THE UNIVERSITY *of* EDINBURGH

Segment-level Evaluation of Machine Translation Metrics

Nikita Moghe

Doctor of Philosophy

Institute for Language, Cognition and Computation
School of Informatics
University of Edinburgh

2024

Abstract

Most metrics evaluating Machine Translation (MT) claim their effectiveness by demonstrating their ability to distinguish the quality of different MT systems over a large corpus (system-level evaluation). However, their evaluation on determining good translations from bad for one instance (segment-level evaluation) is largely understudied and overlooked. Segment-level evaluation influences system-level evaluation and is crucial in applications that use MT as a part of their technology stack. In this thesis, we offer a new perspective on evaluating segment-level metrics through their use in extrinsic tasks and challenge sets allowing us to identify their drawbacks and subsequently provide suggestions to improve them. Our first approach evaluates a metric's ability to correlate translation quality with translation utility in an extrinsic task. We find that contemporary MT metrics exhibit negligible correlation with the outcomes of a downstream task indicating their inability to identify useful translations. We observe that the scores provided by neural metrics are not interpretable, in large part due to having undefined ranges. We further find that different tasks show varying sensitivity to MT errors. To assess the capability of individual metrics in identifying various machine translation errors, we create a contrastive challenge set. ACES consists of 68 phenomena ranging from simple perturbations at the word/character level to more complex errors based on discourse and real-world knowledge spanning 146 language pairs. We evaluate 47 metrics on the ACES dataset belonging to different design paradigms. We also investigate claims that Large Language Models (LLMs) are effective as MT evaluators, addressing the limitations of previous studies by providing a more holistic evaluation that covers a range of linguistic phenomena and language pairs and includes both low- and medium-resource languages. Our results demonstrate that different metric families struggle with different phenomena and that LLM-based methods fail to demonstrate reliable performance. We conduct several analyses and observe that many metrics ignore the information in the source sentence, have a tendency to prefer surface-level overlap and end up incorporating properties of base multilingual models which are not always beneficial. Throughout the thesis, it becomes evident that singular scores produced by metrics are uninformative. We provide several recommendations to improve metric design while advocating MT evaluation based on the prediction of error labels instead of error scores. To facilitate this, we also release SPAN-ACES where the incorrect translations from ACES are annotated at the span level.

To Dr. A.P.J. Abdul Kalam, Dr. Kalpana Chawla, and Amita Teacher

Lay Summary

When machine translation systems are deployed into the real world, they go through internal rigorous checks involving both human evaluation and computer-assisted evaluation. But, what if after deployment, we want to test the quality of that system for a specific sentence or paragraph? Such an evaluation is known as segment-level evaluation. We can ask a professional translator to rate that translation. However, scaling this evaluation is challenging. This resulted in the development of computer-assisted evaluation for translation where the resulting methods are termed MT metrics. But how do we know if this computer-assisted evaluation is up to the mark?

This thesis introduces new ways to evaluate MT evaluation. We introduce an evaluation method that that looks at identifying if the respective MT metrics can identify the usefulness of the translated content in a specific task. For example, if you are a French speaker talking to an English chatbot, we expect the chatbot to include an MT system that will translate from French to English. If the translation is of good quality, the chatbot can process your request but if it is of a poor quality, the chat bot will get confused. Typically, MT metrics offer a numerical score for predicting translation quality. We find that many of the current evaluation methods cannot effectively judge whether a translation is truly helpful for the end task. The scores provided by certain metrics are also found to be unclear and hard to interpret. For example, if a method specifies that a translation has a score of 63, it is hard to understand this score without context - which most metrics do not offer. We build a catalogue of common machine translation errors and test several old and new MT metrics on it to understand which errors can be handled by these metrics, akin to a doctor's report. Additionally, it allows us to get an idea of the good trends and incorrect design strategies in metric development.

The thesis concludes by suggesting ways to improve the design of these evaluation methods and recommends a shift towards evaluating translations based on marking the incorrect terms in the sentence rather than assigning scores.

Acknowledgements

Writing the acknowledgements section often feels like giving a speech at the Oscars, but it is so much more than that. First and foremost, I would like to express my sincere gratitude to my primary advisor, Alexandra Birch for her supervision, unending support, and constant encouragement. Her detail-oriented approach and her long-term vision for research amaze me and I am thankful to her for making me a better researcher. She has been kind and patient throughout the process even identifying when I needed external help. I thank Mark Steedman for being an excellent and hands-on second supervisor. His "And why is that" when explaining any results will stay with me for the rest of my career. This thesis belongs to both of them as much as it belongs to me.

I thank my examiners Barry Haddow and Carolina Scarton for an intense yet fun viva and thoughtful suggestions that have improved this thesis. I would like to thank Ivan Titov, Barry Haddow, and Edoardo Ponti for offering feedback during the annual reviews which has shaped several discussions in this work.

I would like to thank everyone involved in CDT in NLP and IGS, especially Sally Galloway for making bureaucracy easier and cheering for every small win. 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. We also thank Huawei for their support. I thank Adam Lopez for teaching DRNLP and offering insights in the hallway which contributed to my growth as a researcher.

This thesis is a result of multiple interesting collaborations with researchers within and outside the university. I express gratitude towards Korhonen, Arnisa Fazla, Chantal Amrhein, Evgeniia Razumovskaia, Christian Hardmeier, Ivan Vulić, Liane Guillou, Rachel Baweden, Rico Sennrich, Tom Kocmi, and Tom Sherborne. Special mention to Chantal, Genie, Liane, and Tom S for their rants and support during and after the projects. I would also like to thank the organisers and the participants of the WMT shared task for testing our benchmark. I thank the amazing people I met during the conferences who have greatly shaped my research intuitions. Special thanks to Raj Dabre for his uplifting chat during the final thesis push.

My internship experiences at PolyAI and Microsoft Semantic Machines enriched my research experience. I thank Ben Levin, Ivan Vulić, and everyone at PolyAI for making the virtual internship enjoyable. I thank you Harsh Jhamtani, Patric Xia, Jacob Andreas, and Jason Eisner for the amazing in-person summer internship at Microsoft.

With the privilege of being jointly supervised I received lots of tips, guidance, support, and lunches from two distinct research groups - StatMT: Arturo, Biao, Christos,

Denis, Guillem, Guilio, Gustavo, Jelmer, Jonas, Kadi, Kenneth, Laurie, Mateusz, Maxi, Nick B, Patrick, Proyag, Tsz Kin, Vivek; and Semantx: Elizabeth, Iona, Javad, Kasia P, Liane, Liang, Louis, Matt, Miloš, Nick M, Ratish, Sabine, Sander, Tianyi (Teddy) and Tianyang.

The CDT in NLP and the Forum in general provide a productive and entertaining environment. Special mention to the CDT-19 cohort for being the "best friends" I could ever ask for - Dan, Emelie, Faheem, Georgia, Henry, Irene, Jie, Laurie, Nina, Nicole, Parag, Rimvydas, Rohit, Ronald, and Tom H. Dealing with the pandemic was easier with them with the constant virtual ranting. I thank the remaining cohorts in the CDT and especially Agostina, Aida, Anna, Coleman, Eddie, Gautier, Matthias, Nick F, Nick S, Radi, Sandrine, Verna, and Zheng for their support. Moving on to the folks beyond my immediate research/CDT groups; thank you Andreas, Babita, Kasia S, Ola, Matus, Fady, Peter, Bogdan, Jonathan, Ruchika, Raman, Tarini, Ameer, and Viktor for helping me survive the PhD! Special mention to Kasia S and Ruchika for their support during the corrections period. I am grateful to my office mates Adarsh, Faheem, Ondrej, Resul, and Will for making it a fun workplace too. I thank Adarsh, Arushi, and Octave for offering love and support through food.

Outside of the Forum, Edinburgh blessed me with so many lovely people, especially Durga, Isha, and Kartik. My first flatmate Fahima made adjusting to Edinburgh easier. My flatmates in the ever changing Marchmont environment - Leila, Ayesha, Hannah, Fraser, Cam, and Anna C; thank you for a peaceful home. I thank Frank, Kate, and Mick for being very co-operative landlords. I am grateful to Dance Base and Edinburgh Leisure Centre for keeping me healthy. My Sitar teacher Alec Cooper and his wonderful set of people at The Sitar Project, especially Ioana helped me through their music.

I thank Balaraman Ravindran and Mitesh M. Khapra for introducing me and helping me take the first steps in the field of research. I express gratitude towards the amazing people who continue to support me across time zones and borders - Surbhi, Darshan, Ayesha, Tarun, Deepak, Shweta, Stanley, Kanika, Shikha, Yash, Priyesh, Rohan, Pavan, and Gandhalee. I am grateful to Dr. Vrushali Situt and Dr. Ankita Mishra for their timely support. I would like to thank my parents Vinay Moghe and Sangeeta Moghe, my grandmother Sunanda, and my sister Ankita Moghe for everything. Thank you, Kaustubh, Kyra, and the rest of my family for their constant encouragement.

There are a bunch of other people missing from this page who deserve a shoutout (and a dessert from me). Lastly, I thank all the strangers on the internet for their memes, tweets, blogs, and reddit posts that offered inspiration and joy when I needed the most.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*Nikita Moghe*)

Contents

Chapter 1

Introduction

Language technologies have made significant progress over the last decade. A substantial portion of these advancements has been driven by data-centric approaches. This has resulted in tremendous progress for a handful of languages where such data is available (high-resource) while leaving other languages behind (Joshi et al., 2020). To encourage the democratisation of language technologies, several applications incorporate Machine Translation (MT) models within their technical pipeline. MT systems are computer programs that convert information from one language into another language. Using MT models enables building applications that cater to a global and multilingual user base.

To achieve success in building multilingual technologies, it is crucial that the underlying MT systems consistently demonstrate *high quality* in their translations. This definition of high quality has varied over the years to accommodate the development and progress of MT systems. A popular paradigm in this evaluation of MT systems is *human evaluation* where the end users of MT systems or translation professionals are asked to perform a quantitative or qualitative evaluation of the generated outputs (White et al., 1994). For example, rating translations on a scale of 1 to 5 according to their grammatical accuracy.

The development of MT systems has relied on a combination of several movable parts responsible for a particular function to achieve the said translation. For example, phrase-based MT systems have phrase tables, re-ordering algorithms, and phrase-based decoders (Koehn et al., 2003). At the same time, modern deep learning-based MT systems produce different quality MT systems based on data, tokenisation, hyperparameters, encoding/decoding algorithms (Popel and Bojar, 2018). Performing human evaluation with every minor architectural change made the human evaluation process tedious, labour-intensive, and expensive. This led to the exploration of building automatic

methods that can emulate such human evaluation, termed MT metrics.

