ACES: Translation Accuracy Challenge Sets for Evaluating Machine Translation Metrics

Chantal Amrhein[1*] and **Nikita Moghe**[2*] and **Liane Guillou**[2*]
[1]Department of Computational Linguistics, University of Zurich
[2]School of Informatics, University of Edinburgh
amrhein@cl.uzh.ch, nikita.moghe@ed.ac.uk, lguillou@ed.ac.uk

Abstract

As machine translation (MT) metrics improve their correlation with human judgement every year, it is crucial to understand the limitations of these metrics at the segment level. Specifically, it is important to investigate metric behaviour when facing accuracy errors in MT because these can have dangerous consequences in certain contexts (*e.g.,* legal, medical). We curate ACES, a translation accuracy challenge set, consisting of 68 phenomena ranging from simple perturbations at the word/character level to more complex errors based on discourse and real-world knowledge. We use ACES to evaluate a wide range of MT metrics including the submissions to the WMT 2022 metrics shared task and perform several analyses leading to general recommendations for metric developers: consider a) combining metrics with different strengths, b) explicitly modelling additional language-specific information beyond what is available via multilingual embeddings.

1 Introduction

Challenge sets have been developed for measuring the success of systems or metrics on a particular phenomenon of interest for a range of NLP tasks, including but not limited to: Sentiment Analysis[1] (Li et al., 2017; Mahler et al., 2017; Staliūnaitė and Bonfil, 2017), Natural Language Inference (McCoy and Linzen, 2018; Rocchietti et al., 2021), Question Answering (Ravichander et al., 2021), Machine Reading Comprehension (Khashabi et al., 2018), Machine Translation (MT) (King and Falkedal, 1990; Isabelle et al., 2017), and the more specific task of pronoun translation in MT (Guillou and Hardmeier, 2016). They are useful to compare the performance of different systems, or to identify performance improvement/degradation between a modified system and a previous iteration.

We describe the University of Zurich - University of Edinburgh submission to the *Challenge Sets* subtask of the WMT 2022 metrics shared task. Our translation accuracy challenge sets (ACES) consist of 36,499 examples covering 146 language pairs and representing challenges from 68 phenomena. We focus on translation accuracy errors and base the phenomena covered in our challenge set on the Multidimensional Quality Metrics (MQM) ontology (Lommel et al., 2014). We include phenomena ranging from simple perturbations involving the omission/addition of characters or tokens, to more complex examples involving mistranslation e.g. ambiguity and hallucinations in translation, untranslated elements of a sentence, discourse-level phenomena, and real-world knowledge.

We evaluate the metrics submitted to the WMT 2022 metrics shared task and a range of baseline metrics on ACES. Additionally, we perform an extensive analysis, which aims to reveal:

1. The extent to which reference-based and reference-free metrics take into account the source sentence context.
2. The extent to which reference-based metrics rely on surface-level overlap with the reference.
3. Whether using multilingual embeddings results in better metrics.

Based on our analysis, we recommend that metric developers consider: a) combining metrics with different strengths in the form of ensemble models, b) explicitly modelling additional language-specific information beyond what is available via multilingual embeddings. We also propose that ACES be used as a benchmark for developing evaluation metrics for MT to a) monitor which error categories can be identified better, and b) whether there are any categories for which metric performance worsens.

*Equal contribution by all authors.
[1]Submitted to the EMNLP 2017 "Build It Break It" shared task on sentiment analysis

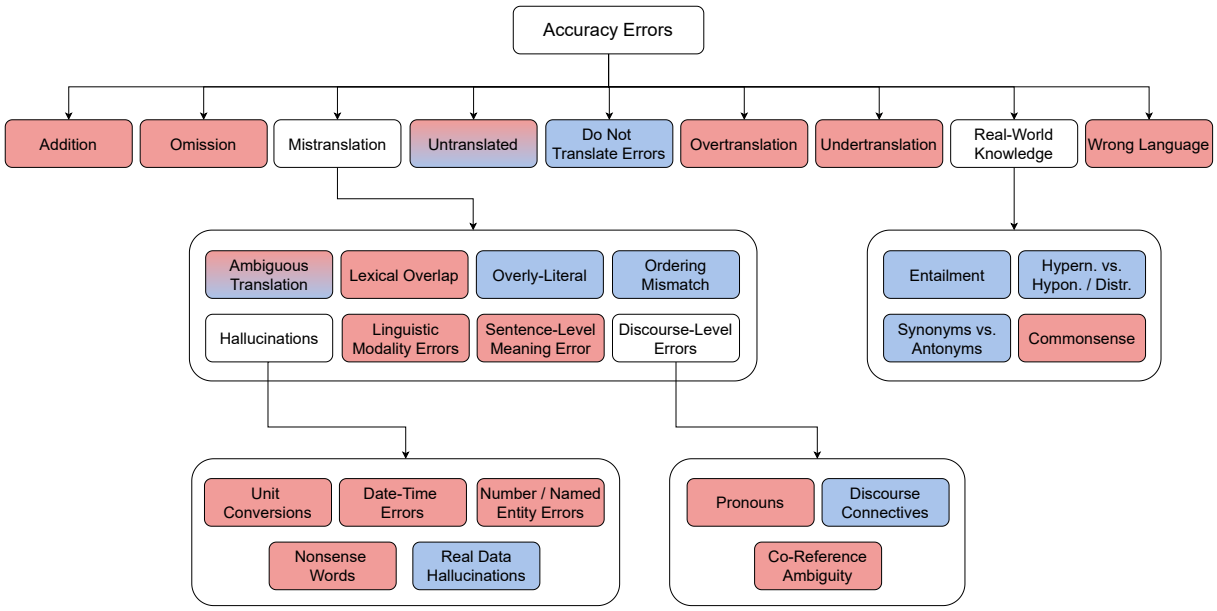

Figure 1: Diagram of the error categories on which our collection of challenge sets is based. Red means challenge sets are created automatically, blue means challenge sets are created manually.

2 Motivation

With the advent of neural networks and especially Transformer-based architectures (Vaswani et al., 2017), machine translation outputs have become more and more fluent (Bentivogli et al., 2016; Toral and Sánchez-Cartagena, 2017; Castilho et al., 2017). Fluency errors are also judged less severely than accuracy errors by human evaluators (Freitag et al., 2021a) which reflects the fact that accuracy errors can have dangerous consequences in certain contexts, for example in the medical and legal domains (Vieira et al., 2021).

For these reasons, we decide to build a challenge set focused on accuracy errors. Specifically, we use the hierarchy of errors under the class *Accuracy* from the MQM ontology to design these challenge sets. We extend this ontology by two error classes (translations defying real-world knowledge and translations in the wrong language) and specify several more specific subclasses such as discourse-level errors or ordering mismatches. A full overview of all error classes can be seen in Figure 1. Our challenge set consists of synthetically generated adversarial examples, examples from repurposing contrastive MT test sets (both marked in red), and manually annotated examples (marked in blue). To create the challenge sets, we use test sets from tasks such as adversarial paraphrase detection, natural language inference, and contrastive MT test sets created independently of the WMT shared tasks to avoid overlap with the data that is used to train neural evaluation metrics.

Another aspect we focus on is including a broad range of language pairs in ACES. Whenever possible we create examples for all language pairs covered in a source dataset when we use automatic approaches. For phenomena where we create examples manually, we also aim to cover at least two language pairs per phenomenon, but are of course limited to the languages spoken by the authors.

Finally, we aim to offer a collection of challenge sets covering both easy and hard phenomena. While it may be of interest to the community, to continuously test on harder examples to check where machine translation evaluation metrics still break, we believe that easy challenge sets are just as important to ensure that metrics do not suddenly get worse at identifying error types that we previously considered as "solved". Therefore, we take an holistic view when creating ACES and do not filter out individual examples or exclude challenge sets based on baseline metric performance or other factors.

We first discuss previous efforts to create challenge sets (Section 3), before giving a broad overview of the datasets used to construct ACES (Section 4) and discussing the individual challenge sets in more detail and presenting examples for each (Section 5).

3 Related Work

Challenge sets are used to study a particular phenomenon of interest rather than the general distribution of phenomena in standard test sets (Popović and Castilho, 2019). The earliest introduction of challenge sets was by King and Falkedal (1990) who probed acceptability of machine translations for different domains. Challenge sets have been prevalent in different fields within NLP such as parsing (Rimell et al., 2009), NLI (McCoy and Linzen, 2018; Rocchietti et al., 2021), question answering (Ravichander et al., 2021), reading comprehension (Khashabi et al., 2018) and sentiment analysis (Li et al., 2017; Mahler et al., 2017; Staliūnaitė and Bonfil, 2017), to name a few. These challenge sets provide insights on whether the state-of-the-art models are robust to domain shifts, linguistic phenomena like negation/commonsense or identify whether these models rely on shallow heuristics. Another line of work under "adversarial datasets" also focuses on creating examples by perturbing the standard test test to fool the model (Smith (2012); Jia and Liang (2017), *inter-alia*).

Challenge sets for evaluating MT models have focused on the translation models' ability to generate the correct translation under the phenomenon of interest. These include word sense ambiguity (Vamvas and Sennrich, 2021), gender bias (Rudinger et al., 2017; Zhao et al., 2018; Stanovsky et al., 2019), structural divergence (Isabelle et al., 2017) and discourse level phenomena (Guillou and Hardmeier, 2016; Emelin and Sennrich, 2021).

While such challenge sets focus on evaluating specific machine translation models, it is necessary to identify whether the existing machine translation evaluation metrics also perform well under these and related phenomena. Developing challenge sets for machine translation metric evaluation has gained considerable interest because recently neural MT evaluation metrics showed improved correlation with human judgements (Freitag et al., 2021c; Kocmi et al., 2021). However, their weaknesses remain relatively unknown and only a small number of works like Hanna and Bojar (2021) and Amrhein and Sennrich (2022) have proposed systematic analyses in uncovering them.

Previous challenge sets for metric evaluation focused on negation and sentiment polarity (Specia et al., 2020) and synthetic perturbations such as antonym replacement or punctuation (Freitag et al., 2021c). Avramidis et al. (2018) developed a manually constructed test suite of linguistically motivated perturbations for identifying weaknesses in reference-free evaluation. However, these challenge sets for metrics are only focused on high-resource language pairs such as English↔German and English→Chinese. In this work, we repurpose existing machine translation challenge sets to evaluate machine translation evaluation metrics and we introduce several synthetically generated and manually created challenge sets that broadly focus on translation accuracy errors for 146 language pairs.

4 Datasets

The majority of the examples in our challenge set were based on data extracted from three main datasets: FLORES-101, PAWS-X, and XNLI (with additional translations from XTREME).

The **FLORES-101** evaluation benchmark (Goyal et al., 2022) consists of 3,001 sentences extracted from English Wikipedia and translated into 101 languages by professional translators. **FLORES-200** (NLLB Team et al., 2022) expands the set of languages in FLORES-101. Originally intended for multilingual and low-resource MT evaluation, these datasets have a particular focus on low-resource languages.

PAWS-X (Yang et al., 2019), a cross-lingual dataset for paraphrase identification, consists of pairs of sentences that are labelled as true or adversarial paraphrases. It comprises the Wikipedia portion of the PAWS corpus (Zhang et al., 2019) translated from English into six languages: French, Spanish, German, Chinese, Japanese, and Korean. The development and test sets (23,659 sentences total) were manually translated by professional translators, and the training set was translated using NMT systems via Google Cloud Translation[2].

XNLI (Conneau et al., 2018) is a multilingual Natural Language Inference (NLI) dataset consisting of 7,500 premise-hypothesis pairs with their corresponding inference label. The English examples were generated by crowd source workers before being manually translated into 14 languages: French, Spanish, German, Greek, Bulgarian, Russian, Turkish, Arabic, Vietnamese, Thai, Chinese, Hindi, Swahili and Urdu. In addition, we use the automatic translations from **XTREME** (Hu et al., 2020) of the XNLI test set examples from these 14 languages into English.

For the *mistranslation* phenomena *Gender in*

[2] https://cloud.google.com/translate

Occupation Names and *Word Sense Disambiguation* (Sections 5.2.1 and 5.2.2) we leveraged the WinoMT and MuCoW datasets. **WinoMT** (Stanovsky et al., 2019), a challenge set developed for analysing gender bias in MT, contains 3,888 English examples extracted from the Winogender (Rudinger et al., 2017) and WinoBias (Zhao et al., 2018) coreference test sets. WinoMT sentences cast participants into non-stereotypical gender roles and the dataset has an equal balance of male and female genders, and of stereotypical and non-stereotypical gender-role assignments (e.g., a female doctor vs. a female nurse). **MuCoW** (Raganato et al., 2019) is a multilingual contrastive, word sense disambiguation test suite for machine translation. The dataset covers 16 language pairs with more than 200,000 contrastive sentence pairs. It was automatically constructed from word-aligned parallel corpora and BabelNet's (Navigli and Ponzetto, 2012) wide-coverage multilingual sense inventory.

