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 such 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[1], a Translation **A**ccuracy **C**halleng**E** **S**et, 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. We recommend: a) combining metrics with different strengths, b) developing metrics that give more weight to the source and less to surface-level overlap with the reference and c) explicitly modelling additional language-specific information beyond what is available via multilingual embeddings.

1 Introduction

Challenge sets have already been created 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[2] (Li et al., 2017; Mahler et al., 2017; Staliūnaitė and Bonfil, 2017), Natural Language Inference (McCoy and Linzen, 2019; 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.

In this work, we describe the University of Zurich - University of Edinburgh submission to the *Challenge Sets* subtask of the Conference on Machine Translation (WMT) 2022 Metrics shared task. Our Translation **A**ccuracy **C**halleng**E** **S**et (ACES) consists of 36,476 examples covering 146 language pairs and representing challenges from 68 phenomena (see Appendix A.4 for the distribution of examples across language pairs and Appendix A.5 for the distribution of language pairs across 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.

[*] Equal contribution by all authors.

[1] Our dataset is available at https://huggingface.co/datasets/nikitam/ACES and the corresponding evaluation scripts at https://github.com/EdinburghNLP/ACES

[2] Submitted to the EMNLP 2017 "Build It Break It" shared task on sentiment analysis

Proceedings of the Seventh Conference on Machine Translation (WMT), pages 479–513
Abu Dhabi, December 7–8, 2022. ©2022 Association for Computational Linguistics

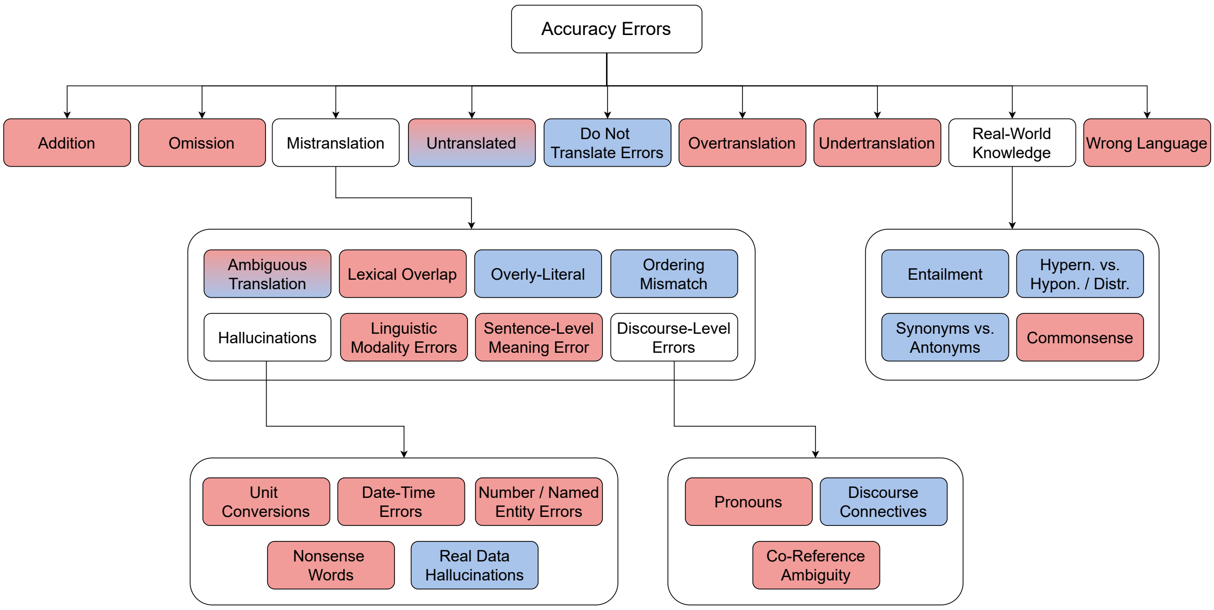

Figure 1: Diagram of the error categories on which our collection of challenge sets is based. Red means challenge sets are created automatically, blue means challenge sets are created manually.

Based on our analysis, we recommend that metric developers consider: a) combining metrics with different strengths, e.g. in the form of ensemble models, b) paying more attention to the source and avoiding reliance on surface-overlap with the reference, and c) 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 monitor which error categories can be identified better, and also whether there are any categories for which metric performance degrades.

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 decided 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 repurposed 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 become worse at identifying error types that were previously considered “solved”. Therefore, we take a 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 (Section 5). We then introduce the metrics that participated in the shared task (Section 6), present an overview of their performance on ACES (Section 7) and detailed analyses (Section 8) that lead to a set of recommendations for future metric development (Section 9).

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, 2019; 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 state-of-the-art models are robust to domain shifts, and whether they have some understanding of linguistic phenomena like negation/commonsense or they simply rely on shallow heuristics. Another line of work under “adversarial datasets” also focuses on creating examples by perturbing the standard test set to fool the model (Smith (2012); Jia and Liang (2017), *inter-alia*).

Challenge sets for evaluating MT systems have focused on the translation models’ ability to generate the correct translation given a 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 have shown improved correlation with human judgements (Freitag et al., 2021b; Kocmi et al., 2021). However, their weaknesses remain relatively unknown and only a small number of works (e.g. Hanna and Bojar (2021) and Amrhein and Sennrich (2022)) have proposed systematic analyses to uncover them.

Previous challenge sets for metric evaluation focused on negation and sentiment polarity (Specia et al., 2020) and synthetic perturbations such as antonym replacement, word omission, number swapping, punctuation removal, etc. (Freitag et al., 2021b). 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. 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[3].

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 Occupation Names and Word Sense Disambiguation, 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 nurse vs. a female doctor). **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, 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 for Commonsense Co-Reference Disambiguation. 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: one which contains an error or phenomenon of interest (the "incorrect" translation) and one which 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 (or two identical copies of the same reference) and insert errors into one of them to form an incorrect translation while the uncorrupted version can be used as the good 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. 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 focus on the 46 languages

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

for which there exists a stanza parser[4] and create datasets for all languages paired with English plus ten additional language pairs that we selected 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.

Only partial translations that can be constructed by deleting spans from the full translation are considered. For translation, we use the M2M100[5] model with 1.2B parameters (Fan et al., 2021).

We 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 it doesn't, we discard the example, if it does, 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 of 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. The metrics then 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 identify the better translation. Since many reference-based metrics, by design, do not include the source to compute evaluation scores, we believe that this presents a challenging test set.

Our method for creating examples is 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): Der Manager 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 explanatory subordinate clauses using a sequence of regular expressions, so that the sentence becomes "The manager fired the baker." where the gender 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 which are all languages that mark gender morphologically on most nouns that refer to a person. For translation, we use Google Translate[6] because we find that this system produces gendered occupation

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

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

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

names that are largely faithful to the disambiguation cues. Finally, we remove explicit translations of "female" and "male" from the German, French or Italian output that would help the disambiguation beyond morphological cues. We predict the gender of the occupation names using the scripts provided by Stanovsky et al. (2019) and only keep translation pairs where both 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 above.