The development of MT metrics is guided by the breakthroughs in Natural Language Processing (NLP) research as well as the progress in building MT systems. An MT metric measures the quality of a translation given a source sentence and/or reference translation(s). The metrics which include reference in their setup are called reference-based metrics while the ones which do not use reference in their process are reference-free or quality estimation (QE) metrics. One of the earliest attempts at offering a proxy for human evaluation has been the BLEU metric (Papineni et al., 2002). It calculates a score based on the number of words or phrases (n-grams) that overlap between the translation and the reference sentence. As the quality of translations from MT systems improved, relying only on such word overlap was no longer sufficient (Callison-Burch et al., 2006). Then such metrics were used in combination with embedding-based metrics. These metrics calculated scores based on similarity in the geometric space based on the advancements in representation learning for NLP (Gupta et al., 2015; Lo, 2019). Simultaneously, the community has led several human evaluation campaigns for MT evaluation (Koehn and Monz (2006) to Freitag et al. (2023), Callison-Burch et al. (2012) to Blain et al. (2023)). With the availability of human annotations for MT evaluation, a parallel shift in MT metric development was to develop metrics that learn evaluation directly from these annotations (Song and Cohn, 2011; Rei et al., 2020). More recently, Large Language Models (LLMs) have been used for performing MT evaluation with a few demonstration examples (Kocmi and Federmann, 2023b).

With the surge of metrics for evaluating MT, there have been collective efforts to assess the effectiveness of these metrics as well. This assessment is conducted through two prominent directions - segment-level evaluation and system-level evaluation. In segment-level evaluation, these metrics evaluate a fixed number of translations from different MT systems. The predicted evaluations, typically scores in a pre-defined range, are then correlated against the human evaluation of these translations. This measures the effectiveness of the MT metric irrespective of the system that produced the translation. In system-level evaluation, the metrics are evaluated on their ability to distinguish if one MT system is collectively better at generating translations than the other. This is typically done by averaging the segment-level predictions over a corpus and then comparing these averages. Often, metrics are used to rank different MT systems (Koehn and Monz, 2006; Freitag et al., 2023).

Most of the new metrics claim their effectiveness by evaluating at the system-level. The methods claim state-of-the-art on standard benchmarks generally, the WMT shared

tasks (Koehn and Monz (2006) to Freitag et al. (2023)). The system-level evaluation only provides an overview of the metric's ability to rank MT systems without any information on its robustness to specific MT errors. There is no mechanism to verify the reliability of a metric's prediction for a randomly selected translation. Even with the current techniques for segment-level evaluation, the results are offered as an overview across language-pairs without any in-depth analyses about the metric's merits and shortcomings (Ma et al., 2019). Further, human evaluation which is the gold standard for the benchmarks is noisy as it is subjective, leading to low inter-annotator agreement (Popović, 2021).

We emphasise that segment-level evaluation is extremely important as it influences the system-level evaluation. It is crucial to access the quality of a translation in an online setting. For example, a chat bot deployed in a particular language needs to interpret a user's request in a different language by using automatic translation in real time. As MT systems are not perfect, it is crucial to determine if the current translation is suitable for continued processing or if it contains significant errors that could potentially compromise the overall user experience.

In this thesis, we use this setup to identify if a metric can correlate translation quality with translation utility in downstream tasks. This usability of translations is application dependent. For example, lower quality translations are helpful in crisis relief (Bansal, 2019) hence, should not be discarded while fluent translations are essential when translating responses from a dialogue system (Lin et al., 2021). We note that this setup inherently requires QE metrics as references are unavailable at test time. Yet, this thesis includes experiments with reference-based metrics due to their prevalence in the field and their foundational role in many QE metric designs. We also demonstrate alternate ways of using reference-based metrics in an online setting with round trip translation.

To understand if a metric can be reliable for specific MT errors, we need to detect its ability to predict such errors. To that end, we create a diagnostic dataset of 68 MT errors and benchmark 47 metrics. This type of meta-evaluation offers a fine-grained evaluation of every metric as well as an overview of trends in current metric design. Based on our observations, we make several recommendations for future metric development.

In this thesis, we offer a critical assessment of metrics performing segment-level evaluation through new perspectives. We look at techniques that do not involve human evaluation as well as provide an in-depth analysis of the strengths and the weaknesses of several MT metrics. We list the contributions of our thesis as follows -

1.1 Contributions

We highlight that the use of MT systems in downstream tasks is as important as using MT systems as standalone application. To that end, we identify if existing metrics can asses the quality of these intermediate translations. Specifically, we ask the following question -

Can metrics reliably identify the impact of translation quality on translation utility in downstream tasks?

We list our first contribution as follows:

Task-based evaluation of MT metrics: We propose a method for evaluating MT metrics by correlating their predictions of translation quality with the success of using these translations in downstream tasks. In our setup, we have access to a parallel task-oriented dataset, a task-specific monolingual model, and a translation model that can translate from the test language into the language of the monolingual model (task language). The examples from the test language are translated to the task language and passed through the task-specific model. When the translated example produces an incorrect prediction for the end task, we use a label as *breakdown* and *not breakdown* when the prediction for the task is accurate. We thus obtain a binary classification benchmark for breakdown detection. As this benchmark is solely dependent on the predictions of the downstream task, there is no need for human evaluation. We evaluate the metrics on their ability to correlate the quality of translation with respect to its utility in the downstream task. We evaluate nine different metrics for three extrinsic tasks: Semantic Parsing, Question Answering, and Dialogue State Tracking. We find that segment-level scores provided by all the metrics show negligible correlation with the success/failure outcomes of the end task across different language pairs. This outcome suggests that segment scores produced by these metrics are uninformative. We further find that different extrinsic tasks demonstrate varying levels of sensitivity to diverse MT errors.

The above exploration only covers three tasks and a handful of MT errors. MT systems deployed across diverse applications beyond academic benchmarks will likely produce varying errors as they are incorporated into newer tasks. Existing evaluation of MT metrics largely focuses on ranking MT systems based on their quality. However, there are limited efforts towards fine-grained evaluation.

How do we identify if metrics are robust to specific translation errors?

This brings us to our second contribution:

Diagnostic Dataset for MT evaluation: We construct a benchmark to evaluate individual metrics holistically. We curate ACES, a translation Accuracy ChallengE Set, consisting of 68 phenomena across 146 language pairs that influence the meaning of the translation. ACES is a contrastive challenge set consisting of a source sentence, a good translation, an incorrect translation containing an error corresponding to a specific phenomenon, a reference sentence, and the label for the phenomenon. These phenomena can include MT errors such as incorrect named entities, ambiguous translations, errors relating to discourse, and real-world knowledge. ACES measures the ability of a metric to distinguish a good translation from an incorrect translation. This extensive and diverse catalogue of errors makes a good benchmark for metric developers for identifying the properties of their metrics. We use ACES to evaluate a wide range of MT metrics including the submissions to the WMT 2022 and WMT 2023 metrics shared task leading to 47 metrics corresponding to different paradigms in metric development. These include metrics that rely on word level overlap, leveraging similarity in geometric space (also referred as neural metrics), and those metrics that learn to predict from historical human evaluation of MT systems. Additionally, we evaluate the emerging trend of using LLMs for MT evaluation with ACES.

After benchmarking these different metrics on ACES, we obtain a good overview of trends in metric design, leading us to the question -

What does holistic evaluation of 47 MT metrics tell us about the general trends in MT evaluation?

Meta-evaluation through ACES: We find different metric families exhibit different strengths, thus, leaving no single winning metric across all categories. While neural metrics tend to do better than metrics with word overlap, these metrics are brittle under some MT errors. We find that reference-based metrics tend to disregard information in source sentences and fail at ambiguous translations. Further, even neural metrics are considerably influenced by word overlap with the reference. As several metrics use multilingual embeddings to enhance MT evaluation, we find that certain properties of multilingual embeddings produce undesirable effects on MT evaluation. We further

find that using LLMs for MT evaluation is far from perfect as their evaluation on ACES is worse than surface-overlap metrics.

Based on our observations in a task-based evaluation setup and through the meta-evaluation on ACES benchmark, we make general recommendations for building segment-level metrics:

Recommendations: Our observations show that a single summary score is uninterpretable and uninformative. We recommend developing metrics that mark the erroneous text within the translation while simultaneously describing the type of error made in that text. There should be efforts to collect MT evaluation data with error labels following a common annotation scheme like Multi-dimensional Quality Metrics (MQM) (Lommel et al., 2014). To encourage the development of more metrics that produce error labels instead of scores, we develop SPAN-ACES by annotating the tokens in the incorrect translation of ACES that exhibit the corresponding phenomenon. The labels in SPAN-ACES are obtained automatically and through human annotation. We benchmark SPAN-ACES on recent methods that predict erroneous spans in translations. Their poor results highlight the need for further advancements in label-based evaluation for MT.

In terms of improved metric design, we recommend: combining metrics with different strengths, developing metrics that give more weight to the source and less to surface-level overlap with the reference, adding diverse references/translations during training of the metrics, and explicitly modelling additional language-specific information beyond what is available via multilingual embeddings.

1.2 Thesis Outline

This chapter introduces the readers to the goal of this thesis. We now list the contents of the remaining chapters of the thesis.

1. Chapter 2 provides the background required to understand the contents of this thesis. We discuss developments in human evaluation of MT followed by different design paradigms for automatic evaluation. We then explore the recent advancements in the meta-evaluation of machine translation.

2. Chapter 3 introduces a new setup to evaluate MT metrics on their ability to predict the utility of translations in downstream tasks. This setup is conducted on three

downstream tasks for nine metrics.

3. Chapter 4 discusses the curation of the ACES benchmark and the subsequent SPAN-ACES dataset. We discuss in detail the datasets used for the construction of these challenge sets. We describe automatic and semi-automatic methods for the creation and annotation of various challenge sets within those datasets.

4. Chapter 5 provides a comprehensive evaluation on ACES on the metrics submitted to the Metrics shared task at WMT 2022 and 2023. We experiment with LLMs for MT evaluation and provide baselines for SPAN-ACES. We conduct a comprehensive analysis of these results and report general trends in MT evaluation.

5. Chapter 6 concludes our findings and provides recommendations for MT evaluation based on the results in the previous chapter. We also suggest directions for future work.

The publications included in this thesis are (* denotes equal contribution):

- "Extrinsic Evaluation of Machine Translation Metrics"
  Nikita Moghe, Tom Sherborne, Mark Steedman, and Alexandra Birch. 2023. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 13060–13078, Toronto, Canada. Association for Computational Linguistics. (Moghe et al., 2023b) [Chapter 3]
  I led this work in all aspects - framing the breakdown detection task, designing experiments, and writing the paper. Tom contributed by providing the models for the parsing task and offered comments on the paper. Lexi and Mark provided supervision and feedback on the paper draft.

- "ACES: Translation Accuracy Challenge Sets for Evaluating Machine Translation Metrics." Chantal Amrhein*, Nikita Moghe*, and Liane Guillou*. 2022. In Proceedings of the Seventh Conference on Machine Translation (WMT), pages 479–513, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics. (Amrhein et al., 2022) [Chapters 4 and 5]
  All the authors contributed equally during brainstorming of different categories of challenge sets, designing analyses experiments, and writing the draft. Specifically, I focused on bringing insights from multilingual NLU literature to the development of challenge sets.