For the *discourse-level* phenomena (Section 5.8) we relied on *annotated* resources developed specifically to support work on those phenomena in an MT setting. The **WMT 2018 English-German pronoun translation evaluation test suite** (Guillou et al., 2018) contains 200 examples of the ambiguous English pronouns *it* and *they* extracted from the TED talks portion of ParCorFull (Lapshinova-Koltunski et al., 2018). The example sentences were translated into German by the 16 English-German systems submitted to WMT 2018, and the (German) pronoun translations were manually judged by human annotators as "good/bad". **Wino-X** (Emelin and Sennrich, 2021) is a parallel dataset of German, French, and Russian Winograd schemas, aligned with their English counterparts. It was developed for commonsense reasoning and coreference resolution and used for this purpose to generate examples in Section 5.8.3. The **Europarl ConcoDisco** corpus (Laali and Kosseim, 2017) comprises the English-French parallel texts from Europarl (Koehn, 2005) over which automatic methods were used to perform PDTB-style discourse connective annotation. Discourse connectives are labelled with their sense type and are aligned between the two languages.

5 Challenge Sets

Creating a contrastive challenge set for evaluating a machine translation evaluation metric requires a source sentence, a reference translation, and two translation hypotheses of which one contains an error or phenomenon of interest (the "incorrect" translation) and the other one is a correct translation in that respect (the "good" translation). One possible way to create such challenge sets is to start with two alternative references and to insert errors in one of them to form an incorrect translation. This limits the full evaluation scope to translation hypotheses that only contain a single error. To create a more realistic setup, we also create many challenge sets where the good translation is not free of errors, but it is a better translation than the incorrect translation. For automatically created challenge sets, we put measures in place to ensure that the incorrect translation is indeed a worse translation than the good translation.

5.1 Addition and Omission

We create a challenge set for addition and omission errors which are defined in the MQM ontology as "Target content that includes content not present in the source." and "Errors where content is missing from the translation that is present in the source.", respectively. For general cases of addition and omission, we focus on the level of constituents and use an implementation by Vamvas and Sennrich (2022) to create synthetic examples of addition and omission errors.

To generate examples, we use the concatenated dev and devtest sets from the FLORES-101 evaluation benchmark. We choose the 46 languages for which there exists a stanza parser[3] and create datasets for all languages paired with English plus ten additional language pairs that we choose randomly. The script by Vamvas and Sennrich (2022) randomly drops constituents from the source sentence and then generates two translations, one of the full source and one of the partial source without the constituent. Here is an example of two resulting translations:

Full:	For example, castle visits in the Loire Valley, the Rhine Valley, or a cruise **to interesting cities on the Danube** or **a** boat ride along the Erie Canal.
Partial:	For example, castle visits in the Loire Valley, the Rhine Valley, or a cruise or boat ride along the Erie Canal.

[3] https://stanfordnlp.github.io/stanza/available_models.html

Only partial translations that can be constructed by deleting spans from the full translation are considered. For translation, we use the M2M100[4] model with 1.2B parameters (Fan et al., 2021).

We can create **omission** examples by taking the original source and reference and using the translation of the full source as a good translation and the translation of the partial source as an incorrect translation. For **addition** errors, we test if the deleted span also occurs in the reference. If not, we discard the example, if yes, we delete that span from the reference and pair this partial reference with the partial source. Then, the good translation is the translation of the partial source and the incorrect translation is the translation of the full source. For language pairs with a BLEU score less than 13 between the good translation and the reference, we manually check the examples to ensure the challenge set features appropriate examples of additions and omissions.

5.2 Mistranslation - Ambiguous Translation

This error type is defined in the MQM ontology as a case where "an unambiguous source text is translated ambiguously". For this error type, we create challenge sets where MT metrics are presented with an unambiguous source and an ambiguous reference and need to choose between two disambiguated translation hypotheses where only one meaning matches the source sentence. Therefore, these challenge sets test whether metrics consider the source when the reference is not expressive enough to choose the better translation. Since most reference-based metrics do not include the source to compute evaluation scores by design, we believe that this presents a challenging test set.

To create examples, we are inspired by Vamvas and Sennrich (2021) who score a translation against two versions of the source sentence, one with an added correct disambiguation cue and one with a wrong disambiguation cue to determine whether a translation model produced the correct translation or not. Instead of adding the disambiguation cues to the source, we use an unambiguous source and add disambiguation cues to an ambiguous reference to create two contrasting translation hypotheses.

5.2.1 Ambiguity - Occupation Names Gender

First, we create a challenge set based on WinoMT, where the challenge is to choose either a translation with a "female" or "male" disambiguation cue based on the source sentence:

SRC (de):	Die Managerin feuerte **die** Bäcker**in**.
REF (en):	The manager fired the baker.
✓:	The manager fired the **female** baker.
✗:	The manager fired the **male** baker.

We take all English sentences from the WinoMT dataset where either a pro-stereotypical or an anti-stereotypical occupation name occurs. The original sentences in WinoMT contain additional context from which the gender in the English sentence can be inferred. For example, the sentence above exists in the dataset once as "The manager fired the baker because she was too rebellious." from which it is clear that the baker is female, and once as "The manager fired the baker because he was upset." from which it is clear that the manager is male. To make the English sentences ambiguous, we remove the additional context patterns using a sequence of regular expressions, so the sentence becomes "The manager fired the baker" where the genders of the manager and the baker are ambiguous.

We then add the disambiguation cues ("female" or "male") to the ambiguous English sentences and translate them into German, French and Italian. For translation, we use Google Translate[5] because we find that this model produces gendered occupation names that are largely faithful to the disambiguation cues. Finally, we remove explicit translations of "female" and "male" from the output and predict the gender of the occupation names using the scripts provided by Stanovsky et al. (2019). We only keep translation pairs where the translation of the male-disambiguated source is predicted to be male and the translation of the female-disambiguated source is predicted to be female. We then use either the German, French or Italian translation as the source sentence, the disambiguated English sentences as the translation candidates, and the ambiguous English sentence as the reference as shown in the example.

5.2.2 Ambiguity - Word Sense Disambiguation

Second, we create a challenge set based on MuCoW, where the challenge is to choose a translation with a sense-matching disambiguation cue based on the unambiguous source sentence:

[4] https://huggingface.co/facebook/m2m100_1.2B

[5] https://translate.google.com/

SRC (de):	Was heisst "**Brühe**"?
REF (en):	What does "**stock**" mean?
✓:	What does "**vegetable stock**" mean?
✗:	What does "**penny stock**" mean?

We start with disambiguation cues that were automatically extracted by Vamvas and Sennrich (2021) via masked language modelling. Initial screening of the data shows that some disambiguation cues are not sense-specific enough. Therefore, we decide to manually check all disambiguation cues and ensure they are sense-specific and if needed, replace them with other cues. We generate three pairs of contrasting disambiguation cues per example and use the question of "What does X mean?" as a pattern to create the challenge set examples. We decided against using sentences where ambiguous words occur naturally since it may be possible to infer the correct sense from the context of the English sentence rather than by looking at the unambiguous source word. We annotate each example as to whether the correct sense is the more frequent or less frequent sense using frequency counts by Vamvas and Sennrich (2021). Following this methodology, we create a challenge set for German into English (255 examples) and Russian into English (216 examples).

5.2.3 Ambiguity - Discourse Connectives

Third, we create a challenge set where the challenge is to choose a translation with the correct discourse connective based on the unambiguous source sentence:

SRC (fr):	On estime qu'un million et demi de personnes sont mortes **depuis** la mise en uvre des sanctions.
REF (en):	It is estimated that 1.5 million people have died **since** the sanctions were introduced.
✓:	It is estimated that 1.5 million people have died **from the time** the sanctions were introduced.
✗:	It is estimated that 1.5 million people have died **because** the sanctions were introduced.

The English discourse connective "since" can have either causal or temporal meaning, which is expressed explicitly in French and German. Exploiting this fact, we use the ambiguous "since" in the reference and create two contrastive translations one with "because" for causal meaning and one with "from the time" for temporal meaning. The correct translation is determined by looking at the French or German source sentence where this information is marked explicitly. We use the discourse connective annotations in the Europarl ConcoDisco corpus for this challenge set. We use an automatic-guided search based on the French discourse connective "depuis" (which has temporal meaning) to identify candidate translation pairs. We then manually construct valid contrasting examples for causal and temporal "since" based on the English reference. This results in a challenge set for French-English with 53 examples where "since" has a causal meaning and 53 examples where it has a temporal meaning. We also create a German-English version of the challenge set, where we translate the French source sentences to German and manually correct them.

5.3 Mistranslation - Hallucinations

In this category, we group together several subcategories of mistranslation errors that happen at the word level and could occur due to hallucination of an MT model. Such errors are wrong units, wrong dates or times, wrong numbers or named entities, as well as hallucinations at the subword level that result in nonsensical words. We also present one challenge set of annotated hallucinations in real MT outputs. These challenge sets test whether the machine translation evaluation metrics can reliably identify hallucinations when presented with a correct alternative translation.

5.3.1 Hallucination - Unit Conversion

We create a challenge set for unit conversions where the challenge is to identify the correct unit conversion:

SRC (de):	Auf einem **100 Fuß** langen Teilabschnitt läuft Wasser über den Damm.
REF (en):	Water is spilling over the levee in a section **100 feet** wide.
✓:	On a **30.5 metres** long section, water flows over the dam.
✗:	On a **100 metres** long section, water flows over the dam.

We take all source sentences, reference sentences and translations of the FLORES-101 sets from Section 5.1. We only use the 45 language pairs into English since the Python packages we use for unit conversion only work for English. We first

use the Python package `quantulum3`[6] to extract unit mentions from text. We only consider sentences where we identify the same unit mentions in the translation as in the reference and we remove self-disambiguating unit mentions, like "645 miles (1040 km)" from the reference and translation. Then, we use the Python package `pint`[7] to convert unit mentions in the translation into different units. The allowed conversions can be found in Appendix A.2. The sentence with the converted amount and unit is considered to be the good translation. Based on this sentence, we construct two incorrect versions, one where the amount matches the reference but the unit is still converted (see example above) and one where the amount is the converted amount but the unit is copied from the reference. We pair each incorrect translation with the good translation and add both examples to the challenge set individually. Combining all language pairs, we construct a challenge set with 5,399 examples for each incorrect translation type.

5.3.2 Hallucination - Date-Time Errors

We also create a challenge set for the category of "date-time errors". To do this, we collect month names and their abbreviations for several language pairs. We then form a good translation by swapping a month's name with its abbreviation. The corresponding incorrect translation is generated by swapping the month name with another month name:

SRC (pt):	Os manifestantes esperam coletar uma petição de 1,2 milhão de assinaturas para apresentar ao Congresso Nacional em **novembro**.
REF (en):	Protesters hope to collect a petition of 1.2 million signatures to present to the National Congress in **November**.
✓:	The protesters expect to collect a petition of 1.2 million signatures to be submitted to the National Congress in **Nov.**
✗:	The protesters expect to collect a petition of 1.2 million signatures to be submitted to the National Congress in **August**.

To create this dataset, we use the FLORES-101 dataset from Section 5.1. We choose all pairs with target languages for which we know the abbreviations for months[8] which results in 70 language pairs. As a measure of control, we check that the identified month names in the translation also occur in the reference. If they do not, we ignore the example.

5.3.3 Hallucination - Numbers and Named Entities

We create a challenge set for numbers and named entities where the challenge is to identify translations with incorrect numbers or named entities. Following the analysis by Amrhein and Sennrich (2022), we perform character-level edits (adding, removing or substituting digits in numbers or characters in named entities) as well as word-level edits (substituting whole numbers or named entities). In last year's WMT metrics shared task, number differences were not a big issue for most neural metrics (Freitag et al., 2021c). However, we argue that simply changing a number in an alternative translation and using this as an incorrect translation is an overly simplistic setup and does not cover the whole translation hypothesis space.

SRC (es):	Sin embargo, Michael Jackson, Prince y **Madonna** fueron influencias para el álbum.
REF (en):	Michael Jackson, Prince and **Madonna** were, however, influences on the album.
Level-1 ✓:	However, Michael Jackson, Prince, and **Madonna** were influences on the album.
Level-1 ✗:	However, Michael Jackson, Prince, and **Garza** were influences on the album.
Level-2 ✓:	However, Michael Jackson, Prince, and **Madonna** were influences on the album.
Level-2 ✗:	Michael Jackson, Prince and **Garza** were, however, influences on the album.
Level-3 ✓:	The record was influenced by **Madonna**, Prince, and Michael Jackson though.
Level-3 ✗:	Michael Jackson, Prince and **Garza** were, however, influences on the album.