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:

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 necessary, replace them with other cues. We generate three pairs of contrasting disambiguation cues per example and use the question "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 provided by Vamvas and Sennrich (2021). Following this methodology, we create challenge sets for German into English and Russian into English.

5.2.3 Ambiguity - Discourse Connectives

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

SRC (fr):	Aucun test de qualité de l'air n'ait été réalisé dans ce bâtiment **depuis** notre élection.
REF (en):	No air quality test has been done on this particular building **since** we were elected.
✓:	No air quality test has been done on this particular building **from the time** we were elected.
✗:	No air quality test has been done on this particular building **because** we were elected.

The English discourse connective "since" can have either causal or temporal meaning, which is expressed explicitly in both 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 but we also create a German-English version of the challenge set, where we translate the French source sentences into 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 by 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 a 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 - Date-Time Errors

We 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 automatic translations of the FLORES-101 dataset from Section 5.1. We choose all pairs with target languages for which we know the abbreviations for months[7] 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 exclude the example.

5.3.2 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 the 2021 WMT metrics shared task, number differences were not a big issue for most neural metrics (Freitag et al., 2021b). However, we believe that simply changing a number in an alternative translation and using this as an incorrect translation as done by Freitag et al. (2021b) is an overly simplistic setup and does not cover the whole translation hypothesis space.

To address this shortcoming, we propose a three-level evaluation (see examples below). The first, easiest level follows Freitag et al. (2021b) 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 reference 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 between the incorrect translation and the reference increases.

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.

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 Levenshtein distance[8] at the character-level and use spacy[9] (Honnibal et al., 2020) for identifying named entities of type "person". To substitute whole named entities, we make use of the `names`[10] 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.

[7] https://web.library.yale.edu/cataloging/months

[8] https://github.com/life4/textdistance

[9] https://spacy.io/

[10] https://github.com/treyhunner/names

5.3.3 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`[11] 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`[12] to convert unit mentions in the translation into different units. The permitted conversions are listed in Appendix A.2.

The sentence with the converted amount and new 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. We are aware that this challenge set lies beyond the ability of current MT systems and evaluation metrics, however, we believe challenge sets such as these incentivise future work on such capabilities which would reduce the workload in post-editing.

5.3.4 Hallucination - Nonsense Words

We also consider more natural hallucinations at the subword level. Because recent MT systems are trained with subwords (Sennrich et al., 2016), an 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.

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 below, "mass" is broken down as "mas" and "##s" using subwords 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. We use the multilingual BERT (Devlin et al., 2019) tokeniser to replace the subwords.

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.

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.

For this dataset, we manually check the translations of the FLORES-101 dev and devtest 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 part with its correct translation to form the good translation. If possible, we create one good translation by copying the corresponding

[11] https://github.com/nielstron/quantulum3

[12] https://github.com/hgrecco/pint

token(s) from the reference and one with a synonymous token that does not match the reference:

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.

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 has been 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 a high degree of 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 degree of 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 in PAWS-X to obtain the "source". Following this methodology we create examples for each 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"), the 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 manually 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 an *epistemic* meaning (expressing probability or prediction). We also construct examples in which the modal verb is omitted from the incorrect translation.

We employ two strategies to create examples: one in which the modal auxiliary is substituted, and another where it is deleted. We use 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, for example literal translations of figurative language. Here, we look specifically at idioms and at real-data errors.

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[13] 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.

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, 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 to create such examples. We 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 an entailment label are excluded as some examples in the dataset are paraphrases of each other and there would be no sentence-level meaning change. We discuss ef-

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

fects of entailment in Section 5.12.1. We use either the premise or the hypothesis as the reference and an automatic translation as the "good translation". The corresponding premise or hypothesis from the remaining 14 languages is used 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 word order in a way that changes meaning:

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.

This challenge set is created manually by changing translations from the FLORES-101 dataset and covers de→en, en→de and fr→de.

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 automatically 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 instances of "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. we use "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 examples for both the *Comparison.Contrast* sense and 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. Note that while in the previous challenge set the correct translation can be identified by looking at the source, here metrics can only rely on context to identify the correct discourse connective.

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 resolutions. We are interested in whether such commonsense coreference resolutions pose a challenge for MT evaluation metrics:

SRC (en):	It took longer to clean the fish tank than the dog cage because **it** was dirtier.
REF (de):	Das Reinigen des Aquariums dauerte länger als das des Hundekäfigs, da **es** schmutziger war.
✓:	Das Reinigen des Aquariums dauerte länger als das des Hundekäfigs, da **das Aquarium** schmutziger war.
✗ :	Die Reinigung des Aquariums dauerte länger als die des Hundekäfigs, da **er** schmutziger war.

The English sentences in the Wino-X challenge set were sampled from the Winograd schema. All contain the pronoun *it* and were 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 instance of *it* into the correct disambiguating filler, while the second translation contains the pronoun referring to the incorrect antecedent (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 dev and devtest 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 the 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.[14] 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 incorrect translation will be the same, thus leading to a higher score.

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 into 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 as undertranslation while the latter is referred to as overtranslation. For example, “car” may be substituted with “vehicle” (undertranslation) or “BMW” (overtranslation). To automate the generation of such errors, we use Wordnet (Miller, 1994). In our setup 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 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.

5.12 Real-world Knowledge

We manually constructed 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 by the “good” translation. 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

[14] Through observations of Swahili → English translation; unpublished work

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. When constructing our examples we focus solely on leveraging *directional entailments*. We specifically exclude paraphrases as these are bidirectional.

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).

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

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"):

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[15] (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 construct examples in which the good translation contains a hypernym (here "pet") of the word in the reference (here "dog"). We form the incorrect translation by replacing the original word in the source/reference with a different member from the same class (here "cat"; both cats and dogs belong to the class of pets). For example:

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 present in the reference translation.

5.12.4 Real-world Knowledge - Antonyms

Similar to the generation of over- and undertranslations, we also constructed "incorrect" translations by replacing words with their corresponding antonyms from Wordnet. We construct challenge sets for both nouns and verbs.

For nouns, we automatically constructed "incorrect" translations by replacing nouns in the reference with their antonyms. The "good" translation is not amended. This method may result in noisy replacement of nouns with their respective antonyms.

In the case of verbs, we manually constructed a more challenging set of examples intended to be used to assess whether the metrics are able to distinguish between translations that contain a synonym versus an antonym of a given word. We replaced verbs in the reference with a synonym to produce the good translation, and with their antonym to produce the incorrect translation:

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!

For the verbs challenge set, 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

[15] https://wordrel.com/

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 Luft im Haus war kühler als in der Wohnung.
REF (de):	The air in the house was cooler than in the apartment.
✓:	The air in the house was cooler than in the apartment because **the apartment** had a broken air conditioner.
✗:	The air in the house was cooler than in the apartment because **the house** had a broken air conditioner.

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 identifier (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 in a similar language (same typology/same script) to the reference (e.g. a Catalan translation may be used as the incorrect translation if the target language is Spanish). 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.