- "ACES: Translation Accuracy Challenge Sets at WMT 2023" Chantal Amrhein*, Nikita Moghe*, and Liane Guillou*. 2023. In Proceedings of the Eighth Conference on Machine Translation, pages 695–712, Singapore. Association for Computational Linguistics. (Amrhein et al., 2023) [Chapter 5]
  This paper required re-running analyses from the previous paper. Hence, contributions remain the same. While all authors contributed writing equally during submission, Liane and I substantially rewrote the camera-ready version based on the comments of the reviewers.

- "Machine Translation Meta Evaluation through Translation Accuracy Challenge Sets" Nikita Moghe, Arnisa Fazla, Chantal Amrhein, Tom Kocmi, Mark Steedman, Alexandra Birch, Rico Sennrich, and Liane Guillou. 2024. Under review. (Moghe et al., 2024) [Chapters 4 and 5]
  In addition to the contributions from the previous ACES works, I led the work on extending ACES for the journal submission. I worked on the section on LLMs - designing and running experiments using open-source LLMs. I helped in developing the manual annotation guidelines and inter-annotator evaluation with Arnisa, Liane, and Chantal. I contributed to the baselines for SPAN-ACES and developed evaluation methods. I have largely written and edited the draft of the journal version. Arnisa worked on automatic annotation in SPAN-ACES and developed two baselines for SPAN-ACES. I mentored Arnisa for the development of SPAN-ACES baselines and partially for automatic annotation. Chantal's contribution is the same as that of ACES and she mentored Arnisa for automatic annotation. Tom helped us with collecting data for manual annotation. He contributed to the GPT-based experiments and provided feedback on the draft. Lexi, Rico, and Mark provided supervision and comments on the paper draft. Liane's contribution is the same as that of ACES and she led the section on inter-annotation. She contributed to the writing of the paper.

The recommendations in the final chapter are based on observations of all the above publications.

These contributions emerged from a parallel line of work focusing on building multilingual dialogue systems with limited data. We looked at this problem by actively building datasets in a low-resource setting (Moghe et al., 2023a), developing dialogue-specific translation models (Moghe et al., 2020), and cross-lingual transfer learning through intermediate fine-tuning for dialogue tasks (Moghe et al., 2021). Another

direction in reducing data labelled data annotation was to use MT metrics to detect if a translation was "good enough" to sufficiently use as an alternate data sample or if human correction was required for the same. However, as seen later in the thesis, these scores are uninformative, which led to the development of this thesis.

Thus, the papers published in addition to the ones included in the thesis are as follows:

- "The University of Edinburgh-Uppsala University's Submission to the WMT 2020 Chat Translation Task." Nikita Moghe, Christian Hardmeier, and Rachel Bawden. 2020. In Proceedings of the Fifth Conference on Machine Translation, pages 473–478, Online. Association for Computational Linguistics. (Moghe et al., 2020)

- "Cross-lingual Intermediate Fine-tuning improves Dialogue State Tracking." Nikita Moghe, Mark Steedman, and Alexandra Birch. 2021. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 1137–1150, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics. (Moghe et al., 2021)

- " Multi3NLU++: A Multilingual, Multi-Intent, Multi-Domain Dataset for Natural Language Understanding in Task-Oriented Dialogue." Nikita Moghe*, Evgeniia Razumovskaia*, Liane Guillou, Ivan Vulić, Anna Korhonen, and Alexandra Birch. 2023.In Findings of the Association for Computational Linguistics: ACL 2023, pages 3732–3755, Toronto, Canada. Association for Computational Linguistics. (Moghe et al., 2023a)

Chapter 2

Background

Evaluating translation quality has been of interest ever since the development of machine translation systems. This evaluation generally measures how good is the provided translation with respect to the source and/or reference translations(s) and whether such output is usable in its current form. Conducting regular evaluation allows developers to track progress and improvements in MT systems over time. It enables comparison between different MT systems or versions, helping users and developers choose the most effective solution. With fine-grained calibration of errors made by the system, it can highlight specific weaknesses of that system, guiding future development efforts. After deploying systems that surpass a certain quality standard, the user's trust in the system is also improved encouraging wider adoption.

As there is no agreed definition of translation quality, this has resulted in development of multiple concurrent measures of perceived quality of the translation. For example, developing a method that provides a notion of quality produced by the system on an average or creating a list of errors that are specific to a given task/process. Lommel and Melby (2018) propose the following typology of translation quality metrics suiting different purposes when using machine translation systems. These metrics can be automatic, manually evaluated, or a combination of different techniques. The categories in this typology are Holistic vs. Analytic, Fine-grained vs. Coarse, Objective vs. Subjective metrics, and Reference-based vs. Reference-free.

The holistic metrics provide a single result by examining the entire translated text. For example, you might see a statement like, "This translation has a score of 96.5". In contrast, analytic metrics look at several dimensions of qualities and provide a decomposition of a single score. For example, that same score of 96.5 can be a weighted score of individual scores of accuracy, fluency, and style. Fine-grained metrics provide

a further breakdown down of the individual categories like accuracy can be decomposed as a score of mistranslation/did not translate *etc*. Coarse metrics, in contrast, offer a singular rating for high-level aspects such as accuracy or fluency.

Objective metrics are based on observable facts that are not influenced by individual subjectivity, making them more data-driven. In contrast, subjective metrics depend on personal reactions and individual preferences, often influenced by taste or non-objective factors.

Reference-based metrics involve comparing translations against one or more "gold standard" reference texts with or without the source text. Reference-free metrics on the other hand determine the quality of the generated translation based on the source alone.

The selection of various categories within this typology is often determined by the intended uses of that translation. In the rest of this thesis, we will discuss metrics that only provide holistic scores except for a few metrics that provide fine-grained evaluation.

In this chapter, we will discuss traditional and contemporary campaigns for evaluating MT quality. We will provide an overview of the evaluation of these diverse metrics, which constitute the core focus of this thesis.[1]

2.1 Human Evaluation of Machine Translation

Human evaluation of machine translation is a hard problem as there are multiple correct translations and the judgement of *correctness* of that translation is dependent on its use. Throughout the years, various attempts have been made to simplify this problem. The earliest evidence of large-scale human evaluation of MT systems is in Carroll (1966) where a nine-point scale was used to measure intelligibility and informativeness of the presented translation. Every point on the scale had a specific definition. Moving to the research in the 21st century, the annual Conference on Machine Translation (WMT) has been one of the sources to document different strategies of human evaluation through their shared tasks. The work in Koehn and Monz (2006) suggested a five-point scale for adequacy - whether the meaning is preserved and fluency, whether the translation is grammatically correct.

For several years(Callison-Burch et al. (2007) to Bojar et al. (2016b)), the default human evaluation was termed as "Relative Ranking". It involved obtaining translations

[1]We thank Nitika Mathur for writing a comprehensive background chapter in her thesis (Mathur, 2021).

for the same example from different MT systems which were then ranked among themselves while allowing for ties. However, this method constantly displayed low inter-annotator agreement (Bojar et al., 2016b). It was constrained to the randomly selected systems during the ranking, allowing unfair advantage to certain systems (Lopez, 2012) and lacking a universal score.

The more recent human evaluation strategies namely Direct Assessment and Expert Annotation using Multidimensional Quality Metrics offer some solutions to the development of more unified human evaluation. We discuss these in detail below-

2.1.1 Direct Assessment

Graham et al. (2013) proposed using the continuous scale of 0-100 for evaluating machine translations which is commonly referred to as Direct Assessment (DA). The annotator is shown a sliding scale to mark the respective translations. These annotators can be workers from popular crowdsourcing platforms, or the participants in the WMT shared tasks, and as the annotators may develop their internal scale while scoring the translations, the general practice is to (i) standardise the scores for every worker by subtracting the mean and dividing by the standard deviation and (ii) use multiple annotators per sentence (Graham et al., 2015). These measures are employed to reduce the noise in the human annotation. This method has been the official evaluation method since the 2017 shared task (Bojar et al., 2017).

The advantage of this method is that it scales linearly with the number of MT systems submitted to the shared task as this evaluation is performed independently for every system. This is unlike relative ranking which scales quadratically and is only dependent on the submissions of that year. Further, these DA scores can be converted into Relative Ranking by ranking the outputs from different systems on the value of the respective DA scores.

Typically, adequacy (whether the translation preserves the meaning) is scored by showing the translation and a ground truth reference. Fluency is measured by only showing the generated translations. Both the raw scores from the annotators and the corresponding z-scores are made available when releasing the data. The method to obtain DA scores has been continuously updated throughout the years to accommodate changes. For example, Ma et al. (2019) included additional context (wherever available) to ensure that the generated translation was coherent with the current context (Läubli et al., 2018). For the WMT 2021 task (Akhbardeh et al., 2021) included a second round

of evaluation to focus on higher quality evaluation of the top scoring systems.

2.1.2 Multidimensional Quality Metrics

While direct assessment scores offer a solution, these scores do not provide any fine-grained information about the type of error that is being made in the translation. With internally calibrated scales per annotator, it is hard to interpret the significance of the score obtained for a particular translation.

To overcome these drawbacks, Lommel et al. (2014) proposed Multi-dimensional Quality Metrics (MQM) framework which consists of a hierarchy of MT errors. The MQM framework is inspired from several quality assurance systems based in the industry as well as evaluation practices of expert human translators. The MQM typology can be adequately used for evaluating machine generated and/or human written translation. As there is a hierarchy of errors, different errors can be weighted according to their intended use. One of popular weighting schemes is the use of "minor" ($w_{minor} = 1$), "major" ($w_{major} = 5$), and "critical" ($w_{critical} = 10$) errors. Minor errors do not impact the usability or meaning of the translation while translation with major errors deviates from the source meaning but it does not impact usability. Translation with critical errors severely impact the usability of the text.

There has been a resurgence in using MQM annotation for human evaluation of MT outputs most notably in the work driven by Freitag et al. (2021a). This work collects the then largest amount of MQM annotated labelled data for the two language pairs English $\rightarrow$ German and Chinese $\rightarrow$ English. Typically, the error(s) in the translation are marked on a span-level (contiguous chunks of text) with their coarse-grained error category (example: Accuracy) and the corresponding fine-grained category (example: Omission). Beyond the availability of data, this work first shows that ranking based on MQM evaluation is superior than the DA style evaluation.

This work adapted the MQM definitions more suited to the automatic translation task which is illustrated in Table 2.1. While the categories in this figure are not exhaustive, they give a general idea of the MQM hierarchy. Note, we mostly use the errors under the accuracy category in the rest of the thesis. The presence of this well-defined hierarchy helps in designing better metrics as explained in the subsequent chapters.