To address this, we propose a three-level evaluation (see examples above). The first, easiest level follows Freitag et al. (2021c) and applies a change to an alternative translation to form an incorrect translation. The second level uses an alternative translation that is lexically very similar to the refer-

[6] https://github.com/nielstron/quantulum3
[7] https://github.com/hgrecco/pint
[8] https://web.library.yale.edu/cataloging/months

ence as the good translation and applies a change to the reference to form an incorrect translation. The third, and hardest level, uses an alternative translation that is lexically very different from the reference as the good translation and applies a change to the reference to form an incorrect translation. In this way, our challenge set tests whether number and named entity differences can still be detected as the surface similarity between the two translation candidates decreases and the surface similarity of the incorrect translation to the reference increases.

We use cross-lingual paraphrases from the PAWS-X dataset as a pool of alternative translations to create this challenge set. For levels 2 and 3, we measure surface-level similarity with the Levenshtein distance[9] on character-level and use spacy[10] (Honnibal et al., 2020) for identifying named entities of type "person". To substitute whole named entities, we make use of the `names`[11] Python library. We only consider language pairs for which we can use a spacy NER model on the target side, which results in 42 language pairs.

5.3.4 Hallucination - Nonsense Words

We also consider more natural hallucinations on subword level. Because recent MT systems are trained with byte pair encoding (BPE) (Sennrich et al., 2016), the MT model may choose a wrong subword at a specific time step such that the resulting token is not a known word in the target language. With this challenge set, we are interested in how well neural MT evaluation metrics that incorporate subword-level tokenisation can identify such "nonsense" words:

SRC (de):	Die **Massen**produktion von elektronischen und digitalen Filmen war bis zum Aufkommen der pornographischen Videotechnik direkt mit der Mainstream-Filmindustrie verbunden.
REF (en):	The **mas**s production of electronic and digital films was directly linked to the mainstream film industry until the emergence of pornographic video technology.
✓:	Until the advent of pornographic video technology , the mass production of electronic and digital films was tied directly to the mainstream film industry.
✗:	The **in**s production of electronic and digital films was directly linked to the mainstream film industry until the emergence of pornographic video technology.

To create this challenge set, we consider tokens which are broken down into at least two subwords and then randomly swap those subwords with other subwords to create nonsense words. In the example above, "mass" is broken down as "mas" and "s" using BPE and the new word is created by swapping "mas" with "in" while retaining "s", creating "ins" as the nonsense word. We use the paraphrases from the PAWS-X dataset as good translations and randomly swap one subword in the reference to generate an incorrect translation. This perturbation is language-agnostic and in this work, we consider fr→ja, ko→ja, de→en, en→ko, and ko→en as the language pairs. We use the multilingual BERT (Devlin et al., 2019) tokeniser to replace the subwords.

5.3.5 Hallucination - Real Data Hallucinations

The previously discussed hallucination challenge sets were all created automatically. In addition to these challenge sets, we also create one with real data hallucinations:

[9] https://github.com/life4/textdistance
[10] https://spacy.io/
[11] https://github.com/treyhunner/names

SRC (de): Es wird angenommen, dass dieser voll gefiederte warmblütige Raubvogel aufrecht auf zwei Beinen lief und **Krallen** wie der Velociraptor hatte.

REF (en): This fully feathered, warm blooded bird of prey was believed to have walked upright on two legs with **claws** like the Velociraptor.

✓ (copy): It is believed that this fully feathered warm-blooded predator ran upright on two legs and had **claws** like the Velociraptor.

✓ (syn.): It is believed that this fully feathered warm-blooded predator ran upright on two legs and had **talons** like the Velociraptor.

✗: It is believed that this fully feathered warm-blooded predator ran upright on two legs and had **crumbs** like the Velociraptor.

For this dataset, we manually check the translations of the FLORES-101 dev and test sets for four language pairs: de→en, en→de, fr→de and en→mr. We consider both cases where a more frequent, completely wrong word occurs and cases where the MT model started with the correct subword but then produced random subwords as hallucinations. Translations with a hallucination are used as incorrect translations. We manually replace the hallucination with its correct translation to form the good translation. If possible, we create one good translation by copying the corresponding token from the reference and one with a synonymous token that does not match the reference.

5.4 Mistranslation - Lexical Overlap

Language models trained with the masked language modelling objective are successful on downstream tasks because they model higher-order word co-occurrence statistics instead of syntactic structures (Sinha et al., 2021). Although this was shown for a monolingual English model, we expect that multilingual pre-trained models, as well as MT metrics finetuned on such models, exhibit such behaviour. Similarly, existing surface-level metrics rely on n-gram matching between the hypothesis and the reference. Thus, we are interested in whether MT evaluation metrics can reliably identify the incorrect translation if it shares high lexical overlap with the reference:

SRC (fr): En 1924, il a été porte-parole invité de l'ICM à Toronto, à Oslo en 1932 et à Zurich en 1936.

REF (en): In 1924 he was an invited spokesman for the ICM in Toronto, in **Oslo in 1932** and in **1936 in Zurich.**

✓: He served as a guest speaker for ICM in 1924, 1932 and 1936 in Toronto, Oslo and Zurich.

✗: He was an invited spokesman for the ICM in Toronto in 1924, in **Zurich in 1932** and in **Oslo in 1936.**

In this example, Oslo and Zurich are swapped in the "incorrect translation" making the sentence factually incorrect. To create such examples, we use the PAWS-X dataset for which adversarial paraphrase examples were constructed by changing the word order and/or the syntactic structure while maintaining a high lexical overlap. We only consider examples in the development set that are adversarial paraphrases.

We automatically translate the first example in a pair (fr→en, en→fr, en→ja) and then manually correct the translations for en, fr, and ja to obtain 100 "good translations" per language. We use the corresponding first paraphrase as the "reference" and the second (adversarial) paraphrase as the "incorrect translation". We then pair these examples with the first paraphrase in the remaining six languages to obtain the "source". Following this methodology, we create 600 examples per target language (xx→en, xx→fr, xx→ja).

5.5 Mistranslation - Linguistic Modality

Modal auxiliary verbs signal the function of the main verb that they govern. For example, they may be used to denote possibility ("could"), permission ("may"), giving of advice ("should"), or necessity ("must"). We are interested in whether MT evaluation metrics can identify when modal auxiliary verbs are incorrectly translated:

SRC (de): Mit der Einführung dieser Regelung **könnte** diese Freiheit enden.

REF (en): With this arrangement in place, this freedom **might** end.

✓: With the introduction of this regulation, this freedom **could** end.

✗: With the introduction of this regulation, this freedom **will** end.

We focus on the English modal auxiliary verbs: *must* (necessity), and "may", "might", "could" (possibility). We begin by identifying parallel sentences where there is a modal verb in the German source sentence and one from our list (above) in the English reference. We then translate the source sentence using Google Translate to obtain the "good" translation and manually replace the modal verb with an alternative with the same meaning where necessary (e.g. "have to" denotes necessity as does "must"; also "might", "may" and "could" are considered equivalent). For the incorrect translation, we substitute the modal verb that conveys a different meaning or *epistemic strength* e.g. in the example above *might* (possibility) is replaced with *will*, which denotes (near) certainty. Instances of "may" with deontic meaning (e.g. expressing permission) are excluded from the set, leaving only those with epistemic meaning. We also construct examples in which the modal verb is omitted from the incorrect translation.

We extract 50 examples for which the modal auxiliary is substituted and 50 where it is deleted, using a combination of the FLORES-200 and PAWS-X datasets as the basis of the challenge sets.

5.6 Mistranslation - Overly Literal Translations

MQM defines this error type as translations that are overly literal translations of the source content, for example of figurative language.

5.6.1 Overly Literal - Idioms

Idioms tend to be translated overly literally (Dankers et al., 2022) and it is interesting to see if such translations are also preferred by neural machine translation evaluation metrics, which likely have not seen many idioms during finetuning:

SRC (de):	Er hat versucht, mir die Spielregeln zu erklären, aber **ich verstand nur Bahnhof**.
REF (en):	He tried to explain the rules of the game to me, but **I did not understand them**.
✓:	He tried to explain the rules of the game to me, but **it was all Greek to me**.
✗:	He tried to explain the rules of the game to me, but **I only understood train station**.

We create this challenge set based on the PIE[12] parallel corpus of English idiomatic expressions and literal paraphrases (Zhou et al., 2021). We manually translate 102 parallel sentences into German for which we find a matching idiom that is not a word-by-word translation of the original English idiom. Further, we create an overly-literal translation of the English and German idioms. We use either the German or English original idiom as the source sentence. Then, we either use the correct idiom in the other language as the reference and the literal paraphrase as the good translation or vice versa. The incorrect translation is always the overly-literal translation of the source idiom.

[12] https://github.com/zhjjn/MWE_PIE

5.6.2 Overly-Literal - Real Data Errors

We are also interested in overly-literal translations occurring in real data:

SRC (de):	Today, the only insects that cannot fold back their wings are **dragon flies** and mayflies.
REF (en):	Heute sind **Libellen** und Eintagsfliegen die einzigen Insekten, die ihre Flügel nicht zurückklappen können.
✓ (copy) :	Heute sind die einzigen Insekten, die ihre Flügel nicht zurückbrechen können, **Libellen** und Mayflies.
✓ (syn.):	Heute sind die einzigen Insekten, die ihre Flügel nicht zurückbrechen können, **Wasserjungfern** und Mayflies.
✗:	Heute sind die einzigen Insekten, die ihre Flügel nicht zurückbrechen können, **Drachenfliegen** und Mayflies.

For this challenge set, we manually check MT translations of the FLORES-101 datasets. If we find an overly-literal translation, we manually correct it to form the good translation. We create one good translation where we copy the part of the reference that corresponds to the overly-literal part and if possible another good translation where we use a synonym of the reference token. This challenge set contains examples for four language pairs: de→en, en→de, fr→de and en→mr.

5.6.3 Mistranslation - Sentence-Level Meaning Error

We also consider a special case of sentence-level semantic error that arises due to the nature of the task of natural language inference (NLI). The task of NLI requires identifying where the given hypothesis is an entailment, contradiction of, or neutral, with respect to a given premise. As a result, the premise and hypothesis have substantial overlap

but they vary in meaning. We are interested in whether MT evaluation metrics can pick up on such sentence-level meaning changes:

SRC (el):	Ο πραγματικός θόρυβος ελκύει τους ηλικιωμένους.
REF (en):	Real noise appeals to the old. (premise)
✓:	The real noise attracts the elderly.
✗:	Real noise appeals to the young and appalls the old. (hypothesis)

We use the XNLI dataset for creating such examples and consider examples where there is at least 0.5 chrF score between the English premise and hypothesis and where the labels are either contradiction or neutral. Examples with the label of entailment are excluded because some examples in the dataset are paraphrases of each other and there would be no sentence-level meaning change. We discuss effects of entailment in detail in section 5.12.1

We use either the premise or the hypothesis as the reference and an automatic translations as the "good translations" in our challenge set. Then, we use the corresponding premise or hypothesis from the remaining 14 languages as the source. The "incorrect translation" is either the premise if the reference is the hypothesis, or vice versa.

5.7 Mistranslation - Ordering Mismatch

We also investigate the effects of changing the order of words in a way that changes meaning. This challenge set is created manually by changing translations from the FLORES-101 dataset and covers de→en, en→de and fr→de.

SRC (de):	Erfülle Dein Zuhause mit einem köstlichem **Kaffee** am Morgen und etwas entspannendem **Kamillentee** am Abend.
REF (en):	Fill your home with a rich **coffee** in the morning and some relaxing **chamomile tea** at night.
✓:	Fill your home with a delicious **coffee** in the morning and some relaxing **chamomile tea** in the evening.
✗:	Fill your home with a delicious **chamomile tea** in the morning and some relaxing **coffee** in the evening.

5.8 Mistranslation - Discourse-level Errors

We introduce a new subclass of mistranslation errors that specifically cover discourse-level phenomena.

5.8.1 Discourse-level Errors - Pronouns

First, we are interested in how MT evaluation metrics handle various discourse-level phenomena related to pronouns. To create these challenge sets, we use the English-German pronoun translation evaluation test suite from the WMT 2018 shared task as the basis for our examples.

We extract all translations (by the English-German WMT 2018 systems) that were marked as "correct" by the human annotators, for the following six categories derived from the manually annotated pronoun function and attribute labels: pleonastic *it*, anaphoric subject and non-subject position *it*, anaphoric *they*, singular *they*, and group *it/they*. In the case of anaphoric pronouns, we select only the inter-sentential examples (i.e. where the sentence contains both the pronoun and its antecedent). We use the MT translations as the "good" translations and automatically generate "incorrect" translations using one of the following strategies: *omission* - the translated pronoun is deleted from the MT output, *substitution* - the "correct" pronoun is replaced with an "incorrect" form.