Note that if we were to compare references for different languages and not an automatic translation vs. a reference, this challenge set should be considered unsolvable for reference-free metrics if there is no way to specify the desired target language. But in this case, we expect reference-free metrics to prefer the reference that we use as the "incorrect translation" since there may be translation errors in the automatically translated "good translation".

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.

5.14.1 Punctuation

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

In strategies 1 and, especially, 3 and 4, some of the examples may also contain accuracy-related errors. For example, the meaning of the sentence could be changed in the incorrect translation if we remove a comma, e.g. in the (in)famous example "Let's eat, Grandma!" vs. "Let's eat Grandma!". 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 by a number of 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.

spBLEU (Goyal et al., 2022) is BLEU computed over text tokenised with a single language-agnostic SentencePiece subword model. The spBLEU baselines, F101SPBLEU and F200SPBLEU, are named according to whether the SentencePiece tokeniser (Kudo and Richardson, 2018) was trained using data from the FLORES-101 or FLORES-200 languages.

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.

BLEURT20 (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 as this is a reference-free metric, only the source and the hypothesis are combined to produce a final score.

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 the relevant system description papers for further details.

COMET-22 (Rei et al., 2022) 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 a 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).

Metric-X is a massive multi-task metric, which fine tunes large language model checkpoints such as mT5 on a variety of human feedback data such as Direct Assessment, MQM, QE, NLI and Summarization Eval. Scaling up the metric is the key to unlocking quality and makes the model work in difficult settings such as evaluating without a reference, evaluating short queries, distinguishing high quality outputs, and evaluating on other generation tasks such as summarisation. The four metrics are referred to according to the mT5 model variant used (xl or xxl) and the fine-tuning data: METRICX_*_DA_2019 only used 2015-19 Direct Assessment data for fine-tuning, whereas METRICX_*_MQM_2020 used a mixture of Direct Assessment 2015-19 and MQM 2020 data.

MS-COMET-22 and **MS-COMET-QE-22** (Kocmi et al., 2022) are built on top of the COMET (Rei et al., 2020) architecture. They are trained on a several times larger set of human judgements covering 113 languages and covering 15 domains. Furthermore, the authors propose filtering of human judgements with potentially low quality. MS-COMET-22 receives the source, the MT hypothesis and the human reference as input, while MS-COMET-QE calculates scores in a quality estimation fashion with access only to the source segment and the MT hypothesis.

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. These are referred to in this paper as UNITE-SRC, UNITE-REF, and UNITE respectively.

COMET-Kiwi (Rei et al., 2022) 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 a hyperparameter search that weights different features extracted from these two QE models and combines them into a single score.

Huawei submitted several metrics to the shared task (Liu et al., 2022). **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. **KG-BERTScore** is a reference-free machine translation evaluation metric, which incorporates a multilingual knowledge graph into BERTScore by linearly combining the results of BERTScore and bilingual named entity matching.

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 Freitag et al. (2021a). MATESE is reference-based, while **MATESE-QE** is its reference-free version, with the source sentence used in place of the reference.

MEE (Mukherjee et al., 2020) is an automatic evaluation metric that leverages the similarity between embeddings of words in candidate and reference sentences to assess translation quality, focusing mainly on adequacy. Unigrams are matched based on their surface forms, root forms and meanings which aims to capture lexical, morphological and semantic equivalence. Semantic evaluation is achieved by using pretrained fasttext embeddings provided by Facebook to calculate the word similarity score between the candidate and reference words. MEE computes an evaluation score using three modules namely exact match, root match and synonym match. In each module, fmean-score is calculated using the harmonic mean of precision and recall by assigning more weightage to recall. The final translation score is obtained by taking average of fmean-scores from individual modules.

MEE2 and **MEE4** (Mukherjee and Shrivastava, 2022b) are improved versions of MEE, focusing on computing contextual and syntactic equivalences along with lexical, morphological and semantic similarity. The intent is to capture fluency and context of the MT outputs along with their adequacy. Fluency is captured using syntactic similarity and context is captured using sentence similarity leveraging sentence embeddings. The final sentence translation score is the weighted combination of three similarity scores: a) Syntactic Similarity achieved by modified BLEU score; b) Lexical, Morphological and Semantic Similarity: measured by explicit unigram matching similar to MEE score; c) Contextual Similarity: Sentence similarity scores are calculated by leveraging sentence embeddings of Language-Agnostic BERT models.

REUSE (Mukherjee and Shrivastava, 2022a) is a REeference-free UnSupervised quality Estimation Metric. This is a bilingual untrained metric. It estimates the translation quality at chunk-level and sentence-level. Source and target sentence chunks are retrieved by using a multi-lingual chunker. Chunk-level similarity is computed by leveraging BERT contextual word embeddings and sentence similarity scores are calculated by leveraging sentence embeddings of Language-Agnostic BERT models. The final quality estimation score is obtained by mean pooling the chunk-level and sentence-level similarity scores.

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 include in the final challenge set 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., 2021b) 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 equal to or lower than the incorrect translation (discordant):

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

Ties are considered as discordant. Note that a higher τ indicates a better performance and that the values can range between -1 and 1.

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 24 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 the last row of Table 1, and without taking outliers into account, we might conclude that addition, undertranslation, real-world knowledge, and wrong language (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 and do not translate (with an average Kendall tau-like correlation of > 0.7) 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.

7.2 ACES Score

To analyse general, high-level, performance trends of the metrics on the ACES challenge set, we define a weighted combination of the top-level categories to derive a single score. We call this score the "ACES - Score":

$$\text{ACES} = sum \left\{ \begin{array}{r} 5 * \tau_{\text{addition}} \\ 5 * \tau_{\text{omission}} \\ 5 * \tau_{\text{mistranslation}} \\ 1 * \tau_{\text{untranslated}} \\ 1 * \tau_{\text{do not translate}} \\ 5 * \tau_{\text{overtranslation}} \\ 5 * \tau_{\text{undertranslation}} \\ 1 * \tau_{\text{real-world knowledge}} \\ 1 * \tau_{\text{wrong language}} \\ 0.1 * \tau_{\text{punctuation}} \end{array} \right\} \quad (1)$$