WMT has been adopting MQM annotation for human evaluation since 2021 (Freitag et al., 2021b). Beyond WMT campaigns, Rei et al. (2020) used a proprietary MQM data for European languages, Sai B et al. (2023) carried out an MQM annotation

Error Category		Description
Accuracy	Addition	Translation includes information not present in the source.
	Omission	Translation is missing content from the source.
	Mistranslation	Translation does not accurately represent the source.
	Untranslated text	Source text has been left untranslated.
Fluency	Punctuation	Incorrect punctuation (for locale or style).
	Spelling	Incorrect spelling or capitalization.
	Grammar	Problems with grammar, other than orthography.
	Register	Wrong grammatical register (eg, inappropriately informal pronouns).
	Inconsistency	Internal inconsistency (not related to terminology).
	Character encoding	Characters are garbled due to incorrect encoding.
Terminology	Inappropriate for context	Terminology is non-standard or does not fit context.
	Inconsistent use	Terminology is used inconsistently.
Style	Awkward	Translation has stylistic problems.
Locale convention	Address format	Wrong format for addresses.
	Currency format	Wrong format for currency.
	Date format	Wrong format for dates.
	Name format	Wrong format for names.
	Telephone format	Wrong format for telephone numbers.
	Time format	Wrong format for time expressions.
Other		Any other issues.
Source error		An error in the source.
Non-translation		Impossible to reliably characterize the 5 most severe errors.

Table 2.1: MQM hierarchy as illustrated in Freitag et al. (2021a). There is a top-level error category followed by fine-grained errors within that category. The hierarchy is borrowed from Lommel et al. (2014).

campaign for Indian languages which was made publicly available. Similarly, Wang et al. (2023) offered MQM annotations for typologically diverse African languages. These works indicate the growing adoption of using MQM for human evaluation. The recommendations based on the observations in this thesis advocate the use of MQM for future evaluation.

2.1.3 Criticism

We briefly mention the prevalent drawbacks of existing methods for human evaluation. Firstly, human evaluation is inherently noisy due to subjectivity, level of expertise of annotators, and varying guidelines (Fomicheva and Specia, 2019; Freitag et al., 2021a). The inter-annotator agreement is found to be low for both relative ranking and direct assessments (Koehn and Monz, 2006; Castilho, 2020; Popović, 2021). The direct

assessment scores are only reliable if multiple judgements for the same translation are available (Ma et al., 2019).

The majority of the human annotation data is made available through shared tasks. Thus, the choice of language pairs in the dataset is dependent on the language pairs present in the translation task of that year. This also restricts the evaluation to the domain of the shared task (eg., news) as well as systems submitted to the shared task.

2.1.4 Task-based evaluation

The discussion so far has looked at measuring translation quality intrinsically. To overcome the subjectivity in human evaluation, several works have replaced measuring MT quality with objective end tasks. Doyon et al. (1999) detailed the notion of task-based evaluation using tasks such as summarisation, data filtering, *etc* based on the translated output. We refer the reader to (Scarton, 2016) while providing a high level overview of different tasks used for MT evaluation.

Reading comprehension tests form a good setup for task-based evaluation - if the generated text is of poor quality, the participants cannot infer the correct answer from the translated passage. Tomita et al. (1993) used passages from the Test of English as a Foreign Language (TOEFL) exams, where the questions were manually translated into the target language (Japanese) and the participants were asked to answer using the translated output. Laoudi et al. (2006) evaluate the translation quality on the task of event extraction by asking human annotators wh-questions about the translated text. Berka et al. (2011) used a binary questions on passages across various domains to rank different MT systems. More recently, Forcada et al. (2018) formulated MT evaluation as a cloze test and reading comprehension, by asking the human annotators to fill in the gap or answer the question based on the information from the translated output. Similarly, Han et al. (2022) evaluate simultaneous translation models by presenting the participants with a question that gets translated word-by-word into the target language and the participants are asked to answer the questions immediately.

Eye-tracking experiments are useful in determining the quality of the generated translations (Doherty and O'Brien, 2009; Stymne et al., 2012; Colman et al., 2022) . The participants are provided with the translated passages and their reading patterns such as gaze time, fixation count and average fixation duration are measured. Essentially, a translation is considered to be of poor quality if the participant had a higher gaze time.

Another popular alternative, especially in the quality estimation literature is post-editing time (Tatsumi, 2009; Temnikova, 2010; Scarton et al., 2019; Gladkoff and Han, 2022). These works typically measure the time taken by a human annotator to correct the translation. The translations that require more time for the process of post-editing are regarded as less favourable compared to the translations that can be swiftly rectified.

In addition to the above tasks, more critical tasks like software installation (Castilho et al., 2014), technical support (Del Gaudio et al., 2015) or cross-lingual retrieval (Zhang et al., 2022) in e-commerce engines are also used for MT evaluation.

Note, that this literature attributes the use of tasks to judge the quality of individual MT systems. Further, these tasks are generally solved by human participants and not automatic systems. The use of downstream tasks to evaluate the quality of MT metrics is a first of its kind and is discussed in Chapter 3.

2.2 Automatic Evaluation of Machine Translation

As seen in the previous section, the collection of human judgements is not trivial. It is tedious, time-consuming, and expensive. This has given rise to several designs for automatic evaluation. We largely focus on the metrics discussed in the rest of the thesis in the following subsections. [2] We provide brief descriptions of these metrics and encourage the reader to look into their respective papers for further details. We divide these metrics based on their design paradigms. Note that this distinction is not rigid and some metrics may belong to multiple categories. We include a subsection on the background of quality estimation (QE) systems at the end of the section.

2.2.1 Surface-level overlap

The earliest automatic metrics relied on comparing the generated translation with one or multiple ground truth references only based on lexical overlap. The motivation to develop such metrics was to obtain a method that was quick, language-agnostic and correlated with then human judgement (Papineni et al., 2002). These metrics are reference-based and have dominated MT evaluation. A few examples of such metrics are as follows:

- BLEU (Papineni et al., 2002) is one of the earliest surface-level compares the token-level n-grams of the hypothesis with the reference translation and then

[2] See Lee et al. (2023) for a comprehensive survey of MT metrics.

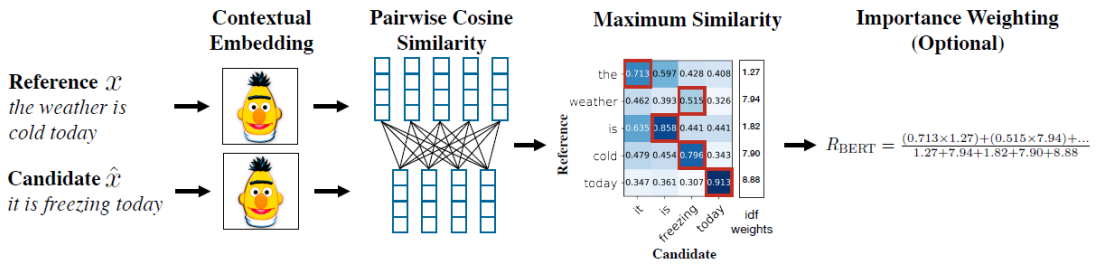

Figure 2.1: Example of translation score calculation in BERTSCORE. Both the candidate and the reference translation are encoded with a pretrained model. Then, a pairwise cosine similarity between every token is conducted. The final score is F1 calculated with maximally matching reference tokens per candidate tokens. Additionally, rare words can be given a separate weight through importance weighting. The image is credited to Zhang et al. (2020)

obtains a precision score weighted by a brevity penalty.

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

- SPBLEU (Goyal et al., 2022) is BLEU score over text tokenised with a language-agnostic subword model (Sennrich et al., 2016). 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.

- TOKENGRAM_F (Dreano et al., 2023b) is similar to CHRF but it uses token-grams obtained from contemporary tokenization algorithms (like subwords tokenization) to capture similarities between words sharing the same semantic roots.

2.2.2 Embedding similarity

As machine translation systems improved their diversity of words within translations, these translations were lexically different yet semantically similar to their source sentence and/or reference. To score such semantically similar sentences, metrics resorted to incorporating deep representations obtained from then state-of-the-art multilingual pre-trained models. Metric designs within this paradigm have looked at different ways of obtaining such representations and alternatives for calculating the similarity between them. Generally, a multilingual encoder is used to encode the translation, source/reference, and cosine similarity between the two representations is calculated as a score. We now discuss some examples.

- 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.

- EBLEU (ElNokrashy and Kocmi, 2023) operates on similar principal as BLEU. Instead of matching lexical overlap between n-grams, it calculates embedding similarities between the n-grams.

- BERTSCORE (Zhang et al., 2020) uses contextual embeddings from pre-trained language models like multilingual BERT (Devlin et al., 2019) 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. See Figure 2.1 for an illustration.

- HWTSC-TEACHER-SIM (Liu et al., 2022) uses sentence level representations from a knowledge-distilled multilingual model (Reimers and Gurevych, 2019) to obtain cosine similarity between the source and the translation.

- EMBED_LLAMA (Dreano et al., 2023a) computes cosine distance between embeddings of hypothesis and reference sentence that are derived from lower layers of an LLM, notably the LLaMA-2 model (Touvron et al., 2023).

- HWTSC-TLM (Liu et al., 2022) measures the quality of a translation (without source or reference) by averaging the probability of every token in the translation under a pre-trained model like XLM-R (Conneau et al., 2020).

2.2.3 Learning from human evaluation data

As seen from Section 2.1, various datasets on human evaluations of MT systems have been collected either through shared tasks or independently. Some metrics have leveraged these human annotations as training data to build metrics that mimic such human judgment. These metrics are neural and also use similarity from learned representations to predict a score.

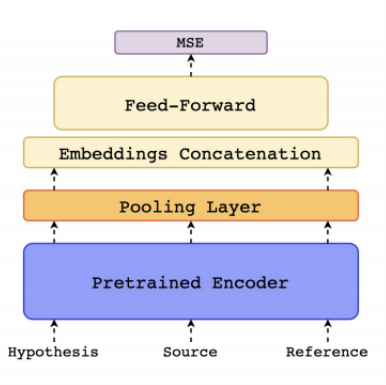

Figure 2.2: COMET framework as illustrated in Rei et al. (2020). It encodes the hypothesis, source, and reference using a cross-lingual encoder. The respective representations are concatenated to predict a score. The training data for these models is generally the WMT human evaluation data.

2.2.3.1 Estimator Models

These models are regression-based models that encode the source and the translation occasionally with the reference using a deep multilingual encoder followed by a feedforward layer that predicts a score. RUSE proposed by Shimanaka et al. (2018), DPMF-COMB (where comb is short for combined) developed by Yu et al. (2015) form some of the earlier metrics based on this paradigm. [3]

We first discuss the COMET framework which forms the basis of several metrics discussed in this thesis.

COMET Framework: Cross-lingual Optimized Metric for Evaluation of Translation (COMET) (Rei et al., 2020) is a machine translation evaluation metric that has been trained on historical human evaluation of machine translation. As seen before, this historical human evaluation data consists of a source, reference, a hypothesis (translation), and the corresponding "direct assessment" score assigned to the quality of the translation which is between 0-100 (Graham et al., 2015).

A cross-lingual encoder like XLM-R (Conneau et al., 2020) is used to encode the source, the hypothesis, and a reference (if any). Unlike most applications that use the representations from the last layer of the cross-lingual encoder, COMET uses a layer-wise attention mechanism to select the most important token representations for every token in the input sentence. The final sentence embedding for the sentence is obtained by applying average pooling over these token embeddings. The respective sentences for the source, hypothesis, and reference are concatenated and projected

[3] We borrow the term "estimator models" from Predictor-Estimator model in Kim et al. (2017)

into a single embedding. The model is trained to produce scores (z-scores) with a feedforward network trained with mean squared error loss. See Figure 2.2 for an illustration of the architecture. Typically, the QE variant of the metric is obtained by excluding the reference encoder in the architecture. COMET-QE variants are referred as COMETKIWI (Rei et al., 2022c).