For *anaphoric* pronouns, when translated from English into a language with grammatical gender, such as German, the pronoun translation must a) agree in number and gender with the translation of its antecedent, and b) have the correct grammatical case. We propose "incorrect" translations as those for which this agreement does not hold:

SRC (en):	I have a *shopping bag*; ***it*** is red.
REF (de):	Ich habe eine *Einkaufstüte*; ***sie*** ist rot.
✓:	Ich habe einen *Einkaufsbeutel*; ***er*** ist rot.
✗ (subs.):	Ich habe einen *Einkaufsbeutel*; ***sie*** ist rot.
✗ (omit):	Ich habe einen *Einkaufsbeutel*; ***Ø*** ist rot.

Conversely, for *pleonastic* uses of "it" no agreement is required, instead, the correct translation in German requires a simple mapping: "it" → "es". An 'incorrect" translation of pleonastic 'it' in German could be "er" (masc. sg.) or "sie" (fem. sg., or pl.). We create, for each "correct" translation a set of possible "incorrect" values and automati-

cally select one at random to replace the "correct" pronoun. For example, in the pleonastic case:

SRC (en):	**It** is raining
REF (de):	**Es** regnet
✓:	**Es** regnet
✗ (subs.):	**Er** regnet
✗ (omit):	**Ø** regnet

5.8.2 Discourse-level Errors - Discourse Connectives

The English discourse connective "while" is ambiguous – it may be used with either a *Comparison.Contrast* or *Temporal.Synchrony* sense – as are two of its possible translations into French: "tandis que" and "alors que". We leverage a corpus of parallel English/French sentences with discourse connectives marked and annotated for sense, and select examples with ambiguity in the French source sentence. We construct the good translation by replacing instances of "while" Temporal with "as" or "as long as" and "while" Comparison as "whereas" (ensuring grammaticality is preserved). For the incorrect translation, we replace the discourse connective with one with the alternative sense of "while" e.g. "whereas" (Comparison) where a Temporal sense is required:

SRC (fr):	Dans l'UE-10, elles ont progressé de 8% **tandis que** la dette pour l'UE-2 a augmenté de 152%.
REF (en):	In EU-10 they grew by 8% **while** the debt for the EU-2 increased by 152%.
✓:	In the EU-10, they increased by 8% **when** the debt for the EU-2 increased by 152%.
✗:	In the EU-10, they increased by 8% **whereas** the debt for the EU-2 increased by 152%.

We extract our examples from the Europarl ConcoDisco dataset. We automatically selected the sentence pairs that contain an instance of "while" in English and either "alors que" or "tandis que" in French. Our dataset contains 50 examples for the *Comparison.Contrast* sense and 21 for the *Temporal.Synchrony* sense.

This challenge set complements the discourse connectives set in section 5.2.3, in which the English discourse connective "since" is ambiguous, but the corresponding connectives in French and German are not. In future work, we might aim to expand the set of ambiguous discourse connectives for English-French, and/or expand to other language pairs.

5.8.3 Discourse-level Errors - Commonsense Co-Reference Disambiguation

One of the greater challenges within computational coreference resolution is referring to the correct antecedent by using commonsense/real-world knowledge. Emelin and Sennrich (2021) construct a benchmark to test whether multilingual language models and neural machine translation models can perform such commonsense coreference resolution. We are interested in whether such commonsense coreference resolution poses a challenge for MT evaluation metrics:

SRC (en):	The woman looked for a different *vase* for the bouquet because **it** was too small.
REF (de):	Die Frau suchte nach einer anderen Vase für den Blumenstrauß, weil **sie** zu klein war.
✓:	Die Frau suchte nach einer anderen Vase für den Blumenstrauß, weil **die Vase** zu klein war.
✗:	Die Frau suchte nach einer anderen Vase für den Blumenstrauß, weil **der Blumenstrauß** zu klein war.

The English sentences for this challenge set are sampled from the Winograd schema. All contain the *it* pronoun and are then manually translated into two contrastive translations for de, fr, and ru. Based on this data, we create our challenge sets covering two types of examples: For the first, the good translation contains the pronoun referring to the correct antecedent while the incorrect translation contains the pronoun referring to the incorrect antecedent. For the second, the correct translation translates the *it* into the correct disambiguating filler while the second translation contains the adversarial filler (see example above).

The sentences for en→de were common across both the challenge sets developed by Emelin and Sennrich (2021). Hence, the corresponding correct translations from the two challenge sets were used as the "good" translation for our evaluation setup. For en→ru and en→fr, the source containing the ambiguous pronoun was machine translated and then verified by human annotators to form the

"good" translation.

5.9 Untranslated

MQM defines this error type as "errors occurring when a text segment that was intended for translation is left untranslated in the target content". In ACES, we consider both word-level and sentence-level untranslated content.

5.9.1 Untranslated - Word-Level

For word-level untranslated content, we manually annotate translations of the FLORES-101 test and dev sets:

SRC (fr):	À l'origine, l'émission mettait en scène des **comédiens de doublage** amateurs, originaires de l'est du Texas.
REF (de):	Die Sendung hatte ursprünglich lokale Amateur**synchronsprecher** aus Ost-Texas.
✓ (copy):	Ursprünglich spielte die Show mit Amateur**synchronsprechern** aus dem Osten von Texas.
✓ (syn.):	Ursprünglich spielte die Show mit Amateur-**Synchron-Schauspielern** aus dem Osten von Texas.
✗:	Ursprünglich spielte die Show mit Amateur-**Doubling-Schauspielern** aus dem Osten von Texas.

We do not only count complete copies as untranslated content but also content that clearly comes from the source language but was only adapted to look more like the target language (as in the example above). If we encounter an untranslated span we use this translation as the incorrect translation and create a good translation by copying the correct span from the reference and – if possible – a second good translation where we use a synonym for the correct reference span. We manually annotate such untranslated errors for en→de, fr→de, de→en, en→mr.

5.9.2 Untranslated - Full Sentences

In case of underperforming machine translation models, sometimes the generated output contains a majority of the tokens from the source language to the extent of copying the entire source sentence.[13] We create a challenge set by simply copying the entire source sentence as the incorrect translation. We used a combination of examples from the FLORES-200, XNLI, and PAWS-X datasets to create these examples. We expect that this challenge set is likely to break embedding-based reference-free evaluation because the representation of the source and the hypothesis will be the same, thus leading to a higher score.

[13] Through observations of Swahili English translation; unpublished work

5.10 Do Not Translate Errors

This category of errors is defined in MQM as content in the source that should be copied to the output in the source language but was mistakenly translated to the target language. Common examples of this error type are company names or slogans. Here, we manually create a challenge set based on the PAWS-X data which contains many song titles that should not be translated:

SRC (en):	Dance was one of the inspirations for the exodus - song **"The Toxic Waltz"**, from their 1989 album "Fabulous Disaster".
REF (de):	Dance war eine der Inspirationen für das Exodus-Lied **„The Toxic Waltz"** von ihrem 1989er Album „Fabulous Disaster".
✓:	Der Tanz war eine der Inspirationen für den Exodus-Song **„The Toxic Waltz"**, von ihrem 1989er Album „Fabulous Disaster".
✗:	Der Tanz war eine der Inspirationen für den Exodus-Song **„Der Toxische Walzer"**, von ihrem 1989er Album „Fabulous Disaster".

To construct the challenge set, we use one paraphrase as the good translation and manually translate an English sequence of tokens (e.g. a song title) into German to form the incorrect translation.

5.11 Overtranslation and Undertranslation

Hallucinations from a translation model can often produce a term which is either more generic than the source word or more specific. Within the MQM ontology, the former is referred to undertranslation while the latter is referred to as overtranslation. For example, "car" is replaced by "vehicle" (undertranslation) or "BMW" (overtranslation). To automate the generation of such errors, we consider using Wordnet (Miller, 1994) where a randomly selected noun from the reference translation is replaced by its corresponding hypernym or hyponym to simulate undertranslation or overtranslation errors, respectively:

SRC (de):	Bob und Ted waren Brüder. Ted ist der **Sohn** von John.
REF (en):	Bob and Ted were brothers. Ted is John's **son**.
✓:	Bob and Ted were brothers, and Ted is John's **son**.
✗:	Bob and Ted were brothers. Ted is John 's **male offspring**.

During the implementation, we only replaced the first sense listed in the Wordnet for the corresponding noun, which may not be appropriate in the given translation. We constructed this challenge set for hypernyms and hyponyms using the PAWS-X dataset, only considering the language pairs where the target language is English (es→en, fr→en, de→en, ja→en, ko→en, zh→en). Though this work only discusses automatic construction for English, we are releasing the code which supports the Multilingual WordNet. The caveat is that the noun in the target language first looks at its English translation, looks up the corresponding hypernym/hyponym and then finds the corresponding translation in the target language.

5.12 Real-world Knowledge

We manually constructed 20 examples each for en→de and de→en for the first four phenomena described in this section. We used German-English examples from XNLI, plus English translations from XTREME as the basis for our examples. Typically, we select a single sentence, either the premise or hypothesis from XNLI, and manipulate the MT translations.

5.12.1 Real-world Knowledge - Textual Entailment

We test whether the metrics can recognise textual entailment – that is, whether a metric can recognise that the meaning of the source/reference is entailed (i.e. can be inferred) by the "good" translation:

SRC (de):	Ein Mann **wurde ermordet**.
REF (en):	A man **was murdered**.
✓:	A man **died**.
✗ (omit):	A man **was attacked**.

We construct examples for which the good translation entails the meaning of the original sentence (and its reference). For example, we use the entailment *was murdered* → *died* (i.e. if a person is murdered then they must have died) to construct the good translation in the example above. We construct the incorrect translation by replacing the entailed predicate (*died*) with a related but non-entailed predicate (here *was attacked*) – a person may have been murdered without being attacked, i.e. by being poisoned for example. In cases where an antonymous predicate is available, we use that predicate in the incorrect translation. For example, if "lost" is in the source/reference, we use "won" in the incorrect translation (lost $\not\rightarrow$ won).

5.12.2 Real-world Knowledge - Hypernyms and Hyponyms

We consider a translation that contains a hypernym of a word to be better than one that contains a hyponym. For example, whilst translating "Hund" ("dog") with the broader term "animal" results in some loss of information, this is preferable over hallucinating information by using a more specific term such as "labrador" (i.e. an instance of the hyponym class "dog"). We consider a translation that contains a *hypernym* of a word to be better than one that contains a *hyponym*:

SRC (de):	..., dass der **Hund** meiner Schwester gehört.
REF (en):	... the **dog** belonged to my sister.
✓ (hypernym):	... the **pet** belonged to my sister.
✗ (hyponym):	... the **labrador** belonged to my sister.

We used Wordnet and WordRel.com[14] (an online dictionary of words' relations) to identify hypernyms and hyponyms of nouns within the reference sentences, and used these as substitutions in the MT output: hypernyms are used in the "good" translations and hyponyms in the "incorrect" translations.

5.12.3 Real-world Knowledge - Hypernyms and Distractors

Similar to the hypernym vs. hyponym examples, we include 20 examples in which the good translation contains a hypernym (here "pet"), and the incorrect translation contains a different member from the hypernym class (e.g. "cat") to that in the source/reference. For example:

[14] https://wordrel.com/

SRC (de):	..., dass der **Hund** meiner Schwester gehört.
REF (en):	... the **dog** belonged to my sister.
✓ (hypernym):	... the **pet** belonged to my sister.
✗ (hyponym):	... the **cat** belonged to my sister.

As before, we used Wordnet and WordRel.com to identify hypernyms of nouns in the reference translation.

5.12.4 Real-world Knowledge - Antonyms

Similar to the generation of over- and undertranslations in Section 5.11, we also automatically constructed "incorrect" translations by replacing nouns with their corresponding antonyms from Wordnet:

SRC (de):	Ich **hasste** jedes Stück der Schule!
REF (en):	I **hated** every bit of school!
✓ (synonym):	I **loathed** every bit of school!
✗ (antonym):	I **loved** every bit of school!

As this method could result in noisy replacement of the words with their respective antonyms, we also construct a manual and more challenging setup for the metrics. We assess whether the metrics can distinguish between translations that contain a synonym versus an antonym of a given word. We consider a translation that contains a synonym of a word in the reference to be a "good" translation, and one that contains an antonym of that word to be "incorrect". As in the example above the use of synonyms preserves the meaning of the original sentence, and the antonyms introduce a polar opposite meaning.

5.12.5 Real-world Knowledge - Commonsense

We are also interested in whether evaluation metrics prefer translations that adhere to common sense. To test this, we remove explanatory subordinate clauses from the sources and references in the dataset described in Section 5.8.3. This guarantees that when choosing between the good and incorrect translation, the metric cannot infer the correct answer from looking at the source or the reference:

SRC (en):	Die Frau suchte nach einer anderen Vase für den Blumenstrauß.
REF (de):	The woman looked for a different vase for the bouquet.
✓:	The woman looked for a different vase for the bouquet because **the vase** was too small.
✗:	The woman looked for a different vase for the bouquet because **the bouquet** was too small.