The weights correspond to the values under the MQM framework that Freitag et al. (2021a) rec-

	addition	**omission**	**mistranslation**	**untranslated**	**do not translate**	**overtranslation**	**undertranslation**	**real-world knowledge**	**wrong language**	**punctuation**	**ACES-Score**
Examples	*999*	*999*	*24457*	*1300*	*100*	*1000*	*1000*	*2948*	*2000*	*1673*	
BLEU	0.748	0.435	-0.229	0.353	0.600	-0.838	-0.856	-0.768	0.661	0.638	-2.79
f101spBLEU	0.662	0.590	-0.084	0.660	0.940	-0.738	-0.826	-0.405	0.638	0.639	-0.09
f200spBLEU	0.664	0.590	-0.082	0.687	0.920	-0.752	-0.794	-0.394	0.658	0.648	0.06
chrF	0.642	0.784	0.162	**0.781**	**0.960**	-0.696	-0.592	-0.294	**0.691**	0.743	3.71
BERTScore	**0.880**	0.750	0.320	0.767	**0.960**	-0.110	-0.190	0.031	0.563	**0.849**	10.65
BLEURT-20	0.437	0.810	0.429	0.748	0.860	0.200	0.014	0.401	0.533	0.649	12.06
COMET-20	0.437	0.808	0.378	0.748	0.900	0.314	0.112	0.267	0.033	0.706	12.27
COMET-QE	-0.538	0.397	0.378	0.135	0.120	0.622	0.442	0.322	-0.505	0.251	6.61
YiSi-1	0.770	0.866	0.356	0.730	0.920	-0.062	-0.076	0.110	0.431	0.734	11.53
COMET-22	0.333	0.806	0.566	0.536	0.900	0.690	0.538	0.574	-0.318	0.539	16.41
metricx_xl_DA_2019	0.395	0.852	0.545	0.722	0.940	0.692	0.376	**0.740**	0.521	0.670	17.29
metricx_xl_MQM_2020	-0.281	0.670	0.523	0.579	-0.740	0.718	**0.602**	0.705	-0.126	0.445	13.10
metricx_xxl_DA_2019	0.303	0.832	0.580	0.762	0.920	0.572	0.246	0.691	0.250	0.630	15.35
metricx_xxl_MQM_2020	-0.099	0.534	0.578	0.651	0.880	**0.752**	0.552	0.712	-0.321	0.369	13.54
MS-COMET-22	-0.219	0.686	0.397	0.504	0.700	0.548	0.290	0.230	0.041	0.508	10.03
UniTE	0.439	0.876	0.501	0.571	0.920	0.496	0.302	0.624	-0.337	0.793	14.93
UniTE-ref	0.359	0.868	0.535	0.412	0.840	0.640	0.398	0.585	-0.387	0.709	15.52
COMETKiwi	0.361	0.830	**0.631**	0.230	0.780	0.738	0.574	0.582	-0.359	0.490	16.95
Cross-QE	0.163	0.876	0.546	-0.094	0.320	0.726	0.506	0.446	-0.374	0.455	14.43
HWTSC-Teacher-Sim	-0.031	0.495	0.406	-0.269	0.700	0.552	0.456	0.261	-0.021	0.271	10.09
HWTSC-TLM	-0.363	0.345	0.384	0.154	-0.040	0.544	0.474	0.071	-0.168	0.634	7.00
KG-BERTScore	0.790	0.812	0.489	-0.456	0.760	0.654	0.528	0.487	0.306	0.255	**17.49**
MS-COMET-QE-22	-0.177	0.678	0.439	0.388	0.240	0.518	0.386	0.248	-0.197	0.523	9.95
UniTE-src	0.285	**0.930**	0.599	-0.615	0.860	0.698	0.540	0.537	-0.417	0.733	15.70
Average	0.290	0.713	0.389	0.404	0.735	0.312	0.167	0.282	0.075	0.578	10.91

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. The last column shows the ACES-Score, a weighted sum of the correlations. The ACES-Score ranges from -29.1 (all phenomena have a correlation of -1) to 29.1 (all phenomena have a correlation of +1).

ommend for major (weight=5), minor (weight=1) and fluency/punctuation errors (weight=0.1). We determined 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. We also include real-world knowledge under 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.

We caution that our weighting for the ACES-Score is not ideal, as some phenomena within a broad category might be more difficult than others. Still, we believe that an ACES-Score will be helpful to quickly identify changes in performance of a metric (e.g. following modifications), prior to conducting in-depth analyses at the category and sub-category levels. The ACES-Score ranges from -29.1 (all phenomena have a correlation of -1) to 29.1 (all phenomena have a correlation of +1).

The ACES-Score results can be seen in the last column of Table 1. Using the ACES-Score, we can see at a glance that the majority of the metrics submitted to the WMT 2022 shared task outperform the baseline metrics. Interestingly, many reference-free metrics also perform on par with reference-based metrics. The best performing metric is a reference-free metric, namely KG-BERTScore, closely followed by the reference-based metric METRICX_XL_DA_2019. Perhaps unsurprisingly, the worst performing metric is BLEU. However, we caution against making strong claims about which metrics perform *best* or *worst* on the challenge set based on this score alone. 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 and because the scales of different metrics are not comparable. We leave an analysis of metrics' confidence on different error types for future work.

	disco.	**halluci.**	**other**
Examples	*3698*	*10270*	*10489*
BLEU	-0.048	-0.420	-0.251
f101spBLEU	0.105	-0.206	-0.153
f200spBLEU	0.094	-0.191	-0.149
chrF	0.405	-0.137	0.161
BERTScore	0.567	-0.058	0.362
BLEURT-20	0.695	0.142	0.402
COMET-20	0.641	0.016	0.399
COMET-QE	0.666	0.303	0.208
YiSi-1	0.609	0.019	0.368
COMET-22	0.682	0.461	0.542
metricx_xl_DA_2019	0.701	0.493	0.458
metricx_xl_MQM_2020	0.573	0.677	0.394
metricx_xxl_DA_2019	0.768	0.541	0.463
metricx_xxl_MQM_2020	0.716	**0.713**	0.392
MS-COMET-22	0.645	0.148	0.360
UniTE	0.746	0.322	0.424
UniTE-ref	**0.776**	0.396	0.437
COMETKiwi	0.733	0.493	**0.637**
Cross-QE	0.644	0.395	0.563
HWTSC-Teacher-Sim	0.594	0.296	0.330
HWTSC-TLM	0.756	0.306	0.151
KG-BERTScore	0.593	0.387	0.472
MS-COMET-QE-22	0.626	0.243	0.416
UniTE-src	0.772	0.463	0.551
Average	0.586	0.242	0.331

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.

7.3 Mistranslation Results

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. We present additional, more fine-grained, performance analyses for individual phenomena in Section 8.

7.4 Language-level Results

	trained	en-x	x-en	x-y
Examples	8871	12695	17966	5815
BLEU	0.009	0.225	-0.370	-0.121
f101spBLEU	0.148	0.170	-0.290	-0.022
f200spBLEU	0.140	0.442	-0.286	-0.004
chrF	0.325	0.392	-0.047	0.098
BERTScore	0.479	0.031	0.173	0.125
BLEURT-20	0.541	0.327	0.280	0.257
COMET-20	0.495	0.379	0.278	0.121
COMET-QE	0.356	0.166	0.144	0.168
YiSi-1	0.476	0.520	0.185	0.150
COMET-22	0.599	0.486	0.554	0.355
metricx_xl_DA_2019	0.622	0.458	0.456	**0.551**
metricx_xl_MQM_2020	0.608	0.567	0.452	0.509
metricx_xxl_DA_2019	0.631	0.431	0.462	0.528
metricx_xxl_MQM_2020	0.605	**0.572**	0.487	0.502
MS-COMET-22	0.415	0.312	0.323	0.117
UniTE	0.635	0.452	0.406	0.283
UniTE-ref	0.619	0.313	0.413	0.305
COMETKiwi	0.620	0.510	**0.694**	0.468
Cross-QE	0.598	0.401	0.552	0.291
HWTSC-Teacher-Sim	0.497	0.357	0.352	0.149
HWTSC-TLM	0.538	0.519	0.167	0.194
KG-BERTScore	0.485	0.428	0.507	0.347
MS-COMET-QE-22	0.483	0.488	0.411	0.257
UniTE-src	**0.658**	0.445	0.582	0.328
MATESE	-0.281	n/a	n/a	n/a
MEE	-0.078	n/a	n/a	n/a
MEE2	0.340	n/a	n/a	n/a
MEE4	0.391	n/a	n/a	n/a
REUSE	0.430	n/a	n/a	n/a
MATESE-QE	-0.313	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 again 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 (i.e. 2021).