This was the design of the first COMET model. Over the years (Rei et al., 2020, 2022a; Guerreiro et al., 2023), it has evolved into different variants - COMET-$*$-DA generally uses direct assessment data up to that year as its metric training data. [4] COMET-$*$-MQM continues fine-tuning of COMET-$*$-DA with MQM judgements. To convert the MQM spans into scores, a weighted sum of errors is inverted and converted to z-scores. COMET-22 (Rei et al., 2022a) and XCOMET Guerreiro et al. (2023) follow a multi-task setup where the metric tries to predict a segment-level score as well as span(s) in the sentence where the error is likely to have occurred.

The COMET architecture has inspired several other metrics:

- MS-COMET-22(Kocmi et al., 2022) uses the COMET architecture but with larger human judgement corpora that are carefully filtered to reduce noisy human judgement.
- METRIC-X(Juraska et al., 2023) uses mt5 (Xue et al., 2021) encoder-decoder language model as its base model. In addition to the DA and MQM data, the model uses synthetically generated evaluation data for training the metric.
- COMETOID (Gowda et al., 2023) distills the scores from a reference-based COMET model to create a student model that is reference-free. It uses InfoXLM (Chi et al., 2021) as its base model.
- XL-SIM (Mukherjee and Shrivastava, 2023) uses sentence embedding from sentence encoders directly instead of pooling. It compares cosine-similarity and z-score of the DA score during the training of the metrics.
- CROSS-QE (Liu et al., 2022) uses COMET-QE architecture with Monte Carlo dropout (Gal and Ghahramani, 2016) to simulate data augmentation.

Except COMETOID and CROSS-QE, all other metrics have their equivalent reference-free variants. In addition to the above metrics, we discuss a few other metrics discussed in the thesis that follow the estimator design paradigm

[4] $*$ indicates the year in which the metric was trained.

- BLEURT20 (Sellam et al., 2020b) 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.

- UniTE (Wan et al., 2022), Unified Translation Evaluation, is another neural translation metric. Instead of having separate architectures for reference-based and reference-free metrics, it proposes a multi-task setup for the three strategies of evaluation: source-hypothesis, source-hypothesis-reference, and reference-hypothesis in a single model. The pre-training stage is similar to that of COMET-DA models with additional synthetic data constructed using a subset of WMT evaluation data. Further, their fine-tuning uses novel attention mechanisms and aggregate loss functions to facilitate the multi-task setup. The unified architecture for the three setups inspired the design for the XCOMET metric.

- KG-BERTScore (Liu et al., 2022) is a reference-free machine translation evaluation metric, which incorporates a multilingual knowledge graph into BERTScore (and more recently COMET-QE) by linearly combining the results of BERTScore and bilingual named entity matching.

2.2.3.2 LLMs as evaluators

The above metrics require a large amount of training data and their effectiveness is proportional to the amount of training data (See Section 5.4.4). With the success of Large Language Models across multiple tasks, the MT community has recently explored the feasibility of using LLMs for MT evaluation. Due to their extensive pre-training, evaluation using LLMs uses *in-context learning* (Dong et al., 2023). Here, an LLM is given a set of instructions about MT evaluation (zero-shot) and a few demonstration examples (few-shot) explaining MT evaluation. The instructions can include predicting a translation score or error labels or both.

GEMBA-DA (Kocmi and Federmann, 2023b) first experimented use of LLMs for evaluation by performing zero-shot prompting methods using GPT models (Brown et al., 2020). The instruction only contained information about the scale for Direct Assesment and the LLM was asked to produce the corresponding score. Their method achieved state-of-the-art performance on system-level for the WMT22 metrics test set (Freitag et al., 2022) prompting further research in this direction.

EAPROMPT, which was proposed by Lu et al. (2023), instructed the model to produce an MQM like error report to simulate the chain-of-thought prompting (Wei et al., 2022) in the few-shot method with language-specific examples. The model was then instructed to calculate a score based on the predicted errors using the major-minor-critical weighting scheme. Kocmi and Federmann (2023a) improved on the multi-step approach proposed by the previous method by developing GEMBA-MQM. The difference is that the model is prompted only once and contains demonstration examples that are language-agnostic. See Figure 2.3 for an example prompt.

AUTOMQM (Fernandes et al., 2023) also prompts LLMs for MQM like evaluation. Their work includes obtaining MQM like data from LLMs and using that data in addition to human annotated MQM data to fine-tune LLMs for the task of metric evaluation.

```
(System) You are an annotator for the quality of machine translation. Your task is to identify
errors and assess the quality of the translation.

(user) {source_language} source:\n
```{source_segment}```\n
{target_language} translation:\n
```{target_segment}```\n
\n
Based on the source segment and machine translation surrounded with triple backticks, identify
error types in the translation and classify them. The categories of errors are: accuracy
(addition, mistranslation, omission, untranslated text), fluency (character encoding, grammar,
inconsistency, punctuation, register, spelling),
style (awkward), terminology (inappropriate  for context, inconsistent use), non-translation,
other, or no-error.\n
Each error is classified as one of three categories: critical, major, and minor.
Critical errors inhibit comprehension of the text. Major errors disrupt the flow, but what
the text is trying to say is still understandable. Minor errors are technically errors,
but do not disrupt the flow or hinder comprehension.

(assistant) {observed error classes}
```

Figure 2.3: The general prompt for GEMBA-MQM as described in Kocmi and Federmann (2023a). The LLM is asked to produce an MQM-based list of errors for the given test example.

2.2.4 Quality Estimation

QE systems involve prediction of the quality of the translation without relying on references. Use of QE systems is central to the work in the next chapter. As seen later in the thesis, we advocate evaluating translations by marking their erroneous spans in the hypothesis - a recommendation that is inspired from the QE literature. The

previous sections only look at the different architectures for popular QE metrics. We now highlight some of the crucial developments in this field.

The earliest approaches to building QE systems treated the problem as confidence estimation at the word level (Gandrabur and Foster, 2003; Ueffing and Ney, 2005), the phrase (Patel and M, 2016; Logacheva et al., 2016), sentence level (Blatz et al., 2004; Ive et al., 2018) and/or a document level (Soricut and Echihabi, 2010; Scarton, 2016). This confidence estimate is the measure of how likely was the MT system at generating that particular textual unit for the given source without using any references. Such systems often rely on using probabilities from the MT systems or independent components such as hand-crafted features, n-best lists, neural networks, and so on.

These systems generate confidence estimates in two main forms: numerical scores and categorical labels which can then be used for post-editing. The scores typically include likert scores (Mehandru et al., 2023), continuous scores in fixed ranges (Fonseca et al., 2019), or minimum number of edits to fix a translation (Specia and Farzindar, 2010). The categorical labels can involve predicting OK/BAD per textual unit, mostly at the word and phrase level (Blatz et al., 2004; Ueffing and Ney, 2005). Since 2021, the WMT QE shared task has extended this label scheme to a critical error detection task where the models label the severity of the error present in the erroneous span (Specia et al. (2021) to Blain et al. (2023). We make a similar recommendation based on the observations later in this thesis.

We refer the reader to Zhao et al. (2024) for an excellent survey on quality estimation methods.

2.3 Meta evaluation

The discussion in the above sections only pertains to the development of individual MT metrics. We shall now look at the different studies for evaluating these metrics.

2.3.1 Shared Tasks

The most prominent way for evaluating metrics and establishing their effectiveness is by comparing them with a set of human judgements (Papineni et al., 2002). These human judgements are generally obtained through the methods discussed in Section 2.1, despite their limitations. Typically, new metrics report the agreement with human judgment in their description papers or participate in shared tasks to establish their

usefulness. The metrics shared tasks (Specia et al., 2018; Zerva et al., 2022; Agarwal et al., 2023) and more notably the WMT metrics shared tasks have been instrumental in developing datasets and techniques for evaluating new MT metrics (Koehn and Monz (2006) to Freitag et al. (2023)). The two most prominent strategies for meta-evaluation are system-level evaluation and segment-level evaluation.

System-level Evaluation

In this type of evaluation, the metric is judged on its ability to determine the overall quality of an MT system for a given language pair. The determination of this overall quality is generally an average of the scores predicted by the metric for every example in the given test set that is correlated with human judgement. Pearson's correlation remains a popular choice for computing such correlation (Macháček and Bojar, 2014). However, it is highly sensitive to the outlier MT systems as discussed in Mathur et al. (2020a). This prompted the organisers to use a Pairwise ranking of systems (Kocmi et al., 2021) which determines if a metric's binary judgements indicating that one machine translation system is superior to the other resemble that of the human preference. Today, system-level evaluation is more associated with the ability of a metric to rank different MT systems.

Segment-level Evaluation

The purpose of this evaluation is to identify if a metric can evaluate the quality of a translation for individual examples irrespective of the systems that produced the translations. The term segment can include any textual unit - paragraph, document, sentence, *etc*. In most of this thesis, the term segment is used to indicate a sentence unless stated otherwise.

Segment-level evaluation was introduced in the 2008 WMT metrics shared task (Callison-Burch et al., 2008). When relative ranking was a popular way for evaluation, the metrics were evaluated based on whether they produced the same ranking using Kendall tau-like correlation. [5]. The concordant contained the set of comparisons where the ranking within the metric scores and the human judgements was the same while discordant contained the set of disagreements. Ties were excluded from the calculation. See Equation (2.1) for the calculation. While Pearson correlation has been used in some shared tasks (Bojar et al., 2016b), Kendall tau-like tau remains popular across recent

[5]The original definition of Kendall tau involves total ordering when calculating the concordant or the discordant. Most organisers rely on DA scores to obtain relative ranking

shared tasks. DA scores are converted to relative ranking while computing the Kendall tau.

$$\tau = \frac{concordant - discordant}{concordant + discordant} \tag{2.1}$$

In addition to the human evaluation and ranking of different metrics, these shared tasks have documented trends in MT evaluation. For example, Stanojević et al. (2015) highlighted the effectiveness of neural embedding-based metrics; Ma et al. (2019) show that metrics struggle on segment-level performance despite achieving impressive system-level correlation; Mathur et al. (2020b) investigated how different metrics behave under different domains; Freitag et al. (2022) advocate use of neural metrics over BLEU and so on.

While these trends are indeed quite important, most of the time the analysis in these papers is focused on system-level evaluation. Even within the system-level evaluation, the emphasis is on identifying a set of best-performing metrics which is obtained by a weighted average across different language pairs. This results in a singular score of metric performance, thus offering limited insights. Further, the ranking is likely to change depending on the method of correlation (Mathur et al., 2020a). Several works report segment-level evaluation tables and yet only provide observations about how the correlation is lower than system-level evaluation (Ma et al., 2019). There are neither explanations nor focused analyses explaining these lower scores.