We remove the explanatory subordinate clauses using a sequence of regular expressions. We then pair the shortened source and reference sentences with the full translation that follows commonsense as the good translation and the full translation with the other noun as the incorrect translation.

Since we present several challenge sets in Section 5.2 where the good translation can only be identified by looking at the source sentence, we also create a version of this challenge set where the explanatory subordinate clause is only removed from the reference but not from the source. By comparing this setup with the results from the setup described above, we achieve another way of quantifying how much a metric considers the source.

5.13 Wrong Language

Most of the representations obtained from large multilingual language models do not explicitly use the language id as an input while encoding a sentence. Here, we are interested in checking whether sentences which have similar meanings are closer together in the representation space of neural MT evaluation metrics, irrespective of their language. We create a challenge set for embedding-based metrics where the incorrect translation is a translation from a similar language (same typology/same script) to the reference of the translation. Note that this is also a common error with multilingual machine translation models. We constructed these examples using the FLORES-200 dataset where the "good" translation was the automatic translation and the "incorrect" translation was the reference from a language similar to the target language:

SRC (en):	Cell comes from the Latin word cella which means small room.
REF (es):	El término célula deriva de la palabra latina cella, que quiere decir «cuarto pequeño».
✓ (es):	La célula viene de la palabra latina cella que significa habitación pequeña.
✗ (ca):	Cèl·lula ve de la paraula llatina cella, que vol dir habitació petita.

We construct two categories within this challenge set: one where the target language is a higher-resource language and the incorrect language is a lower-resource language and vice-versa. The languages we consider are (`src-tgt-sim`): en-hi-mr, en-es-ca, en-cs-pl, fr-mr-hi, en-pl-cs, and en-ca-es.

5.14 Fluency

Although the focus of ACES is on accuracy errors, we also include a small set of fluency errors for the punctuation category. Future work might consider expanding this set to include other categories of fluency errors.

5.15 Punctuation

We assess the effect of deleting and substituting punctuation characters. We employ four strategies: 1) deleting all punctuation [1,000 examples], 2) deleting only quotation marks (i.e. removing indications of quoted speech) [150 ex.], 3) deleting only commas (i.e. removing clause boundary markers) [508 ex.], 4) replacing exclamation points with question marks (i.e. statement → question) [15 ex.].

In strategies 3 and 4, some of the examples may also contain accuracy-related errors. For example, examples in which the meaning of the sentence is changed in the incorrect translation. We use the TED Talks from the WMT 2018 English-German pronoun translation evaluation test suite and apply all deletions and substitutions automatically.

6 Evaluation Methodology

We shall now briefly describe the metrics that participated in the challenge set shared task. The organisers of the shared task also provided scores of a few baseline metrics as described below.

6.1 Baseline Metrics

BLEU (Papineni et al., 2002) compares the token-level n-grams of the hypothesis with the reference translation and then computes a precision score weighted by a brevity penalty.

chrF (Popović, 2017) evaluates translation outputs based on a character n-gram F-score by computing overlaps between the hypothesis and the reference.

BERTScore (Zhang et al., 2020) uses contextual embeddings from pre-trained language models to compute the similarity between the tokens in the reference and the generated translation using cosine similarity. The similarity matrix is used to compute precision, recall, and F1-scores.

BLEURT-20 (Sellam et al., 2020) is a BERT-based (Devlin et al., 2019) regression model, which is first trained on scores of automatic metrics/similarity of pairs of reference sentences and their corrupted counterparts. It is then fine-tuned on the WMT human evaluation data to produce a score for a hypothesis given a reference translation.

COMET-20 (Rei et al., 2020) uses a cross-lingual encoder (XLM-R (Conneau et al., 2020)) and pooling operations to obtain sentence-level representations of the source, hypothesis, and reference. These sentence embeddings are combined and then passed through a feedforward network to produce a score. COMET is trained on human evaluation scores of machine translation systems submitted to WMT until 2020.

COMET-QE was trained similarly to COMET-20 but only the source and the hypothesis are combined to produce a final score as this is a reference-free metric.

YiSi-1 (Lo, 2019) measures the semantic similarity between the hypothesis and the reference by using cosine similarity scores of multilingual representations at the lexical level. It optionally uses a semantic role labeller to obtain structural similarity. Finally, a weighted f-score based on structural and lexical similarity is used for scoring the hypothesis against the reference.

6.2 Metrics Submitted to WMT 2022

We list the descriptions provided by the authors of the respective metrics and refer the reader to their system description papers for further details.

6.2.1 MATESE and MATESE-QE

MATESE metrics (Perrella et al., 2022) leverage transformer-based multilingual encoders to identify error spans in translations, and classify their severity between MINOR and MAJOR. The quality score returned for a translation is computed following the MQM error weighting introduced in "Experts Errors and Context: A Large-Scale Study of Human Evaluation for Machine Translation" (Freitag et al., 2021b). MATESE is reference-based, while MATESE-QE is its reference-free version, with the source sentence used in place of the reference.

6.2.2 UniTe

UniTE (Wan et al., 2022), Unified Translation Evaluation, is a metric approach where the model-based metrics can possess the ability of evaluating translation outputs following all three evaluation scenarios, i.e. source-only, reference-only, and source-reference-combined.

6.2.3 COMET-22

COMET-22 is an ensemble between a vanilla COMET model trained with Direct Assessment (DA) scores and a Multitask model that is trained on regression (MQM regression) and sequence tagging (OK/BAD word identification from MQM span annotations). These models are ensembled together using an hyperparameter search that weights different features extracted from these two evaluation models and combines them into a single score. The vanilla COMET model is trained with DA's ranging 2017 to 2020 while the Multitask model is trained using DA's ranging from 2017 to 2020 plus MQM annotations from 2020 (except for en-ru that uses TedTalk annotations from 2021).

6.2.4 COMET-Kiwi

COMET-Kiwi ensembles two QE models similarly to COMET-22. The first model follows the classic Predictor-Estimator QE architecture where MT and source are encoded together. This model is trained on DAs ranging 2017 to 2019 and then fine-tuned on DAs from MLQE-PE (the official DA from the QE shared task). The second model is the same multitask model used in the COMET-22 submission but without access to a reference translation. This means that this model is a multitask model trained on regression and sequence tagging.

Both models are ensembled together using an hyperparameter search that weights different features extracted from these two QE models and combines them into a single score.

6.2.5 MS-COMET

Both methods, MS-COMET-22 and MS-COMET-QE-22, are basically COMETs (Rei et al., 2020) trained on mainly internal Microsoft data.

6.2.6 HUAWEI Metrics

Huawei submitted the following metrics to the shared task (Qiao et al., 2022):

HWTSC_EE_BERTScore* (Entropy Enhanced Metrics) are a group of metrics built upon existing metrics. They aim to achieve a more balanced system-level rating by assigning weights to segment-levels scores produced by backbone metrics. The weights are determined by the difficulty of a segment, which is related to the entropy of a hypothesis-reference pair. A translation hypothesis with a significantly high entropy value is considered difficult and receives a large weight in aggregation of EE-Metrics' system-level scores.

KG-BERTScore is a reference-free machine translation (MT) evaluation metric, which incorporates a multilingual knowledge graph into BERTScore by linearly combining the results of BERTScore and bilingual named entity matching.

CROSS-QE is a submission based on the COMET-QE architecture.

HWTSC-Teacher-Sim is a reference-free metric constructed by fine-tuning the multilingual Sentence BERT model: paraphrase-multilingual-mpnet-base-v2 (Reimers and Gurevych, 2019).

HWTSC-TLM is a reference-free metric which only uses a target-side language model and only uses the system translations as input.

We are awaiting the descriptions of DATScore, MEE, Metric-X and REUSE.

6.3 Evaluation of Metrics

For all phenomena in ACES where we generated more than 1,000 examples, we randomly subsample 1,000 examples according to the per language pair distribution to keep the evaluation of new metrics tractable.

We follow the evaluation of the challenge sets from the 2021 edition of the WMT metrics shared task (Freitag et al., 2021c) and report performance with Kendall's tau-like correlation. This metric measures the number of times a metric scores the good translation above the incorrect translation (concordant) and vice versa (discordant):

$$\tau = \frac{concordant - discordant}{concordant + discordant}$$

Note that a higher τ indicates a better performance.

7 Results

7.1 Phenomena-level Results

We start by providing a broad overview of metric performance on the different categories of phenomena. We compute Kendall's tau-like correlation scores (Section 6) for the 22 metrics which a) provide segment-level scores and b) provide scores for all language pairs and directions in ACES. We first compute the correlation scores for all of the individual phenomena and then take the average score over all phenomena in each of the nine top-level accuracy categories in ACES plus the fluency category *punctuation* (see Table 1).

The performance of the metrics varies greatly and there is no clear *winner* in terms of performance across all of the categories. There is also a high degree of variation in terms of metric performance when each category is considered in isolation. Whilst each of the categories proves challenging for at least one metric, some categories are more challenging than others. For example, looking at the average scores in Table 1, and without taking outliers into account, we might conclude that *undertranslation*, *wrong language*, and *undertranslated* (all with average Kendall tau-like correlation of < 0.3) present more of a challenge than the other categories. On the other hand, for *omission* (with an average Kendall tau-like correlation of 0.748) metric performance is generally rather high.

We also observe variation in terms of the performance of metrics belonging to the baseline, reference-based, and reference-free groups. For example, the baseline metrics appear to struggle more on the *overtranslation* and *undertranslation* categories than the metrics belonging to the other groups. Reference-based metrics also appear to perform better overall on the *untranslated* category than the reference-free metrics. This makes sense as a comparison with the reference is likely to highlight tokens that ought to have been translated.

Next, we drill down to the fine-grained categories of the largest category: *mistranslation*. We present metric performance on its sub-level categories in Table 2. Again, we find that performance on the different sub-categories is variable, with no clear *winner* among the metrics. The results suggest that *hallucination* phenomena are generally more challenging than *discourse-level* phenomena. Performance on the *hallucination* sub-category is poor overall, although it appears to be particularly challenging for the baseline metrics.

7.2 ACES Score

To simplify the possible future creation of a leaderboard based on ACES, we define a weighted combination of the top-level categories into a single score which we term the "ACES - Score":

$$ACES = sum \begin{cases} 5 * \tau_{addition} \\ 5 * \tau_{omission} \\ 5 * \tau_{mistranslation} \\ 1 * \tau_{untranslated} \\ 1 * \tau_{donottranslate} \\ 5 * \tau_{overtranslation} \\ 5 * \tau_{undertranslation} \\ 1 * \tau_{real-worldknowledge} \\ 1 * \tau_{wronglanguage} \\ 0.1 * \tau_{punctuation} \end{cases}$$

The weights correspond to the values under the MQM framework that Freitag et al. (2021b) recommend for major (weight=5), minor (weight=1) and fluency/punctuation errors (weight=0.1). We decide that untranslated, do not translate and wrong language errors should be counted as minor errors because they can be identified automatically with language detection tools and should also be easy to spot in post-editing. For real-world knowledge, we also count this category as minor errors since we do not expect that current MT evaluation metrics have any notion of real-world knowledge and we do not want to punish them too severely if they do not perform well on this challenge set. The ACES-Score ranges from -29.1 (all phenomena have a correlation of -1) to 29.1 (all phenomena have a correlation of +1).

	addition	omission	mistranslation	untranslated	do not translate	overtranslation	undertranslation	real-world knowledge	wrong language	punctuation	ACES score
Examples	*999*	*999*	*20479*	*1300*	*100*	*1000*	*1000*	*2948*	*2000*	*1673*	
BERTScore	0.880	0.752	0.284	0.774	**0.960**	-0.111	-0.189	0.032	0.565	0.849	10.49
BLEU	0.748	0.438	0.045	0.680	0.818	-0.835	-0.851	-0.203	0.665	0.709	- 0.24
BLEURT20	0.437	0.810	0.397	0.754	0.860	0.200	0.014	0.401	0.533	0.649	11.90
chrF	0.642	0.784	0.145	**0.787**	**0.960**	-0.696	-0.592	-0.236	0.691	0.810	3.70
COMET-QE	-0.538	0.401	0.415	0.135	0.140	0.622	0.446	0.321	-0.504	0.253	6.85
COMET-20	0.439	0.808	0.336	0.750	0.900	0.312	0.110	0.267	0.034	0.706	12.05
YiSi-1	0.770	0.866	0.327	0.736	0.920	-0.062	-0.076	0.110	0.431	0.734	11.39
COMET-22	0.333	0.806	0.546	0.536	0.900	0.690	0.538	0.574	-0.318	0.539	16.31
DATScore.1.2B.ref.not_weighted	0.974	0.882	0.475	0.404	0.840	0.332	0.120	0.423	0.878	0.898	16.55
DATScore.1.2B.ref.weighted	0.954	0.930	0.462	0.381	0.840	0.292	0.122	0.397	0.844	0.937	16.35
metricx_xl_DA_2019	0.403	0.852	0.522	0.731	0.940	0.692	0.384	**0.739**	0.521	0.670	17.26
metricx_xl_MQM_2020	-0.275	0.670	0.521	0.584	0.740	0.718	**0.602**	0.705	-0.125	0.446	13.13
MS-COMET-22	-0.245	0.624	0.380	0.505	0.640	0.554	0.376	0.249	0.084	0.490	9.97
UniTE	0.423	0.870	0.463	0.443	0.900	0.482	0.268	0.549	-0.361	0.714	14.13
COMET-Kiwi	0.361	0.830	**0.601**	0.230	0.780	**0.738**	0.572	0.581	-0.358	0.491	16.79
CROSS-QE	0.163	0.876	0.507	-0.094	0.320	0.726	0.506	0.446	-0.374	0.455	14.24
DATScore.418M.no_ref.not_weighted	**0.994**	0.936	0.443	-0.554	0.840	0.578	0.318	0.409	**0.908**	0.930	**18.04**
DATScore.418M.no_ref.weighted	0.988	**0.966**	0.411	-0.580	0.800	0.524	0.310	0.393	0.871	**0.954**	17.57
HWTSC-Teacher-Sim	-0.027	0.495	0.374	-0.310	0.720	0.546	0.458	0.238	-0.024	0.297	9.89
HWTSC-TLM	-0.363	0.345	0.420	0.157	-0.040	0.544	0.474	0.071	-0.168	0.634	7.18
KG-BERTScore	0.794	0.812	0.441	-0.493	0.780	0.652	0.528	0.464	0.306	0.283	17.22
MS-COMET-QE-22	-0.231	0.696	0.396	-0.030	0.140	0.640	0.448	0.300	-0.102	0.558	10.11
Average	0.392	0.748	0.405	0.297	0.714	0.370	0.222	0.329	0.227	0.636	12.31