Interestingly, the six metrics that only scored the trained language pairs (last group in the table) do not outperform the other metrics on the "trained" category. Note, however, that the MEE* metrics and REUSE are unsupervised metrics and that the MATESE metrics only used MQM training data. Therefore, we cannot comment on whether creating metrics that are specific to a language pair would result in better metrics. In any case, our findings in Section 8.3.1 suggest that generalisation to unseen language pairs is generally quite good for the multilingual metrics which might be a more desirable property than increased performance on specific language pairs.

8 Analysis

Aside from high-level evaluations of which metrics perform best, we are mostly 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 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, all metrics

	since		female		male		wsd		
	causal	**temp.**	**anti.**	**pro.**	**anti.**	**pro.**	**freq.**	**infreq.**	**AVG**
Examples	*106*	*106*	*1000*	*806*	*806*	*1000*	*471*	*471*	*4766*
BERTScore	-0.434	0.434	-0.614	-0.216	0.208	0.618	0.214	-0.223	-0.001
COMET-20	-0.019	0.302	-0.622	-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**
metricx_xxl_DA_2019	-0.849	0.811	-0.944	-0.228	0.233	**0.942**	0.032	-0.028	-0.004
metricx_xxl_MQM_2020	-1.000	**1.000**	-0.878	0.002	-0.007	0.884	0.083	-0.100	-0.002
MS-COMET-22	-0.604	0.623	0.296	0.640	-0.342	0.046	0.316	-0.155	0.102
UniTE	**0.038**	-0.075	-0.890	-0.213	0.377	0.934	0.270	-0.223	0.027
COMET-QE	-1.000	**0.981**	0.450	0.871	-0.854	-0.382	0.244	-0.210	0.013
COMET-Kiwi	-0.245	0.943	0.964	0.978	0.794	**0.938**	0.648	**0.363**	**0.673**
Cross-QE	0.208	0.830	**0.976**	**0.995**	-0.337	0.364	**0.762**	0.355	0.519
HWTSC-Teacher-Sim	-0.453	0.717	0.916	0.772	-0.283	-0.360	0.295	0.079	0.210
KG-BERTScore	**0.453**	0.830	0.638	0.300	**0.968**	0.682	0.295	0.079	0.531
MS-COMET-QE-22	-0.283	0.792	-0.194	0.320	0.246	0.694	0.465	0.002	0.255
UniTE-src	-0.321	0.906	**0.976**	0.980	0.171	0.736	0.622	0.346	0.552

Table 4: Results on the challenge sets where the good translation can only be identified through the source sentence. Upper block: reference-based metrics, lower block: 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.

that do not cover all language pairs, and the smaller versions of METRIC-X (metricx_xl_DA_2019 and metricx_xl_MQM_2020) from this analysis. This leaves us with seven reference-based metrics and seven 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. Interestingly, most of the metrics that seem to ignore the source do not randomly guess the correct translation (which is a valid alternative 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.

	corr. gain
BERTScore	0.002
COMET-20	0.060
COMET-22	**0.190**
metricx_xxl_DA_2019	0.012
metricx_xxl_MQM_2020	-0.016
MS-COMET-22	0.050
UniTE	0.042
COMET-QE	0.018
COMET-Kiwi	**0.338**
Cross-QE	0.292
HWTSC-Teacher-Sim	0.154
KG-BERTScore	0.154
MS-COMET-QE-22	0.196
UniTE-src	0.216

Table 5: Results on the real-world knowledge commonsense 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.

This finding is also supported by our real-world knowledge commonsense challenge set. If we compare the scores on the examples where the subor-

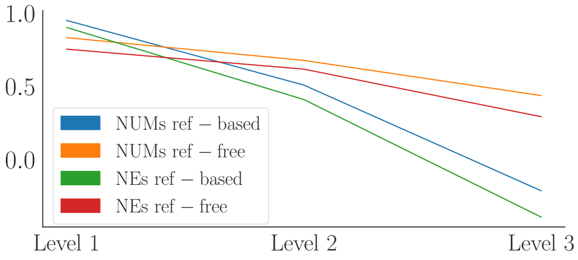

Figure 2: Decrease in correlation for reference-based and reference-free metrics on the named entity and number hallucination challenge sets.

dinate 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. All reference-based model correlation scores improve less than most reference-free correlations when access to the subordinate clause is given through the source. This highlights again that reference-based metrics do not give enough weight to the source sentence.

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. We take the average correlation for all reference-based metrics[16] 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 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.

	reference-based	reference-free
hallucination	-0.22 ± 0.16	+0.04 ± 0.07
overly-literal	-0.32 ± 0.16	+0.12 ± 0.09
untranslated	-0.44 ± 0.18	+0.03 ± 0.06

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.

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 degree of lexical overlap with the reference but does not have the same meaning as an incorrect translation. On average, the reference-based metrics only reach a correlation of 0.05 ± 0.12 on this challenge set, whereas the reference-free metrics reach a correlation of 0.23 ± 0.15. This shows that reference-based metrics are less robust when the incorrect translation has high lexical overlap with the reference.

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. As seen in Table 6, the reference-based metrics show a much larger difference in correlation between the challenge sets with reference-copied good translations and the challenge sets with the synonymous good translations, than the reference-free metrics. For example, for the hallucination test set, reference-free metrics have very similar average performance when the good translation contains the same word as the reference vs. when it contains a synonym (δ of +0.04). On the other hand, the reference-based metrics lose on average -0.22 in correlation when the good translation contains the synonym rather than the same word as the reference. Based on all of 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, such overlap 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 em-

[16] Excluding surface-level baseline metrics: BLEU, SPBLEU and CHRF.

beddings 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 from those phenomena where we have least 100 examples in WMT languages and 100 examples in non-WMT languages, irrespective of the number of examples per individual 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 and the list of language pairs).

	antonym-replacement	real-world knowledge commonsense	nonsense
Examples	*131*	*201*	*239*
BERTScore	0.032	-0.054	1.469
BLEURT-20	0.032	0.201	0.350
COMET-20	0.048	0.067	1.021
COMET-QE	-0.048	-0.188	-0.294
COMET-22	0.080	0.027	0.531
metricx_xl_DA_2019	-0.032	-0.054	0.434
metricx_xl_MQM_2020	-0.048	-0.094	0.182
metricx_xxl_DA_2019	0.016	-0.040	0.266
metricx_xxl_MQM_2020	0.064	-0.067	0.196
UniTE-ref	-0.032	0.013	0.238
UniTE	0.080	0.000	0.643
COMETKiwi	0.048	-0.027	0.042
Cross-QE	0.064	0.188	0.182
HWTSC-Teacher-Sim	0.208	0.081	0.350
UniTE-src	0.096	0.161	-0.028

Table 7: Correlation difference between the performance of WMT and non-WMT language pairs reported for trained metrics across a subset of examples. $\delta = \tau_{\text{WMT}} - \tau_{\text{non WMT}}$. 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.