To remedy this, since 2021, WMT has been conducting the challenge sets subtask that encourages researchers to build test sets to evaluate MT metrics on broader MT errors. Concurrently, the QE shared tasks (Specia et al., 2021; Zerva et al., 2022) introduced critical error detection task/catastrophic error task where the metric is evaluated on its ability to detect whether a translation contains a critical error like hallucination or mistranslation. The next two sections look deeper into the strengths and weaknesses of contemporary metrics beyond comparing them with human judgement.

2.3.2 Criticism of selected metrics

The shared task overview papers offer a summary of the current trends in metric development. Beyond these works, several studies have examined the drawbacks of individual metrics or a set of metrics.

Of these, criticisms of BLEU have been widely studied. For example, Callison-Burch et al. (2006) showed that under certain conditions this metric is neither a sufficient indicator of improved quality of machine translation nor does a higher BLEU score indicates better agreement with human judgement. While judging BLEU in a leader board setting, Mathur et al. (2020a) noted that the observed improvement in correlation with human judgment was an artefact of the setup. It is not robust to outlier MT systems. Kocmi et al. (2021) provided some evidence that solely relying on BLEU score can lead to the selection of lower-quality machine MT systems. Marie et al. (2021) criticise the reporting of MT metrics in academic research and propose guidelines to improve the credibility of the reported evaluation.

The recent surge in the development and adoption of neural metrics within the field has prompted researchers to closely examine their limitations. Amrhein and Sennrich (2022) show that COMET metrics are not robust against hallucinations containing named entities and numbers. BERTSCORE fails at detecting incorrect translations with high lexical overlap and stylistic overlap (Hanna and Bojar, 2021) while Sun et al. (2022) discover that societal biases in neural models propagate in such metrics. Yan et al. (2023) demonstrate that BLEURT and BARTSCORE (Yuan et al., 2021) are not robust to universal translations - generic sentences used as translations that produce high metric score even if the reference sentence has a completely different meaning.

2.3.3 Challenge sets

Challenge sets aim to provide insights on whether state-of-the-art models are robust to domain shifts, or simple textual perturbations, whether they have some understanding of linguistic phenomena such as negation/commonsense or simply rely on shallow heuristics, etc.

The WMT 2021 metrics task evaluated the metrics submitted to that year's task on three challenge sets: negation and sentiment polarity, corrupted references where the corruption included number swapping, antonyms, *etc*, and a linguistically motivated challenge suite developed in Macketanz et al. (2018a). They found that most metrics struggled at detecting corrupted Subordination, Named Entities and Terminology.

Since 2022, the construction of challenge sets has been included as a subtask with the metrics task. The datasets that were submitted to the WMT 2022 challenge sets shared task (Freitag et al., 2022) include: ACES (discussed extensively in this thesis) SMAUG (Alves et al., 2022), the HWTSC challenge set (Chen et al., 2022), and the

	Examples	Language Pairs	Phenomena	Categories
SMAUG	632	2	5	5
HWTSC	721	1	5	5
DEMETR	31,000	10	35	3
DFKI	21,000	2	100	14
MLSC	77388	3	1	1
ACES	36,476	146	68	10

Table 2.2: Comparison of challenge sets for MT metric evaluation in terms of Examples, Language pairs, Phenomena, and Categories. ACES is comprehensive across all the categories offering a holistic meta-evaluation.

DFKI challenge set (Avramidis and Macketanz, 2022). Both SMAUG and HWTSC are relatively small datasets ($<$1000 examples) focusing on a small set of five phenomena, each pertaining to a single category of critical error for meaning change. In comparison, the DFKI challenge set is much larger – it contains 19,347 examples and covers over 100 linguistically motivated phenomena, which are organised into 14 categories. These datasets provide an in-depth focus on specific high-resource language pairs: SMAUG (pt↔en and es→en), DFKI (de↔en), and HWTSC (zh↔en).

Independent of the shared task, DEMETR (Karpinska et al., 2022), which comprises 31K English examples translated from ten languages, was developed for evaluating MT metric sensitivity to a range of 35 different types of linguistic perturbations, belonging to semantic, syntactic, and morphological error categories. These were divided into minor, major, and critical errors according to the type of perturbation. The application of DEMETR in evaluating a suite of baseline metrics revealed that metric performance varies considerably across the different error categories, often with no clear winner as also seen in our experiments later in the thesis. We summarise the statistics of different challenge sets in Table 2.2.

The WMT 2023 Challenge Sets submissions included ACES, MSLC23 (Lo et al., 2023), and an extended version of the DFKI challenge set to include the en→ru language pair plus additional examples and phenomena for the en→de language pair (Avramidis et al., 2023). The MSLC23 dataset covers four language pairs (zh→en, he↔en, and en→de) and includes examples of low-, medium- and high-quality output designed to provide an interpretation of metric performance across a range of different levels of translation quality. The motivation for this is that whilst metric performance may be evaluated on high-quality MT output, these same metrics may later be used to evaluate

low-quality MT output, and it is, therefore, important to understand their performance in the lower-quality setting.

The Chapter 4 discusses the development of the ACES challenge set which covers a comprehensive set of accuracy errors as well as offers broad coverage of language pairs. Whilst there is a clear overlap between the ACES phenomena and those in SMAUG and HWTSC, many of the phenomena in the DFKI dataset are complementary such that in the case of evaluating metrics for the German-English pair, metric developers might consider benchmarking on both datasets. It is worth noting that DEMETR and ACES each have their respective advantages: all examples in DEMETR have been verified by human annotators; ACES provides broader coverage in terms of both languages and linguistic phenomena. We note that the additional annotation of error spans in SPAN-ACES, an extension of the ACES dataset, is a first of its kind. SPAN-ACES promotes the development of metrics that produce error labels instead of scores.

2.4 Summary

In this chapter, we looked at the literature related to the foundation of our work. We discussed different human evaluation schemes for machine translation. As human evaluation is indeed tedious, it resulted in the development of several automatic metrics for MT evaluation. We listed some common design patterns across these metrics. Additionally, we discuss various practices that evaluate the effectiveness of these metrics. In the next chapter, we discuss evaluating the quality of a metric through its utility in a downstream task, a meta-evaluation that is first-of-its-kind.

Chapter 3

Extrinsic Evaluation of MT Metrics

In the previous chapter, we discussed different evaluation methods for Machine Translation in the context of evaluating MT systems. In this chapter, we introduce a new method to evaluate the effectiveness of MT metrics by their ability to correlate translation quality with translation utility in downstream tasks, discussed in Section 3.2. We evaluate the segment-level performance of nine MT metrics (CHRF, COMET, BERTScore, *etc.*) on three downstream cross-lingual tasks (dialogue state tracking, question answering, and semantic parsing). We also provide analyses to understand these results in Section 3.3. The contents of this chapter have been published in Moghe et al. (2023b).

3.1 Introduction

MT models are being more frequently deployed as a component of a complex NLP platform delivering multilingual capabilities such as cross-lingual information retrieval (Zhang et al., 2022) or automated multilingual customer support (Gerz et al., 2021). When an erroneous translation is generated by the MT systems, it may add new errors in the task pipeline leading to task failure and poor user experience. For example, consider the user's request in Chinese 剑桥有牙买加菜吗？ ("*Is there any good Jamaican food in Cambridge?*") machine-translated into English as *"Does Cambridge have a good meal in Jamaica?"*. The model will erroneously consider "Jamaica" as a location, instead of cuisine, and prompt the search engine to look up restaurants in Jamaica [1]. To avoid this *breakdown*, it is crucial to detect an incorrect translation before it causes further errors in the task pipeline.

Recent MT metrics have demonstrated high correlation with human judgements at

[1]Example from the Multi^{2}WoZ dataset (Hung et al., 2022)

the system-level for some language pairs (Freitag et al., 2021b, 2023). These metrics are potentially capable of identifying subtle differences between MT systems that emerge over a relatively large test corpus. These metrics are also evaluated on respective correlation with human judgements at the segment-level, however, there is a considerable performance penalty (Ma et al., 2019; Freitag et al., 2022).

Segment-level evaluation of MT is indeed more difficult and even humans have low inter-annotator agreement on this task (Popović, 2021). Despite MT systems being a crucial intermediate step in several applications, characterising the behaviour of these metrics under task-oriented evaluation has not been explored.

In this chapter, we provide a complementary evaluation of MT metrics. We focus on the segment-level performance of metrics, and we evaluate their performance extrinsically, by correlating the outcome of downstream tasks with scores from the respective metrics. We assume access to a parallel task-oriented dataset, a task-specific monolingual model, and a translation model that can translate from the target language into the language of the monolingual model (task language). We consider the *Translate-Test* setting — where at test time, the examples from the test language are translated to the task language for evaluation. We use the outcomes of this extrinsic task to construct a breakdown detection benchmark for the metrics.

We use dialogue state tracking, semantic parsing, and extractive question answering as our extrinsic tasks. We evaluate nine metrics consisting of string overlap metrics, embedding-based metrics, and metrics trained using scores from human evaluation of MT. Surprisingly, we find our setup challenging for all existing metrics; demonstrating poor capability in discerning good and bad translations across tasks. We present a comprehensive analysis of the failure of the metrics through quantitative and qualitative evaluation.

3.2 Methodology

We aim to determine how reliable MT metrics are for predicting success on downstream tasks. We use tasks and datasets proposed in the academic multilingual NLP literature to simulate an online setup of extrinsic tasks. Briefly, the process is as follows:

Our setup uses a model trained for a specific task (e.g., a dialogue state tracker) only using monolingual data from a *task language*. We chose tasks with parallel test sets from at least three other *test languages.* We use MT to translate a test sentence from a *test language* to the *task language* and then record the output from the task model

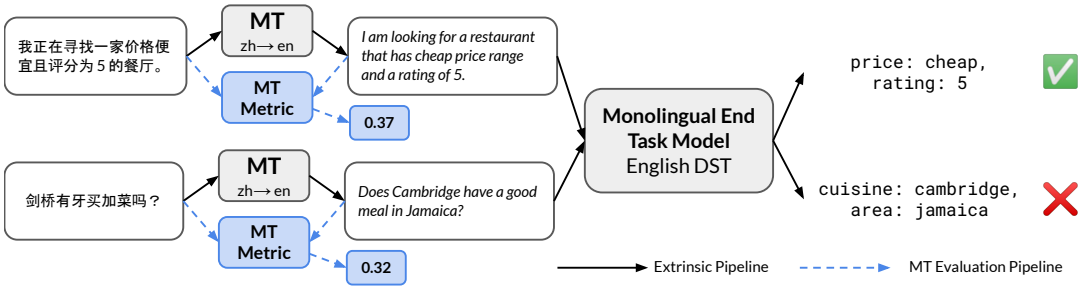

Figure 3.1: The meta-evaluation pipeline. The predictions for the extrinsic task in the test language (Chinese, zh) are obtained using the *Translate-Test* setup — the test language is translated into the task language (English, en) before passing to the task-specific model. The input sentence (zh) and the corresponding translations (en) are evaluated with a metric of interest. The metric is evaluated based on the correlation of its scores with the predictions of the end task.