Table 1: Average Kendall's tau-like correlation results for the nine top level categories in the ACES ontology, plus the additional fluency category: punctuation. The horizontal lines delimit baseline metrics (top), participating reference-based metrics (middle) and participating reference-free metrics (bottom). The best result for each category is denoted by bold text with a green highlight. Note that *Average* is an average over averages. Last column shows a weighted sum of the correlations.

	disco.	halluci.	other
Examples	*3698*	*10292*	*7627*
BERTScore	0.568	-0.053	0.361
BLEU	0.147	-0.190	0.242
BLEURT20	0.695	0.144	0.402
chrF	0.410	-0.129	0.175
COMET-QE	0.663	0.302	0.206
COMET-20	0.643	0.013	0.399
YiSi-1	0.610	0.025	0.368
COMET-22	0.682	0.460	0.542
DATScore.1.2B.ref.not_weighted	0.784	0.237	0.477
DATScore.1.2B.ref.weighted	0.806	0.212	0.447
metricx_xl_DA_2019	0.704	0.495	0.456
metricx_xl_MQM_2020	0.579	**0.679**	0.394
MS-COMET-22	0.728	0.145	0.283
UniTE	0.733	0.302	0.429
COMET-Kiwi	0.734	0.493	**0.638**
CROSS-QE	0.644	0.395	0.563
DATScore.418M.no_ref.not_weighted	0.806	0.216	0.417
DATScore.418M.no_ref.weighted	**0.811**	0.185	0.363
HWTSC-Teacher-Sim	0.561	0.297	0.339
HWTSC-TLM	0.756	0.306	0.151
KG-BERTScore	0.560	0.390	0.481
MS-COMET-QE-22	0.649	0.246	0.394
Average	0.647	0.235	0.388

Table 2: Average Kendall's tau-like correlation results for the sub-level categories in mistranslation: **disco**urse-level, **halluci**nation, and **other** errors. The horizontal lines delimit baseline metrics (top), participating reference-based metrics (middle) and participating reference-free metrics (bottom). The best result for each category is denoted by bold text with a green highlight. Note that *Average* is an average over averages.

The results can be seen in Table 1 in the last column. Almost all metrics of this year's submissions improved quite significantly over the baselines. Interestingly, many reference-free metrics also perform on par with reference-based metrics. The best performing metric is also a reference-free metric, namely `DATScore.418M.no_ref.not_weighted`. However, we caution against making strong claims about which metrics perform *best* or *worst* on the challenge set based on this high-level overview. Instead, we recommend that ACES be used to highlight general trends as to what the outstanding issues are for MT evaluation metrics. More fine-grained analyses are reported in the following sections.

More generally, work on analysing system performance on *ACES* prompts the question: what is the definition of a good metric? One might consider that a *good* metric exhibits a strong correlation with human judgements on whether a translation is good/bad *and* assigns sufficiently different scores to a good vs. an incorrect translation. The latter criterion would provide evidence of the ability of the metric to discriminate reliably between good and incorrect translations, but it may be difficult to establish what this difference should be, especially without knowing to what degree the translations are good/bad without human judgements.

7.3 Language-level Results

	trained	en-x	x-en	x-y
Examples	*8871*	*12701*	*17973*	*5824*
BERTScore	0.480	0.380	0.173	0.127
BLEU	0.295	0.293	-0.067	0.194
BLEURT20	0.541	0.442	0.281	0.258
chrF	0.361	0.358	-0.036	0.126
COMET-QE	0.356	0.224	0.144	0.168
COMET-20	0.496	0.329	0.277	0.120
YiSi-1	0.476	0.393	0.185	0.153
COMET-22	0.599	0.431	0.555	0.354
DATScore.1.2B.ref.not_weighted	0.621	0.633	0.426	0.554
DATScore.1.2B.ref.weighted	0.638	0.640	0.399	0.540
metricx_xl_DA_2019	0.623	0.573	0.457	0.553
metricx_xl_MQM_2020	0.608	0.511	0.453	0.510
MS-COMET-22	0.470	0.347	0.248	0.149
UniTE	0.598	0.420	0.377	0.224
COMET-Kiwi	0.620	0.488	**0.694**	0.467
CROSS-QE	0.598	0.444	0.552	0.291
DATScore.418M.no_ref.not_weighted	0.630	0.669	0.506	**0.589**
DATScore.418M.no_ref.weighted	**0.653**	**0.682**	0.420	0.567
HWTSC-Teacher-Sim	0.477	0.389	0.351	0.148
HWTSC-TLM	0.538	0.428	0.167	0.194
KG-BERTScore	0.467	0.509	0.507	0.349
MS-COMET-QE-22	0.490	0.383	0.402	0.111
MATESE	0.407	n/a	n/a	n/a
MATESE-QE	0.398	n/a	n/a	n/a
MEE	0.098	n/a	n/a	n/a
MEE2	0.220	n/a	n/a	n/a
MEE4	0.294	n/a	n/a	n/a
REUSE	0.430	n/a	n/a	n/a

Table 3: Average Kendall's tau-like correlation results grouped by language pairs: trained language pairs (en-de, en-ru, zh-en), from English (en-x), into English (x-en) and language pairs not involving English (x-y). The horizontal lines delimit baseline metrics (top), all language pairs participating reference-based metrics (second), all language pairs participating reference-free metrics (third) and trained language pairs only metrics (bottom). The best result for each category is denoted by bold text with a green highlight.

Another possible way to evaluate the metrics' performance is not to look at the phenomena but rather at the results on different language pairs. Since ACES covers 146 language pairs and for some of these language pairs we only have very few examples, we decide to split this analysis into four main categories:

- **trained:** language pairs for which this year's WMT metrics shared task provided training material (en-de, en-ru and zh-en). This category also allows us to analyse the metrics that only cover these specific language pairs and not the full set of language pairs in ACES.

- **en-x:** language pairs where the source language is English.

- **x-en:** language pairs where the target language is English.

- **x-y:** all remaining language pairs, where neither the source language nor the target language are English.

Table 3 shows the results for all metrics. It is important to note that the results for different language pair categories cannot be directly compared because the examples and covered phenomena categories are not necessarily the same. However, we can compare metrics on each of the language pair groups individually. First, it can be observed that most submitted metrics outperform the baseline metrics (first group). This shows that the field is advancing and MT evaluation metrics have improved since last year. The best performing metrics on all four language pair categories are reference-free metrics which suggests that quality estimation may now be done more reliably at the segment level even without a reference.

Interestingly, the six metrics that only cover the trained language pairs (last group in the table) do not outperform the other metrics on the "trained" category and even perform worse than most baseline metrics. This indicates that finetuning large multilingual pretrained models does not only allow MT evaluation in more language pairs but also improves performance in those directions where training material is available.

8 Analysis

Aside from high-level evaluations of which metrics perform best, we are also interested in metric-spanning weaknesses that we can identify using ACES. This section shows an analysis of three general questions that we aim to answer using ACES.

8.1 How sensitive are metrics to the source?

We designed our challenge sets for the type of "ambiguous translation" (see Section 5.2) in a way that the correct translation candidate given an ambiguous reference can only be identified through the source sentence. Here, we present a targeted evaluation intended to provide some insights into how important the source is for different metrics. We exclude all metrics that do not take the source as input and all metrics that do not cover all language pairs from this analysis. That leaves us with seven reference-based metrics and eight reference-free metrics. Table 4 shows the detailed results of each metric on the considered phenomena.

The most important finding is that the reference-free metrics generally perform much better on these challenge sets than the reference-based metrics. This indicates that reference-based metrics rely too much on the reference and in the case of some metrics even ignore the source, as is shown by their average correlation of close to 0. Interestingly, most of these metrics do not randomly guess the correct translation (which is a valid choice when the correct meaning is not identified via the source) but rather they strongly prefer one phenomenon over the other. For example, several metrics show a gender bias either towards female occupation names (female correlations are high, male low) or male occupation names (vice versa). Likewise, most metrics prefer translations with frequent senses for the word-sense disambiguation challenge sets, although the difference between frequent and infrequent is not as pronounced as for gender.

Only metrics that look at the source and exhibit fewer such preferences can perform well on average on this collection of challenge sets. COMET-22 performs best out of the reference-based metrics and COMET-Kiwi performs best of all reference-free metrics. It is noteworthy that there is still a considerable gap between these two models, suggesting that reference-based models should pay more attention to the source when a reference is ambiguous to reach the performance of reference-free metrics.

This finding is also supported by our real-world knowledge commonsense challenge set from Section 5.12.5. If we compare the scores on the examples where the subordinate clauses are missing from both the source and the reference to the ones where they are only missing from the reference, we can directly see the effect of disambiguation through the source. The corresponding correlation gains are shown in Table 5. Except for the two DATScore models, all reference-based model corre-

	since		female		male		wsd		
	causal	temp.	anti.	pro.	anti.	pro.	freq.	infreq.	AVG
Examples	*106*	*106*	*1000*	*806*	*806*	*1000*	*471*	*471*	*4766*
BERTScore	-0.462	0.462	-0.626	-0.219	0.218	0.626	0.210	-0.221	-0.001
COMET-20	-0.019	0.302	-0.620	-0.370	**0.586**	0.772	0.202	-0.079	0.097
COMET-22	-0.415	**0.792**	**0.940**	**1.000**	-0.628	0.374	**0.558**	**0.040**	**0.333**
DATScore.1.2B.ref.not_w.	**0.623**	0.415	0.192	0.236	0.323	0.646	0.113	0.015	0.320
DATScore.1.2B.ref.w.	0.075	0.698	0.752	0.593	-0.119	-0.076	0.304	-0.121	0.263
MS-COMET-22	-0.472	0.528	0.526	0.705	-0.650	-0.414	0.384	-0.295	0.039
UniTE	0.302	-0.340	-0.838	-0.184	0.377	**0.878**	0.236	-0.219	0.027
COMET-QE	-1.000	**0.981**	0.454	0.868	-0.849	-0.390	0.244	-0.210	0.012
COMET-Kiwi	-0.245	0.943	0.964	0.978	0.794	**0.938**	0.648	**0.355**	**0.672**
CROSS-QE	0.208	0.830	0.976	**0.995**	-0.337	0.364	**0.762**	**0.355**	0.519
DATScore.418M.no_ref.not_w.	**0.981**	0.642	0.814	0.794	0.447	0.318	0.057	0.121	0.522
DATScore.418M.no_ref.w.	0.925	0.943	**0.982**	0.958	-0.414	-0.512	0.244	0.104	0.404
HWTSC-Teacher-Sim	-0.453	0.717	0.916	0.772	-0.283	-0.360	0.291	0.096	0.212
KG-BERTScore	0.453	0.830	0.638	0.300	**0.968**	0.682	0.291	0.096	0.532
MS-COMET-QE-22	-0.302	0.623	-0.132	0.258	0.390	0.604	0.482	0.079	0.250

Table 4: Results on the challenge sets where the good translation can only be identified through the source sentence. Upper block are reference-based metrics, lower block are reference-free metrics. Best results for each phenomenon and each group of models is marked in bold and green and the average over all can be seen in the last column.

	corr. gain
BERTScore	0.002
COMET-20	0.054
COMET-22	0.190
DATScore.1.2B.ref.not_w.	**0.811**
DATScore.1.2B.ref.w.	0.787
MS-COMET-22	0.046
UniTE	0.048
COMET-QE	0.012
COMET-Kiwi	0.340
CROSS-QE	0.292
DATScore.418M.no_ref.not_w.	0.963
DATScore.418M.no_ref.w.	**0.983**
HWTSC-Teacher-Sim	0.138
KG-BERTScore	0.434
MS-COMET-QE-22	0.194

Table 5: Results on the real-world knowledge common-sense challenge set with reference-based metrics in the upper block and reference-free metrics in the lower block. The numbers are computed as the difference between the correlation with the subordinate clause in the source and the correlation without the subordinate clause in the source. Largest gains are bolded.