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 0.056, indicating a good generalisation to unseen languages. We still note that performance on the phenomena is variable when we compare the results on WMT language pairs versus non-WMT language pairs. In the case of real-world knowledge commonsense, performance is slightly better on the non-WMT language pairs[17], while the opposite is (generally) true for the antonym replacement and, especially, the nonsense phenomena for certain metrics. Further analysis is required to better understand metric behaviour on zero-shot language pairs, especially considering that some of the analysed non-WMT language pairs have a target language that is also the target language in at least one of the WMT language pairs (e.g. English).

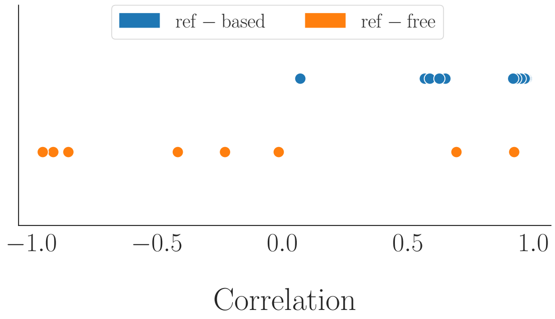

Figure 3: Correlation of reference-based metrics (blue) and reference-free metrics (orange) on the sentence-level untranslated test challenge set.

8.3.2 Language Dependent Representations

Multilingual models often learn cross-lingual representations by abstracting 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 sentence-level untranslated text challenge set (see Figure 3) and wrong language phenomena (see Table 1). 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. The representations of the source and the incorrect translation are identical, leading to a higher surface and embedding similarity, and thus a higher score. We do, however, find some exceptions to this trend

[17] We also observe better performance on non-WMT language pairs for the similar language high phenomenon.

- COMET-KIWI and MS-COMET-QE-22 both have a high correlation on sentence-level untranslated text. This suggests that these metrics could have learnt language-dependent representations.

Most reference-based metrics have good to almost perfect correlation and can identify the copied source quite easily. As reference-based metrics tend to ignore the source (see Section 8.2), the scores are based on the similarity between the reference and the MT output. In this challenge set, the similarity between the good-translation and the reference is likely to be higher than the incorrect-translation and the reference. The former MT output is in the same language as the reference and will have more surface level overlap. We believe the reference here acts as grounding.

However, this grounding property of the reference is only robust when the source and reference languages are dissimilar, as is the case with language pairs in the sentence-level untranslated text challenge set. We find that reference-based metrics struggle on wrong language phenomena (see Table 1) where the setup is similar, but now the incorrect translation and the reference are from similar languages (e.g. one is in Hindi and the other is in Marathi). Naturally, there will be surface level overlap between the reference and both the good-translation and the incorrect-translation. For example, both Marathi and Hindi use named entities with identical surface form, and so these will appear in the reference and also in both the good-translation and the incorrect-translation. Thus, the semantic content drives the similarity scores between the MT outputs and the references. It is possible that the human translation in the similar language (labelled as the incorrect-translation) has a closer representation to the human reference because in the MT output (labelled as the good-translation) some semantic information may be lost. We leave further investigation of this for future work.

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 some groups submitted metrics that combined ensembles of models or multiple components (COMET-22, COMET-KIWI, KG-BERTSCORE, MEE*, REUSE).

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-level overlap. We expect future metrics to use 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 36,476 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 levels 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 be addressed in future research, and that metric developers should consider a) combining metrics with different strengths, e.g. in the form of ensemble models, b) developing metrics that give more weight to the source and less to surface-level overlap with the reference, and c) incorporating strategies to explicitly model additional language-specific information (rather than simply relying on multilingual embeddings).

We have made 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.

As a result of using existing datasets as the basis for many of the examples, errors present in these datasets may be propagated through into ACES. Whilst we acknowledge that this is undesirable, in our methods for constructing the *incorrect translation* we aim to ensure that the quality of the *incorrect translation* is always worse than the corresponding *good translation*.

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 challenge set is based on publicly available datasets and will be released for future use.

Acknowledgements

We thank the organisers of the WMT 2022 Metrics task for setting up this shared task and for their feedback throughout the process, and the shared task participants for scoring our challenge sets with their systems. We are grateful to Stephanie Droop, Octave Mariotti, and Kenya Murakami for helping us with the annotations. We thank the StatMT group at Edinburgh, especially Alexandra Birch, Barry Haddow, and Ulrich Germann, and the attendees at the MT Marathon 2022 for their valuable feedback. We thank Janis Goldzycher, Mark Steedman, Rico Sennrich, and the anonymous reviewers for their insightful comments and suggestions. This work was supported in part by the UKRI Centre for Doctoral Training in Natural Language Processing, funded by the UKRI (grant EP/S022481/1) and the University of Edinburgh (Moghe), by the Swiss National Science Foundation (project MUTAMUR; no. 176727) (Amrhein) and by the ERC H2020 Advanced Fellowship GA 742137 SEMANTAX (Guillou). We also thank Huawei for their support (Moghe).

References

Chantal Amrhein and Rico Sennrich. 2022. Identifying weaknesses in machine translation metrics through minimum bayes risk decoding: A case study for COMET. In *2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing*, Online. Association for Computational Linguistics.

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

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

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.

Tom Kocmi, Hitokazu Matsushita, and Christian Federmann. 2022. MS-COMET: More and Better Human Judgements Improve Metric Performance. In *Proceedings of the Seventh Conference on Machine Translation*, Abu Dhabi. 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.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

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.

Yilun Liu, Xiaosong Qiao, Zhanglin Wu, Su Chang, Min Zhang, Yanqing Zhao, shimin tao Song Peng, Hao Yang, Ying Qin, Jiaxin Guo, Minghan Wang, Yinglu Li, Peng Li, and Xiaofeng Zhao. 2022. Partial Could Be Better Than Whole: HW-TSC 2022 Submission for the Metrics Shared Task. In *Proceedings of the Seventh Conference on Machine Translation*, Abu Dhabi. 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.

Richard T McCoy and Tal Linzen. 2019. Non-entailed subsequences as a challenge for natural language inference. *Proceedings of the Society for Computation in Linguistics (SCiL)*, pages 358–360.

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*.

Ananya Mukherjee, Hema Ala, Manish Shrivastava, and Dipti Misra Sharma. 2020. Mee : An automatic metric for evaluation using embeddings for machine translation. In *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 292–299.