(for example, the corresponding dialogue state) using the translated input sentence. If the model predicts a correct label for the original *task language* input but an incorrect label for the translated *test language* input, then we have observed a *breakdown* due to a material error in the translation pipeline. We then study if the metric could predict if the translation is suitable for the end task. We refer to Figure 3.1 for an illustration. In that figure, we find that the metric has provided scores in the same range for translation which caused a breakdown v/s a perfect translation. This questions if metrics are operating similarly to random chance which we explore in this work. We frequently use the terms *test language* and *task language* to avoid confusion with the usage of *source language* and *target language* in the traditional machine translation setup. In Figure 3.1, the task language is English and the test language is Chinese. We now describe our evaluation setup and the metrics under investigation.

3.2.1 Setup

Our setup consists of four steps:

1. **Train a task model**: We select tasks that have a fixed set of labels. Within these tasks, we select datasets with N-way parallel test sets, where N is the number of languages. The semantic content of respective examples across different languages is identical, and they share the ground truth labels. We train a task-specific model using training data of one of these N-way test sets. In most cases, the chosen task language is a high-resource language. [2] We store predictions of

[2]For most of the tasks, we only have access to a single task language. Most multilingual datasets in the academic literature are created after the monolingual task is popular, as an afterthought. The

the subset of examples which are correctly predicted in the task language to avoid errors that arise from extrinsic task complexity. In the example in Figure 3.1, we train a dialogue state tracking model in English.

2. **Simulate Translate-Test**: Based on the *Translate-Test* paradigm (Hu et al., 2020), we translate the examples from each test language into the task language. The generated translations are then fed to the task-specific monolingual model. We use either (i) OPUS translation models (Tiedemann and Thottingal, 2020), (ii) M2M100 translation (Fan et al., 2021) or (iii) translations provided by the authors of the respective datasets. Note that the examples across all the languages are parallel and we always have access to the correct label for a translated sentence. We obtain the predictions for the translated data only on the subset of examples that had a correct prediction in the task language for the previous step. In the above example, 剑桥有牙买加菜吗？ is translated into English using an MT system. The translated output "Does Cambridge have a good meal in Jamaica" is sent to the English dialogue state tracker and the resulting output {‘‘cuisine’’: ‘‘Cambridge’’, ‘‘area’’: ‘‘Jamaica’’} is stored for further calculations.

3. **Score the translations**: We now look at the translations obtained from the translation models independent of their outputs from the task model. We consider the example from the test language as *source*, the corresponding machine translation as *hypothesis* and the human reference from the task language as *reference*. Thus, in Figure 3.1, the source is 剑桥有牙买加菜吗？, the hypothesis is "Does Cambridge have a good meal in Jamaica?", and the reference will be "Is there any good Jamaican food in Cambridge?". These triples are then scored by the respective metrics. In an ideal online setting, we will not have access to the references. We include references in our setup to benchmark popular reference-based metrics.

4. **Check correlation between quality and utility**: We create a binary breakdown detection benchmark based on the predictions obtained in the first and the second steps. As we have only considered the examples that were correct predictions in the first step, we can safely assume that the incorrect predictions (dialogue states) arise due to the erroneous translations produced by the intermediate translation models. This as far as is practically possible the extrinsic task failure as the fault

efforts are more focused towards creating an evaluation set rather than recreating the entire dataset across multiple languages.

of *only* the MT system. We use these predictions to build a binary classification benchmark—all test language examples that are correctly predicted in the extrinsic task receive a positive label (no breakdown) while the incorrect predictions receive a negative label (breakdown). In the above example, the first dialogue predicts a correct dialouge state receiving "no breakdown" while the second dialogue results in an incorrect prediction and is marked as "breakdown".

We convert the scores produced by the metric into "breakdown" and "no breakdown" predictions by obtaining a threshold. We plot a histogram over the scores with ten bins for every setup per language pair and select the interval with the highest performance on the development set as a threshold. For example, if the threshold for the metric in Figure 3.1 is 0.5, it would mark both examples as bad translations, despite the first example being a usable translation. The metrics are then evaluated on how well their predictions for usable/erroneous translations correlate with the breakdown detection labels.

3.2.2 Tasks

We choose tasks that contain outcomes belonging to a small set of labels, unlike natural language generation tasks which have a large solution space. This discrete nature of the outcomes allows us to quantify the performance of MT metrics based on standard classification metrics. The tasks also include varying types of textual units: utterances, sentences, questions, and paragraphs, allowing a comprehensive evaluation of the metrics. We list an example for each of the three tasks in Table 3.1.

3.2.2.1 Semantic Parsing (SP)

Semantic parsing transforms natural language utterances into logical forms to express utterance semantics in a machine readable format. This format can be a logical formula or a structured query for a database like SQL as seen in the example. The semantic parsing acts as an interpreter between natural language and machine language.

The original ATIS study (Hemphill et al., 1990) collected questions about flights in the USA with the corresponding SQL to answer respective questions from a relational database. The participants were asked to interact with a flight reservation simulator in English and were given hypothetical scenarios to complete the task. These conversations were annotated with corresponding SQL queries to produce the ATIS dataset. We use the MultiATIS++SQL dataset from Sherborne and Lapata (2022), the multilingual version

of the dataset, comprising gold parallel utterances in English, French, Portuguese, Spanish, German and Chinese (from Xu et al. (2020)) paired to executable SQL output logical forms (from Iyer et al. (2017)). The entire dataset is replicated across different languages, giving us N-way parallel train/development/test sets unlike in the remaining tasks where we only have N-way parallel test sets and/or development tests. All the six languages act as respective task languages and the remaining five languages act as test languages in our setup. The monolingual task model follows Sherborne and Lapata (2023), as an encoder-decoder Transformer model based on mBART50 (Tang et al., 2021). The input is user request and the predicted output is the SQL generation. This generated SQL is executed on a database system and the resulting outputs are recorded.

The performance for the extrinsic task is measured as exact-match *denotation accuracy*—the proportion of output queries returning identical database results relative to gold SQL queries. The metric scores are obtained for individual user requests. A translation is considered faulty when the score of the translation of the utterance falls below the chosen threshold for that metric.

3.2.2.2 Extractive Question Answering (QA)

The task of extractive question answering is predicting a span of words from a paragraph/document that directly address the given question. In the example in the table, the answer lies in within the first sentence.

The English SQuAD dataset (Rajpurkar et al., 2016) is a widely used question answering benchmark for evaluating models capable of extractive question answering. This dataset was created with the help of crowdworkers who were shown content from diverse Wikipedia articles and asked to create relevant questions for the article. The answers were marked as a segment of text (span) from the corresponding reading passage. In our setup, we use the XQuAD dataset (Artetxe et al., 2020) for evaluating extractive question answering. The XQuAD dataset was obtained by professionally translating examples from the development set of English SQuAD dataset (Rajpurkar et al., 2016) into ten languages: Spanish, German, Greek, Russian, Turkish, Arabic, Vietnamese, Thai, Chinese, and Hindi.[3] We use the publicly available question answering model that fine-tunes RoBERTa (Liu et al., 2019) on the SQuAD training set consisting of English examples. [4]. Thus, English acts as task language and the remaining language acts as test languages.

[3] SQuAD has a private test set.

[4] https://huggingface.co/csarron/roberta-base-squad-v1

We use the *Exact-Match* metric, i.e., the model's predicted answer span exactly matches the gold standard answer span; for the breakdown detection task. The metrics scores are produced for the question and the context. A translation is considered to be faulty if either of the scores falls below the chosen threshold for every metric.

3.2.2.3 Dialogue State Tracking (DST)

In the dialogue state tracking task, a model needs to map the user's goals and intents in a given conversation to a set of slots and values, known as a *dialogue state*, based on a pre-defined ontology. In the table, we find a dictionary of slot names such as "hotel-area", "hotel-parking" followed by the values requested by the user so far in the conversation. The first part of the slot-name is the domain name.

MultiWoZ 2.1 (Eric et al., 2020) is a popular dataset for examining the progress in dialogue state tracking which consists of multi-turn conversations in English. It features human-human conversations across seven domains (restaurants, hotels, etc.) paired with annotations like user intentions and detailed slot descriptions. Specifically, every user utterance is annotated with a dialogue state which we use in our setup. The conversations consist of restaurant reservation, searching and requesting information about travel, and so on. We consider the Multi^{2}WoZ dataset (Hung et al., 2022) where the development and test set have been professionally translated into German, Russian, Chinese, and Arabic from the MultiWoZ 2.1 dataset. We use the dialogue state tracking model trained on the English dataset by Lee et al. (2019). This model uses a pre-trained English encoder to obtain contextual semantic vectors for the utterances, slot-names, and slot values. It then uses a multi-head attention network to learn the relationship between slot-names and slot-values appearing in the text to predict the dialogue states. The task language is English and the remaining languages act as test languages.

We consider the *Joint Goal Accuracy* where the inferred label is correct only if the predicted dialogue state is exactly equal to the ground truth to provide labels for the breakdown task. We use oracle dialogue history and the metric scores are produced only for the current utterance spoken by the user.

3.2.3 Metrics

We use the following metrics in our study. See Section 2.2 for their description.

1. Surface-level overlap: BLEU(Papineni et al., 2002) and CHRF(Popović, 2017)

Task	Input	Output
Semantic Parsing	what flights go from dallas to phoenix	`SELECT DISTINCT flight . flight_id FROM flight WHERE (flight . from_airport IN (SELECT airport_service . airport_code FROM airport_service WHERE airport_service . city_code IN (SELECT city . city_code FROM city WHERE city . city_name = 'DALLAS')) AND flight . to_airport IN (SELECT airport_service . airport_code FROM airport_service WHERE airport_service . city_code IN (SELECT city . city_code FROM city WHERE city . city_name = 'PHOENIX'))) ;`
Extractive Question Answering	**Question**: What is the usual source of heat for boiling water in a steam engine? **Passage**: The heat required for boiling the water and supplying the steam can be derived from various sources most commonly from **burning combustible materials** with appropriate supply .... the heat source can be an electric heating element	burning combustible materials
Dialogue State Tracking	**User**: I need to book a hotel in the east that has 4 stars. **Agent**: I can help you with that. What is your price range? **User**: That doesn't matter as long as it has free wifi and parking.	`['hotel-area-east', 'hotel-parking-yes', 'hotel-stars-4', 'hotel-internet-yes']`

Table 3.1: Examples for the three extrinsic tasks taken from the respective datasets: MultiATIS++SQL (Sherborne and Lapata, 2022), XQuAD (Artetxe et al., 2020), and Multi^{2}WoZ (Hung et al., 2022).

2. Embedding based: BERTSCORE (Zhang et al., 2020), COMET (Rei et al., 2020), and UniTE(Wan et al., 2022). For UniTE, we use the source-reference and the source-only version. For COMET, we consider the following variants: COMET20-DA (Rei et al., 2020), COMET21-QE-DA, COMET21-MQM, and COMET21-QE-MQM (Rei et al., 2021).