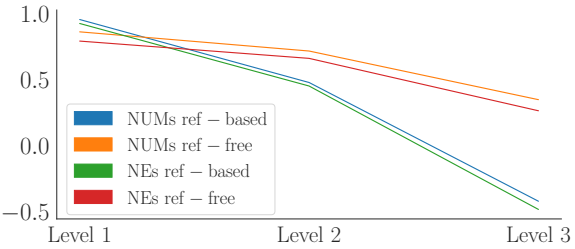

Figure 2: Decrease in correlation for reference-based and reference-free metrics on the named entity and number hallucination challenge sets.

lation scores improve less than most reference-free correlations when access to the subordinate clause is given through the source.

8.2 How much do metrics rely on surface-overlap with the reference?

Another question we are interested in is whether neural reference-based metrics still rely on surface-level overlap with the reference. For this analysis, we use the dataset we created for hallucinated named entities and numbers described in Section 5.3.3. We take the average correlation for all reference-based metrics[15] and the average correlation of all reference-free metrics that cover all languages and plot the decrease in correlation with increasing surface-level similarity of the incorrect translation to the reference. The result can be seen in Figure 2.

We can see that on average reference-based metrics have a much steeper decrease in correlation than the reference-free metrics as the two translation candidates become more and more lexically diverse and the surface overlap between the incorrect translation and the reference increases. This

[15] Excluding surface-level baseline metrics: BLEU and chrf.

	reference-based	reference-free
hallucination	0.37 ± 0.3	-0.03 ± 0.1
overly-literal	0.43 ± 0.2	-0.08 ± 0.1
untranslated	0.56 ± 0.2	-0.06 ± 0.1

Table 6: Average correlation difference and standard deviation between the challenge sets with reference-copied good translations and the challenge sets with the synonymous good translations.

	antonym-replacement	coreference-based-on-commonsense	nonsense
Examples	*133*	*201*	*274*
BERTScore	0.053	-0.092	1.630
BLEURT20	-0.014	-0.227	0.399
COMET-20	0.023	-0.179	1.144
COMET-QE	0.037	-0.451	-0.302
UniTE	0.090	-0.575	0.708
COMET-22	0.060	-0.642	0.546
CROSS-QE	0.161	-0.299	0.166
HWTSC-Teacher-Sim	0.075	-0.015	0.244
COMET-Kiwi	0.015	-0.536	-0.008

Table 7: Correlation difference between the performance of WMT and non-WMT language pairs reported for trained metrics across a subset of examples. δ= τ_{WMT} - τ_{nonWMT}. WMT language pairs consist of a subset of languages seen during training of the metrics, while non-WMT language pairs are unseen. Results show that the metrics are able to generalise to unseen languages.

indicates a possible weakness of reference-based metrics: If one translation is lexically similar to the reference but contains a grave error while others are correct but share less surface-level overlap with the reference, the incorrect translation may still be preferred.

We also show that this is the case for the challenge set where we use an adversarial paraphrase from PAWS-X that shares a high lexical overlap with the reference but does not have the same meaning as an incorrect translation (see Section 5.4). On average, the reference-based metrics only reach a correlation of 0.121 on this challenge set, whereas the reference-free metrics reach a correlation of 0.326.

Finally, we can also see a clear effect of surface-level overlap with the source on three real error challenge sets where we have different versions of the good translation: some where the error was corrected with the corresponding correct token from the reference and some where the error was corrected with a synonym for the correct token from the reference. The reference-based metrics show a much larger difference in correlation between the challenge sets with the reference-copied good translations and the challenge sets with the synonymous good translations. Based on all these results, we conclude that even though state-of-the-art reference-based MT evaluation metrics are not only reliant on surface-level overlap anymore, it still considerably influences their predictions.

8.3 Do multilingual embeddings help design better metrics?

As the community moves towards building metrics that use multilingual encoders, we investigate if some (un)/desirable properties of multilingual embeddings are propagated in these metrics.

8.3.1 Zero-shot Performance

Similar to Kocmi et al. (2021), we investigate whether there is a difference in the performance of metrics on our challenge sets when evaluated on non-WMT language pairs *i.e.* language pairs unseen during the training of the metrics. For this analysis, we include only those metrics for which the training data consisted of some combination of WMT human evaluation data. As different metrics used data from different years, we consider an intersection of languages across these years as WMT language pairs. For a fair comparison, we consider a subset of examples within the phenomenon where at least 100 examples are available of at least one WMT and one non-WMT language pair. We report some of the phenomena in Table 7, where metrics are compared in terms of the correlation difference between the performance on WMT and non-WMT language pairs. (See Appendix A.3 for the original WMT and non-WMT correlation scores.)

We draw similar conclusions to Kocmi et al. (2021), namely that trained metrics are not overfitted to the WMT language pairs. We observe that the median difference of τ between WMT and non-WMT language pairs is 0.019, indicating a good generalisation to unseen languages. We also observe that performance on *harder* phenomena is variable when we compare the results on WMT language pairs versus non-WMT language pairs. In the case of *coreference based on commonsense* (mistranslation), performance is generally better on the non-WMT language pairs[16], while the opposite is (generally) true for the *antonym replacement*

[16]We also observe better performance on non-WMT language pairs for the *similar language high* (wrong translation) phenomenon

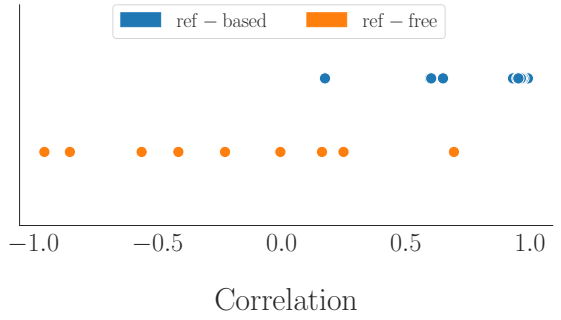

Figure 3: Correlation of reference-based metrics (blue) and reference-free metrics (orange) on the sentence-level untranslated test challenge set.

and *nonsense* phenomena. Further analysis is required to better understand the root cause of this variability, however, in the case of the *coreference based on commonsense* phenomenon, some of the metric training objectives / data may be moving performance in the wrong direction. Note the subset of examples used in this analysis only consists of mid/high resource language pairs; investigation into the performance on low-resource languages is left for future work.

8.3.2 Language Dependent Representations

Multilingual models often learn cross-lingual representations by abstracting away language away from language-specific information (Wu and Dredze, 2019). We are interested in whether the representations are still language-dependent in neural MT evaluation metrics which are trained on such models. For this analysis, we look at the challenge set of sentence-level untranslated text (see Section 5.9.2). We only consider metrics that provided scores for examples in all language pairs.

Figure 3 shows the correlations for all reference-based and reference-free metrics. Unsurprisingly, some reference-free metrics struggle considerably on this challenge set and almost always prefer the copied source to the real translation. This is because the representations of the source and hypothesis are identical, leading to a higher surface and embedding similarity, and thus a higher score. Most reference-based metrics have good to almost perfect correlation and can identify the copied source quite easily. This is expected, as the reference in the target language will act as grounding and the surface-level overlap of the good translation will be much higher in this case. Consequently, the similarity score of the reference to the semantically similar sentence within the same language is likely to be higher than the similarity of semantically similar sentences across languages.

However, there are some exceptions to these trends. UniTE, despite being a reference-based metric only has a correlation of 0.175. This metric may have learned more language-independent representations which will make it harder to identify the untranslated incorrect translation. On the other hand, COMET-Kiwi, which is reference-free, has a correlation of 0.694, which is even better than some reference-based metrics. This suggests that this metric learned to some extent that the translation should be in another language than the source. We also speculate whether the human evaluation data can impart specific knowledge in the embedding space. We leave the investigation of these behaviours as future work.

Thus, while multilingual embeddings help in effective zero-shot transfer to new languages, some properties of the multilingual representation space may need to be altered to suit the task of machine translation evaluation.

9 Recommendations

Based on the metrics results on ACES and our analysis, we derived the following list of recommendations for future MT evaluation metric development:

No metric to rule them all: Both the evaluation on phenomena and on language pair categories in Section 7 showed that there is no single best-performing metric. This divergence is likely to become even larger if we evaluate metrics on different domains. For future work on MT evaluation, it may be worthwhile thinking about how different metrics can be combined to make robust decisions as to which is the best translation. This year's submissions to the metrics shared task already suggest that work in that direction is ongoing as several groups submitted metrics that combined ensembles of several models (COMET-22, COMET-Kiwi, HWTSC_EE_BERTScore*).

The source matters: Our analysis in Section 8.1 highlighted that many reference-based metrics that take the source as input do not consider it enough. Cases, where the correct translation can only be identified through the source, are currently better handled by reference-free metrics. This is a serious shortcoming of reference-based metrics and should be addressed in future research, also considering that many reference-based metrics do not even take the source as input.

Surface-overlap still prevails: In Section 8.2, we showed that despite moving beyond only surface-level comparison to the reference, most reference-based metric scores are still considerably influenced by surface overlap. We expect future metrics to incorporate more lexically diverse references in their training regime to mitigate this issue.

Multilingual embeddings are not perfect: Some properties of multilingual representations, especially, being language-agnostic, can result in undesirable effects on MT evaluation (Section 8.3). It could be helpful for future metrics to incorporate strategies to explicitly model additional language-specific information.

10 Conclusion

We presented ACES, a translation accuracy challenge set based on the MQM ontology. ACES consists of 36,499 examples covering 146 language pairs and representing challenges from 68 phenomena.

We used ACES to evaluate the baseline and submitted metrics from the WMT 2022 metrics shared task. Our overview of metric performance at the phenomena and language level results in Section 7 reveals that there is no single best-performing metric. The more fine-grained analyses in Section 8 highlight that 1) many reference-based metrics that take the source as input do not consider it enough, 2) most reference-based metric scores are still considerably influenced by surface overlap with the reference, and 3) the use of multilingual embeddings can have undesirable effects on MT evaluation. We recommend that these shortcomings of existing metrics should be addressed in future research, and that metric developers should consider a) combining metrics with different strengths in the form of ensemble models, and b) incorporating strategies to explicitly model additional language-specific information (rather than simply relying on multilingual embeddings).

We will make ACES publicly available and hope that it will provide a useful benchmark for MT evaluation metric developers in the future.

Limitations

The ACES challenge set exhibits a number of biases. Firstly, there is greater coverage in terms of phenomena and number of examples for the en-de and en-fr language pairs. This is in part due to the manual effort required to construct examples for some phenomena, in particular those belonging to the discourse-level and real-world knowledge categories. Further, our choice of language pairs is also limited to the ones available in XLM-R. Secondly, ACES contains more examples for those phenomena for which examples could be generated automatically, compared to those that required manual construction/filtering. Thirdly, some of the automatically generated examples require external libraries which are only available for a few languages (e.g. Multilingual Wordnet). Fourthly, the focus of the challenge set is on accuracy errors. We leave the development of challenge sets for fluency errors to future work.

The results and analyses presented in the paper exclude those metrics submitted to the WMT 2022 metrics shared task that provide only system-level outputs. We focus on metrics that provide segment-level outputs as this enables us to provide a broad overview of metric performance on different phenomenon categories and to conduct fine-grained analyses of performance on individual phenomena. For some of the fine-grained analyses, we apply additional constraints based on the language pairs covered by the metrics, or whether the metrics take the source as input, to address specific questions of interest. As a result of applying some of these additional constraints, our investigations tend to focus more on high and medium resource languages than on low resource languages. We hope to address this shortcoming in future work.

Ethics Statement

Some examples within the challenge set exhibit biases, however this is necessary in order to expose the limitations of existing metrics. Wherever external help was required in verifying translations, the annotators were compensated at a rate of £15/hour. Our dataset is based on publicly available datasets and will be released for future use.

References

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

Eleftherios Avramidis, Vivien Macketanz, Arle Lommel, and Hans Uszkoreit. 2018. Fine-grained evaluation of quality estimation for machine translation based on a linguistically motivated test suite. In *Proceedings of the AMTA 2018 Workshop on Translation Quality*

Estimation and Automatic Post-Editing, pages 243–248, Boston, MA. Association for Machine Translation in the Americas.