Ananya Mukherjee and Manish Shrivastava. 2022a. REUSE: REference-free UnSupervised quality Estimation Metric. In *Proceedings of the Seventh Conference on Machine Translation*, Abu Dhabi. Association for Computational Linguistics.

Ananya Mukherjee and Manish Shrivastava. 2022b. Unsupervised Embedding-based Metric for MT Evaluation with Improved Human Correlation. In *Proceedings of the Seventh Conference on Machine Translation*, Abu Dhabi. Association for Computational Linguistics.

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, Lorenzo Proietti, Alessandro Scirè, Niccolò Campolungo, and Roberto Navigli. 2022. Machine Translation Evaluation as a Sequence Tagging Problem. In *Proceedings of the Seventh Conference on Machine Translation*, Abu Dhabi. Association for Computational Linguistics.

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

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.

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, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. COMET-22: Unbabel-IST 2022 Submission for the Metrics Shared Task. In *Proceedings of the Seventh Conference on Machine Translation*, Abu Dhabi. 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 *Proceedings of the Eighth Italian Conference on Computational Linguistics (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 O'Hagan, and Carol O'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, Keqin Bao, Dayiheng Liu, Baosong Yang, Derek F. Wong, Lidia S. Chao, Wenqiang Lei, and Jun Xie. 2022. Alibaba-Translate China's Submission for WMT2022 Metrics Shared Task. In *Proceedings of the Seventh Conference on Machine Translation*, Abu Dhabi. 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	lt	Lithuanian
bg	Bulgarian	lv	Latvian
ca	Catalan	mr	Marathi
cs	Czech	nl	Dutch
da	Danish	no	Norwegian
de	German	pl	Polish
el	Greek	pt	Portuguese
en	English	ro	Romanian
es	Spanish	ru	Russian
et	Estonian	sk	Slovak
fa	Persian	sl	Slovenian
fi	Finnish	sr	Serbian
fr	French	sv	Swedish
ga	Irish	sw	Swahili
gl	Galician	ta	Tamil
he	Hebrew	th	Thai
hi	Hindi	tr	Turkish
hr	Croatian	uk	Ukranian
hu	Hungarian	ur	Urdu
hy	Armenian	vi	Vietnamese
id	Indonesian	wo	Wolof
it	Italian	zh	Chinese

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

A.2 Permitted Unit Conversions

We allow the following unit conversions for the challenge set that covers such errors:

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, real-world knowledge 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. We shall now list the language pairs across the different phenomena:

Antonym Replacement

WMT:	de-en
non-WMT:	ko-en, es-en

Real-world Knowledge - Commonsense

WMT:	de-en, ru-en, en-ru, en-de
non-WMT:	ru-de, fr-ru, ru-fr, de-ru

Nonsense

WMT:	de-en
non-WMT:	fr-ja, ko-ja, en-ko, ko-en

Note that 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.

A.4 Distribution of Examples Across Language Pairs

Table 10 contains the total number of examples per language pair in the challenge set. As can be seen in the table, the distribution of examples is variable across language pairs. The dominant language pairs are: en-de, de-en, and fr-en.

A.5 Distribution of Language Pairs Across Phenomena

Table 11 contains the list of language pairs per phenomena in the challenge set. As can be seen in the table, the distribution of language pairs is variable across phenomena. Addition and omission have the highest variety of language pairs. en-de is the most frequent language pair across all phenomena.

	antonym-replacement		**real-world knowledge -commonsense**		**nonsense**	
	WMT	Non-WMT	WMT	Non-WMT	WMT	Non-WMT
BERTScore	-0.376	-0.408	0.007	0.060	0.790	-0.678
BLEURT-20	0.024	-0.008	0.396	0.195	-0.273	-0.622
COMET-20	0.152	0.104	0.087	0.020	0.706	-0.315
COMET-QE	0.616	0.664	0.168	0.356	0.245	0.538
COMET-22	0.744	0.664	0.584	0.557	0.706	0.175
metricx_xl_DA_2019	0.728	0.760	0.570	0.624	0.790	0.357
metricx_xl_MQM_2020	0.888	0.936	0.517	0.611	0.944	0.762
metricx_xxl_DA_2019	0.312	0.296	0.718	0.758	0.706	0.441
metricx_xxl_MQM_2020	0.696	0.632	0.691	0.758	0.930	0.734
UniTE-ref	0.664	0.696	0.409	0.396	0.091	-0.147
UniTE	0.632	0.552	0.409	0.409	0.441	-0.203
COMETKiwi	0.744	0.696	0.745	0.772	0.510	0.469
Cross-QE	0.680	0.616	0.638	0.450	0.720	0.538
HWTSC-Teacher-Sim	0.504	0.296	0.248	0.168	0.930	0.580
UniTE-src	0.776	0.680	0.651	0.490	0.524	0.552

Table 9: Zero-shot performance of neural metrics on three phenomena to measure the ability of metrics to generalise to new language pairs. 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.

src ↓ \ tgt →	af	ar	be	bg	ca	cs	da	de	el	en	es	et	fa	fi	fr	ga	gl	he	hi	hr	hu	hy	id	it	ja	ko	lt	lv	mr	nl	no	pl	pt	ro	ru	sk	sl	sr	sv	sw	ta	th	tr	uk	ur	vi	wo	zh
af										96			25																																			
ar										361					102				17																													
be										67																																						
bg										393																	40																					
ca										79	175																																					
cs										85																																						
da										83																																						
de										4163	84				394										113	63									104													75
el										387																																						
en	25	5	6	15	347	368	46	6964	21		725	25	20	12	800		16	18	343	27	44	3	31	10	430	545	17	19	52	50	53	349	44	46	698	27	45	15	39		1		10	16	10	25		333
es					88			64		1263					125										117	74																						67
et										70																																						
fa	16									85																																						
fi										79																																						
fr								683		2868	78														403	59			344						46													61
ga										17																																						
gl										70																																						
he										59																													51									
hi		8								367																																						
hr										81																		29																				
hu										53																																						
hy										48																																				13		
id										63																																						
it										801																																						
ja								60		912	67				122											163																						74
ko								70		1004	72				110										358																							73
lt				28						68																																						
lv										61										24																												
mr										63																																						
nl										73																																						
no										53																																						
pl										65																			111							58												
pt										89																												40										
ro										91																																						
ru								106		472	87				42																																	
sk										54																						17																
sl										69																																						
sr										64																							54															
sv										79								28																														
sw										327																																						
ta										39																																						
th										299																																						
tr										386																																						
uk										77																																						
ur										372																																						
vi										391												3																										
wo										11																																						
zh								150		1209	59				113										128	80																						

