		COMET	learned	-25.94	<0.001	1,983.70	100.0%
		BLEURT20	learned	-94.47	<0.001	1,160.41	100.0%
		PRISM-QE	learned	7.59	<0.001	1,991.89	14.3%
		PRISM	learned	-50.57	<0.001	1,159.77	99.4%
		BARTScore	learned	-38.54	<0.001	1,248.78	99.8%

Table A3: A two-samples Welsch t-test is conducted on each metric to compare $\texttt{SCORE}(r, t)$ and $\texttt{SCORE}(r, t')$ (see Section 2.1) of each perturbation type. The tests are implemented in Python using the package `scipy` (Virtanen et al., 2020). Degrees of Freedom (DF) are estimated using the Welch-Satterthwaite equasion for Degrees of Freedom. The accuracy on the baseline perturbation *35* (reference as translation) was reversed, as one can expect the metric to prefer translation identical with the reference.