Luisa Bentivogli, Arianna Bisazza, Mauro Cettolo, and Marcello Federico. 2016. Neural versus phrase-based machine translation quality: a case study. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 257–267, Austin, Texas. Association for Computational Linguistics.

Sheila Castilho, Joss Moorkens, Federico Gaspari, Iacer Calixto, John Tinsley, and Andy Way. 2017. Is neural machine translation the new state of the art? *The Prague Bulletin of Mathematical Linguistics*, 108:109–120.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Verna Dankers, Christopher Lucas, and Ivan Titov. 2022. Can transformer be too compositional? analysing idiom processing in neural machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3608–3626, Dublin, Ireland. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Denis Emelin and Rico Sennrich. 2021. Wino-X: Multilingual Winograd schemas for commonsense reasoning and coreference resolution. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8517–8532, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Michael Auli, and Armand Joulin. 2021. Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research*, 22(107):1–48.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021a. Experts, Errors, and Context: A Large-Scale Study of Human Evaluation for Machine Translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.

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

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

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

Liane Guillou and Christian Hardmeier. 2016. PROTEST: A test suite for evaluating pronouns in machine translation. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 636–643, Portorož, Slovenia. European Language Resources Association (ELRA).

Liane Guillou, Christian Hardmeier, Ekaterina Lapshinova-Koltunski, and Sharid Loáiciga. 2018. A pronoun test suite evaluation of the English–German MT systems at WMT 2018. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 570–577, Belgium, Brussels. 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.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson.

2020. XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 4411–4421. PMLR.

Pierre Isabelle, Colin Cherry, and George Foster. 2017. A challenge set approach to evaluating machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2486–2496, Copenhagen, Denmark. Association for Computational Linguistics.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2021–2031, Copenhagen, Denmark. Association for Computational Linguistics.

Daniel Khashabi, Snigdha Chaturvedi, Michael Roth, Shyam Upadhyay, and Dan Roth. 2018. Looking beyond the surface: A challenge set for reading comprehension over multiple sentences. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 252–262, New Orleans, Louisiana. Association for Computational Linguistics.

Margaret King and Kirsten Falkedal. 1990. Using test suites in evaluation of machine translation systems. In *COLING 1990 Volume 2: Papers presented to the 13th International Conference on Computational Linguistics*.

Tom Kocmi, Christian Federmann, Roman Grundkiewicz, Marcin Junczys-Dowmunt, Hitokazu Matsushita, and Arul Menezes. 2021. To ship or not to ship: An extensive evaluation of automatic metrics for machine translation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 478–494, Online. Association for Computational Linguistics.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X: Papers*, pages 79–86, Phuket, Thailand.

Majid Laali and Leila Kosseim. 2017. Improving discourse relation projection to build discourse annotated corpora. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 407–416, Varna, Bulgaria. INCOMA Ltd.

Ekaterina Lapshinova-Koltunski, Christian Hardmeier, and Pauline Krielke. 2018. ParCorFull: a parallel corpus annotated with full coreference. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Yitong Li, Trevor Cohn, and Timothy Baldwin. 2017. BIBI system description: Building with CNNs and breaking with deep reinforcement learning. In *Proceedings of the First Workshop on Building Linguistically Generalizable NLP Systems*, pages 27–32, Copenhagen, Denmark. Association for Computational Linguistics.

Chi-kiu Lo. 2019. YiSi - a unified semantic MT quality evaluation and estimation metric for languages with different levels of available resources. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 507–513, Florence, Italy. Association for Computational Linguistics.

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

Taylor Mahler, Willy Cheung, Micha Elsner, David King, Marie-Catherine de Marneffe, Cory Shain, Symon Stevens-Guille, and Michael White. 2017. Breaking NLP: Using morphosyntax, semantics, pragmatics and world knowledge to fool sentiment analysis systems. In *Proceedings of the First Workshop on Building Linguistically Generalizable NLP Systems*, pages 33–39, Copenhagen, Denmark. Association for Computational Linguistics.

R. Thomas McCoy and Tal Linzen. 2018. Non-entailed subsequences as a challenge for natural language inference. *CoRR*, abs/1811.12112.

George A. Miller. 1994. WordNet: A lexical database for English. In *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*.

Roberto Navigli and Simone Paolo Ponzetto. 2012. Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

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.

Stefano Perrella, Alessandro Sciré, Lorenzo Proietti, Niccolò Campolungo, and Roberto Navigli. 2022. Matese: Machine translation evaluation as a sequence tagging problem.

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.

Maja Popović and Sheila Castilho. 2019. Challenge test sets for MT evaluation. In *Proceedings of Machine Translation Summit XVII: Tutorial Abstracts*, Dublin, Ireland. European Association for Machine Translation.

Xiaosong Qiao, Chang Su, Yilun Liu, Zhanglin Wu, Ziyang Hui, Peng Li, Yinglu Li, Jiaxin Guo, Minghan Wang, Min Zhang, Shimin Tao, Song Peng, Hao Yang, and Ying Qin. 2022. Hw-tsc submission for wmt2022 metrics task.

Alessandro Raganato, Yves Scherrer, and Jörg Tiedemann. 2019. The MuCoW test suite at WMT 2019: Automatically harvested multilingual contrastive word sense disambiguation test sets for machine translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 470–480, Florence, Italy. Association for Computational Linguistics.

Abhilasha Ravichander, Siddharth Dalmia, Maria Ryskina, Florian Metze, Eduard Hovy, and Alan W Black. 2021. NoiseQA: Challenge set evaluation for user-centric question answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2976–2992, 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.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. 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 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Laura Rimell, Stephen Clark, and Mark Steedman. 2009. Unbounded dependency recovery for parser evaluation. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 813–821, Singapore. Association for Computational Linguistics.

Guido Rocchietti, Flavia Achena, Giuseppe Marziano, Sara Salaris, and Alessandro Lenci. 2021. Fancy: A diagnostic data-set for nli models. In *CLiC-it*.

Rachel Rudinger, Chandler May, and Benjamin Van Durme. 2017. Social bias in elicited natural language inferences. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 74–79, Valencia, Spain. Association for Computational Linguistics.

Thibault Sellam, Amy Pu, Hyung Won Chung, Sebastian Gehrmann, Qijun Tan, Markus Freitag, Dipanjan Das, and Ankur Parikh. 2020. Learning to evaluate translation beyond English: BLEURT submissions to the WMT metrics 2020 shared task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 921–927, Online. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Koustuv Sinha, Robin Jia, Dieuwke Hupkes, Joelle Pineau, Adina Williams, and Douwe Kiela. 2021. Masked language modeling and the distributional hypothesis: Order word matters pre-training for little. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2888–2913, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Noah A. Smith. 2012. Adversarial evaluation for models of natural language. *CoRR*, abs/1207.0245.

Lucia Specia, Zhenhao Li, Juan Pino, Vishrav Chaudhary, Francisco Guzmán, Graham Neubig, Nadir Durrani, Yonatan Belinkov, Philipp Koehn, Hassan Sajjad, Paul Michel, and Xian Li. 2020. Findings of the WMT 2020 shared task on machine translation robustness. In *Proceedings of the Fifth Conference on Machine Translation*, pages 76–91, Online. Association for Computational Linguistics.

Ieva Staliūnaitė and Ben Bonfil. 2017. Breaking sentiment analysis of movie reviews. In *Proceedings of the First Workshop on Building Linguistically Generalizable NLP Systems*, pages 61–64, Copenhagen, Denmark. Association for Computational Linguistics.

Gabriel Stanovsky, Noah A. Smith, and Luke Zettlemoyer. 2019. Evaluating gender bias in machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1679–1684, Florence, Italy. Association for Computational Linguistics.

Antonio Toral and Víctor M. Sánchez-Cartagena. 2017. A multifaceted evaluation of neural versus phrase-based machine translation for 9 language directions. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1063–1073, Valencia, Spain. Association for Computational Linguistics.

Jannis Vamvas and Rico Sennrich. 2021. Contrastive conditioning for assessing disambiguation in MT: A case study of distilled bias. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10246–10265, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jannis Vamvas and Rico Sennrich. 2022. As little as possible, as much as necessary: Detecting over- and undertranslations with contrastive conditioning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 490–500, Dublin, Ireland. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Lucas Nunes Vieira, Minako OâHagan, and Carol Oâ-Sullivan. 2021. Understanding the societal impacts of machine translation: a critical review of the literature on medical and legal use cases. *Information, Communication & Society*, 24(11):1515–1532.

Yu Wan, Dayiheng Liu, Baosong Yang, Haibo Zhang, Boxing Chen, Derek Wong, and Lidia Chao. 2022. UniTE: Unified translation evaluation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8117–8127, Dublin, Ireland. Association for Computational Linguistics.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. 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 833–844, Hong Kong, China. Association for Computational Linguistics.

Yinfei Yang, Yuan Zhang, Chris Tar, and Jason Baldridge. 2019. PAWS-X: A cross-lingual adversarial dataset for paraphrase identification. 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 3687–3692, Hong Kong, China. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Yuan Zhang, Jason Baldridge, and Luheng He. 2019. PAWS: Paraphrase adversaries from word scrambling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1298–1308, Minneapolis, Minnesota. Association for Computational Linguistics.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 15–20, New Orleans, Louisiana. Association for Computational Linguistics.

Jianing Zhou, Hongyu Gong, and Suma Bhat. 2021. PIE: A parallel idiomatic expression corpus for idiomatic sentence generation and paraphrasing. In *Proceedings of the 17th Workshop on Multiword Expressions (MWE 2021)*, pages 33–48, Online. Association for Computational Linguistics.

A Appendix

A.1 Language Codes

Code	Language	Code	Language
af	Afrikaans	ja	Japanese
ar	Arabic	ko	Korean
be	Belarusian	lv	Latvian
bg	Bulgarian	mr	Marathi
ca	Catalan	nl	Dutch
cs	Czech	no	Norwegian
da	Danish	pl	Polish
de	German	pt	Portuguese
el	Greek	ro	Romanian
en	English	ru	Russian
es	Spanish	sk	Slovak
et	Estonian	sl	Slovenian
fa	Persian	sr	Serbian
fi	Finnish	sv	Swedish
fr	French	sw	Swahili
ga	Irish	ta	Tamil
gl	Galician	th	Thai
he	Hebrew	tr	Turkish
hi	Hindi	uk	Ukranian
hr	Croatian	ur	Urdu
hu	Hungarian	vi	Vietnamese
hy	Armenian	wo	Wolof
id	Indonesian	zh	Chinese
it	Italian		

Table 8: ISO 2-Letter language codes of the languages used in the challenge set

A.2 Allowed Unit Conversions

We allow the following unit conversions for the challenge set described in Section 5.3.1:

Distance:

- miles → metres
- kilometres → miles
- kilometres → metres
- metres → feet
- metres → yards
- feet → metres
- feet → yards
- centimetres → inches
- centimetres → millimetres
- inches → centimetres
- inches → millimetres
- millimetres → centimetres
- millimetres → inches
- millimetres → inches

Speed:

- miles per hour → kilometres per hour
- kilometres per hour → miles per hour
- kilometres per second → miles per second
- miles per second → kilometres per second

Time:

- hours → minutes
- minutes → seconds
- seconds → minutes
- days → hours
- months → weeks
- weeks → days

Volume:

- barrels → gallons
- barrels → litres
- gallons → barrels
- gallons → litres

Weight:

- kilograms → grams
- kilograms → pounds
- grams → ounces
- ounces → grams

Area:

- square kilometres → square miles

A.3 Zero Shot Performance Scores

Table 9 contains the Kendall tau-like correlation scores for neural metrics on WMT language pairs (a subset of those seen during training) and non-WMT language pairs (unseen), for three phenomena: antonym replacement, coreference based on commonsense, and nonsense. The table contains the complete set of scores, and complements Table 7, which reports only the difference between the non-WMT and WMT correlation scores. See Section 8.3.1 on zero-shot performance.

	antonym-replacement		**coreference-based -on-commonsense**		**nonsense**	
	WMT	Non-WMT	WMT	Non-WMT	WMT	Non-WMT
BERTScore.	-0.376	-0.429	-0.962	-0.870	0.790	-0.839
BLEURT20.	0.024	0.038	-0.759	-0.532	-0.273	-0.672
COMET-20.	0.136	0.113	-0.722	-0.542	0.706	-0.438
COMET-QE.	0.616	0.579	-0.038	0.413	0.231	0.533
UniTE.	0.488	0.398	-0.570	0.005	0.497	-0.212
COMET-22.	0.744	0.684	-0.608	0.035	0.706	0.161
CROSS-QE.	0.680	0.519	-0.025	0.274	0.720	0.555
HWTSC-Teacher-Sim.	0.504	0.429	0.139	0.154	0.930	0.686
COMET-Kiwi.	0.744	0.729	-0.063	0.473	0.510	0.518

Table 9: Zero-shot performance of neural metrics on three phenomena.