Table 10: Number of examples per language pair. Rows: source language; Columns: target language

phenomena	language pairs	phenomena	language pair
ambiguous-translation-wrong-discourse-connective-since-causal ambiguous-translation-wrong-discourse-connective-since-temporal hallucination-unit-conversion-unit-matches-ref	fr-en, de-en	hallucination-real-data-vs-ref-word	en-de, de-en, fr-de
ambiguous-translation-wrong-discourse-connective-while-contrast	fr-en	hallucination-real-data-vs-synonym	en-mr, de-en, en-de, fr-de
ambiguous-translation-wrong-discourse-connective-while-temporal	fr-en	untranslated-vs-ref-word	en-de, de-en, fr-de
ambiguous-translation-wrong-gender-female-anti	fr-en, de-en, it-en	untranslated-vs-synonym	en-de, de-en, fr-de
ambiguous-translation-wrong-gender-male-anti	fr-en, de-en, it-en	modal_verb:deletion	de-en
ambiguous-translation-wrong-gender-male-pro	fr-en, de-en, it-en	modal_verb:substitution	de-en
ambiguous-translation-wrong-sense-frequent	en-de, en-ru	nonsense	ko-en, ko-ja, en-ko, fr-ja, de-en
ambiguous-translation-wrong-sense-infrequent	en-de, en-ru	ordering-mismatch	en-de, de-en, fr-de
anaphoric_group_it-they:deletion	en-de	overly-literal-vs-correct-idiom	en-de, de-en
anaphoric_group_it-they:substitution	en-de	overly-literal-vs-explanation	en-de, de-en
anaphoric_intra_non-subject_it:deletion	en-de	overly-literal-vs-ref-word	en-de, de-en, fr-de
anaphoric_intra_non-subject_it:substitution	en-de	overly-literal-vs-synonym	en-mr, de-en, en-de, fr-de
anaphoric_intra_subject_it:deletion	en-de	pleonastic_it:deletion	en-de
anaphoric_intra_subject_it:substitution	en-de	pleonastic_it:substitution_pro_trans_different_to_ref	en-de
anaphoric_intra_they:deletion	en-de	punctuation:deletion_all	en-de
anaphoric_intra_they:substitution	en-de	punctuation:deletion_commas	en-de
anaphoric_singular_they:deletion	en-de	punctuation:deletion_quotes	en-de
anaphoric_singular_they:substitution	en-de	punctuation:statement-to-question do-not-translate	en-de
antonym-replacement	fr-en, ko-en, ja-en, es-en, zh-en, de-en	real-world-knowledge-entailment	en-de, de-en
similar-language-high	en-hi, en-cs, en-es	real-world-knowledge-hypernym-vs-distractor	en-de, de-en
similar-language-low	fr-mr, en-pl, en-ca	real-world-knowledge-hypernym-vs-hyponym	en-de, de-en
coreference-based-on-commonsense	en-de, en-ru, en-fr	real-world-knowledge-synonym-vs-antonym	en-de, de-en
hallucination-named-entity-level-1 hallucination-named-entity-level-2 hallucination-named-entity-level-3 hallucination-number-level-1 hallucination-number-level-2 hallucination-number-level-3	en-de, ja-de, en-ko, de-zh, ja-en, es-de, fr-en, es-ko, ko-ja, es-ja, de-ja, zh-es, fr-zh, fr-ja, es-en, fr-ko, zh-en, ko-de, ko-es, de-ko, ko-en, fr-es, ja-es, ja-ko, zh-fr, en-es, de-en, ja-fr, ko-zh, en-fr, de-fr, ko-fr, es-fr, zh-ko, fr-de, ja-zh, de-es, es-zh, en-ja, zh-de, en-zh, zh-ja	undertranslation overtranslation	fr-en, ko-en, ja-en, es-en, zh-en, de-en
lexical-overlap	fr-en, en-fr, de-fr, ko-en, es-ja, ja-en, ko-fr, es-fr, ko-ja, de-ja, zh-en, ja-fr, zh-fr, en-ja, es-en, fr-ja, de-en, zh-ja	xnli-addition-contradiction xnli-addition-neutral xnli-omission-contradiction xnli-omission-neutral	fr-en, vi-en, sw-en, tr-en, zh-en, ru-en, bg-en, el-en, th-en, es-en, hi-en, de-en, ar-en, ur-en
hallucination-unit-conversion-amount-matches-ref hallucination-unit-conversion-unit-matches-ref	et-en, wo-en, da-en, no-en, uk-en, ta-en, fi-en, pl-en, ja-en, hy-en, ur-en, hr-en, fr-en, lt-en, tr-en, he-en, bg-en, ro-en, sv-en, ru-en, es-en, nl-en, zh-en, hu-en, be-en, lv-en, ko-en, ga-en, sk-en, af-en, sl-en, sr-en, ca-en, de-en, mr-en, id-en, vi-en, gl-en, pt-en, fa-en, hi-en, el-en, ar-en, it-en, cs-en	hallucination-date-time	en-de, et-en, ca-es, en-et, hr-lv, da-en, no-en, uk-en, fi-en, en-da, ta-en, pl-en, ja-en, en-hr, hy-en, ur-en, fr-en, hr-en, lt-en, sr-pt, en-sv, tr-en, en-no, en-sl, he-en, pl-sk, ru-en, ro-en, sv-en, en-lt, es-en, en-nl, nl-en, bg-en, he-sv, zh-en, hu-en, be-en, lv-hr, lv-en, bg-lt, en-ro, sk-pl, ko-en, ga-en, sk-en, af-en, sl-en, en-hu, sr-en, en-es, ca-en, en-sk, de-en, mr-en, id-en, vi-en, gl-en, en-fr, de-fr, pt-en, fr-de, en-pt, fa-en, hi-en, el-en, ar-en, it-en, en-pl, cs-en
commonsense-only-ref-ambiguous commonsense-src-and-ref-ambiguous	en-de, fr-en, ru-fr, en-fr, de-fr, ru-de, fr-de, ru-en, en-ru, fr-ru, de-ru, de-en	copy-source	ar-fr, ru-es, ur-en, fr-en, tr-en, zh-de, bg-en, ru-en, es-en, zh-en, sw-en, ja-ko, th-en, de-en, pl-mr, vi-en, hi-en, el-en, ar-en
addition omission	en-ca, en-el, en-et, en-ta, pl-en, hr-en, he-en, pl-sk, en-ar, ru-en, en-fi, zh-en, hu-en, be-en, lv-hr, en-he, ko-en, en-fa, sl-en, ca-en, en-gl, en-tr, en-sk, de-en, en-sr, fa-af, fa-en, ar-en, cs-en, en-de, en-hy, ar-hi, no-en, uk-en, fi-en, en-be, sr-pt, en-ru, sv-en, nl-en, sk-pl, en-hi, en-hu, mr-en, hi-ar, id-en, gl-en, en-fr, en-lv, fr-de, ca-es, en-uk,	addition omission	en-ur, en-hr, ur-en, en-no, en-sl, ro-en, en-vi, en-lt, es-en, en-nl, he-sv, en-it, en-ro, af-fa, en-id, lt-bg, en-af, af-en, es-ca, vi-en, sv-he, de-fr, pt-en, en-pl, et-en, hr-lv, wo-en, da-en, en-ko, en-da, ja-en, hy-en, pt-sr, hy-vi, fr-en, en-cs, lt-en, en-sv, tr-en, bg-en, lv-en, bg-lt, sr-en, en-es, en-bg, en-pt, hi-en, el-en, it-en

Table 11: Collection of list of languages per phenomena