The metrics BERTSCORE, COMET family and UniTE family can run on both GPU and CPU. If run on GPU, the metrics run under 5 minutes for a given task and given language pair. No hyperparameters are required. We follow the standard train-dev-test split as released by the authors for DST (Hung et al., 2022) and SP (Sherborne and Lapata, 2022). As no development set is available for the XQuAD dataset, we use the first 200 examples as the development set to choose the threshold but report the performance on the full test set.

3.2.4 Metric Evaluation

The meta-evaluation for the above metrics uses the breakdown detection benchmark. As the class distribution changes depending on the task and the language pair, we require an evaluation that is robust to class imbalance. We consider using macro-F1 and Matthew's Correlation Coefficient (MCC) (Matthews, 1975) on the classification labels. The range of macro-F1 is from 0 to 1 with equal weight to positive and negative classes. Macro-F1 is useful as the classes can be imbalanced depending on the language pair and the extrinsic task. It considers performance across all classes, not just the majority class.

We include MCC to interpret the MT metric's standalone performance for the given extrinsic task. We describe its formulation based on the confusion matrix in Equation (3.1) where TP is the number of true positives, TN is the number of true negatives, FP is the false positive total, and FN denotes False Negatives. The range of MCC is between -1 to 1. An MCC value near 0 indicates no correlation with the class distribution. Any MCC value between 0 and 0.3 indicates negligible correlation, 0.3 to 0.5 indicates low correlation.

$$MCC = \frac{TP \times TN - FP * FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \tag{3.1}$$

3.3 Results

We report the aggregated results for semantic parsing, question answering, and dialogue state tracking in Table 3.2. We use a random baseline for comparison which assigns the positive and negative labels with equal probability. We report the results for the extrinsic task as well as the results for the respective *translate-test* performance in Appendix A.2. We note the initial monolingual models have strong task-specific performance across all three tasks with varying degradation on the translate-test setup for the remaining languages. This degradation can be attributed to quality of the underlying MT model, data capacity of the language pair (high-resource/low-resource, as well as domain overlap with the MT system. This diverse performance across tasks and language-pairs is beneficial for our setup as it allows us an imbalanced test set for the metric. The language pairs with poor Translate-Test performance are likely to have more "breakdown" labels and vice-versa. We note that a higher monolingual task performance allows us to construct the breakdown detection benchmark in larger scale.

Metric	Semantic Parsing		Question Answering		Dialogue State Tracking	
	F1	MCC	F1	MCC	F1	MCC
Random	0.453	-0.034	0.496	0.008	0.493	0.008
BLEU	0.580	0.179	0.548	0.121	0.529	0.082
CHRF	0.609	0.234	0.554	0.127	0.508	0.067
BERTSCORE	0.590	0.205	0.555	0.127	0.505	0.071
COMET20-DA	0.606	0.228	0.562	0.137	0.608	0.244
COMET21-MQM	0.556	0.132	0.387	0.027	0.597	0.204
UniTE	0.600	0.225	0.375	0.012	0.620	0.262
COMET21-QE-DA	0.556	0.135	0.532	0.100	0.561	0.145
COMET21-QE-MQM	0.597	0.211	0.457	0.033	0.523	0.094
UNITE-QE	0.567	0.155	0.388	0.032	0.587	0.192
Ensemble	0.620	0.251	0.577	0.168	0.618	0.248

Table 3.2: Performance of MT metrics on the classification task for extrinsic tasks Parsing (Multi-ATIS++SQL), Question Answering (XQuad) using an English-trained question answering system, and Dialogue State Tracking (Multi^{2}WoZ) using an English-trained state tracker. Reported Macro F1 scores and MCC scores quantify if the metric detects a breakdown for the extrinsic task. Metrics have a negligible correlation with the outcomes of the end task. MCC and F1 are average over respective language pairs

We find that almost all metrics perform above the random baseline on the macro-F1 metric. We use MCC to identify if this increase in macro-F1 makes the metric usable in the end task. Evaluating MCC, we find that all the metrics show negligible correlation across all three tasks. Contrary to the trends where neural metrics are better than metrics based on surface overlap (Freitag et al., 2021b), we find this breakdown detection to be difficult irrespective of the design of the metric. We also evaluate an ensemble with majority voting of the predictions from the top three metrics per task. Ensembling provides minimal gains suggesting that metrics are making similar mistakes despite varying properties of the metrics. We report fine-grained results in Appendix A.2 for the three tasks as we find consistent poor performance across all the language pairs.

Comparing the reference-based versions of trained metrics (COMET20-DA, COMET21-MQM, UniTE) with their reference-free quality estimation (QE) equivalents, we observe that reference-based versions perform better, or are competitive to, their reference-free versions for the three tasks. We also note that references are unavailable when the systems are in production, hence reference-based metrics are unsuitable for realistic settings. We discuss alternative ways of obtaining references in Section 3.3.4.

Between the use of MQM-scores and DA-scores during fine-tuning COMET vari-

ants, we find that both COMET21-QE-DA and COMET20-DA are strictly better than COMET21-QE-MQM and COMET21-MQM for question answering and dialogue state tracking respectively, with no clear winner for semantic parsing.

Unlike dialogue state tracking and question answering, the parallel train set available in the allows us to test our setup across task languages other than English. The fine-grained results in Tables A.5 and A.6 in Appendix A.2 suggest that the choice of task language has no drastic effect on the metric. Further, the results on per-language pair in Appendix A.2 suggest that no specific language pairs stand out as easier/harder across tasks. As this performance is already poor, we cannot verify if neural metrics can generalise in evaluating language pairs unseen during training.

3.3.1 Case Study

We look at Semantic Parsing with an English-trained parser tested with Chinese inputs for our case study with the well-studied COMET20-DA metric [5]. We select semantic parsing for the case study because the examples in this task are at sentence level and COMET20-DA is predominantly trained with examples at the sentence level. Its fine-tuning set includes examples from Chinese→English MT systems, making the setup particularly favourable for the COMET20-DA. We report the number of correct and incorrect predictions made by the metric across ten equal ranges of scores in Figure 3.2. The bars labelled on the x-axis indicate the end-point of the interval i.e., the bar labelled -0.74 contains examples that were given scores between -1.00 and -0.74.

First, we highlight that the threshold is -0.028, counter-intuitively suggesting that even some correct translations receive a negative score. We expected the metric to fail in the regions around the threshold as those represent the strongest confusion. For example, “周日下午从迈阿密飞往克利夫兰” is correctly translated as “Sunday afternoon from Miami to Cleveland” yet the metric assigns it a score of -0.1. However, the metric makes mistakes throughout the bins. For example, “我需要预订一趟联合航空下周六的从辛辛那提飞往纽约市的航班” is translated as “I need to book a flight from Cincinnati to New York City next Saturday.” and loses the crucial information of “United Airlines”; yet it is assigned a high score of 0.51. This demonstrates that the metric possesses a limited perception of a good or bad translation for the end task.

We suspect this behaviour is due to the current framework of MT evaluation. The development of machine translation metrics largely caters towards the intrinsic task

[5]COMET20-DA the default model choice when this study was conducted.

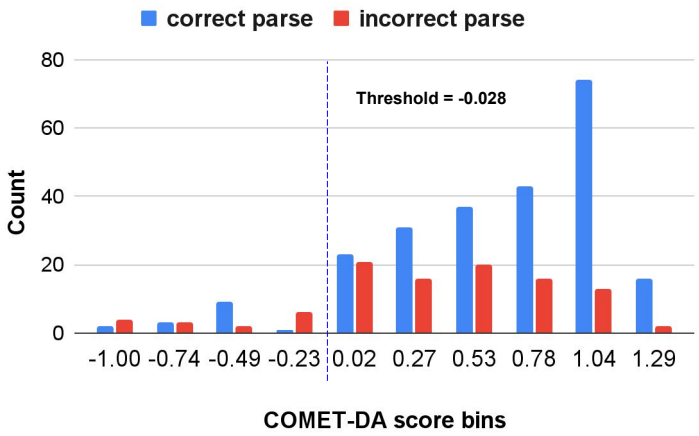

Figure 3.2: Graph of predictions by COMET20-DA (threshold: -0.028), categorised by the metric scores in ten intervals. Task: Semantic Parsing with English parser and test language is Chinese. The bars indicate the count of examples with incorrect parses (red) and correct parses (blue) assigned the scores for the given ranges.

of evaluating the quality of a translated text in the target language. The severity of a translation error is dependent on the guidelines released by the organisers of the WMT metrics task or the design choices of the metric developers. Our findings agree with Zhang et al. (2022); Doyon et al. (1999) that different downstream tasks will demonstrate varying levels of sensitivity to the same machine translation errors.

3.3.2 Qualitative Evaluation

To quantify detecting which translation errors are most crucial to the respective extrinsic tasks, we conduct a qualitative evaluation of the MT outputs and task predictions. We annotate 50 false positives and 50 false negatives for test languages Chinese (SP), Hindi (QA), and Russian (DST) respectively. The task language is English. We annotate the MT errors (if present) in these examples based on the MQM ontology. We tabulate these results in Table 3.3 using COMET20-DA for these analyses.

Within the false negatives, a majority of the errors ($>$48%) are due to the metric's inability to detect translations containing synonyms or paraphrases of the references as valid translations. Further, omission errors detected by the metric are not crucial for DST as these translations often exclude pleasantries. Similarly, errors in fluency are not important for both DST and SP but they are crucial for QA as grammatical errors in questions produce incorrect answers. Mistranslation of named entities (NEs), especially which lie in the answer span, is a false negative for QA since QA models

Task	MT error	Prediction	input	reference	hypothesis	gold task output	translated task output
SP	mistranslation	No Breakdown	哪些航空公司在 多伦多 和 圣地亚哥 之间飞行	which airlines fly between toronto and san diego	Which airlines fly between Toronto and Santiago?	SELECT DISTINCT airline_1 … city1.city_name = 'TORONTO' … city_2 . city_name = 'SAN DIEGO' ;	SELECT DISTINCT airline_1 ...city1.city_name = 'TORONTO'; (city_2 is excluded)
DST	mistranslation	No Breakdown	Я ищу такси из Yu Garden, которое прибудет к 14:30.	I am looking for a taxi from yu garden arriving by 14:30	I'm looking for a taxi from Yu Garden, which will arrive by 2:30.	['taxi-departure-yu garden', 'taxi-arriveby-14:30']	['taxi-departure-yu garden', 'taxi-arriveby-02:30']
QA	fluency	No Breakdown	विस्तारित महानगरीय क्षेत्र कितने हैं?	How many extended metropolitan areas are there?	How much are the extended metropolitan areas?	two	exceed five million in population.
QA	mistranslation	Breakdown	एनर्जीप्रोजेक्ट AB कहाँ स्थित है?	Where is Energiprojekt AB based?	Where is Energyproject AB located?	Sweden	Sweden
SP	none	Breakdown	查询从 底特律 飞往 多伦多 的航班	get flights from detroit to toronto	Query flights from Detroit to Toronto.	SELECT DISTINCT flight_1 .. city1.city_name = 'DETROIT'... city2.city_name = 'TORONTO';	SELECT DISTINCT flight_1 .. city1.city_name = 'DETROIT'... city2.city_name = 'TORONTO';
DST	none	Breakdown	Да. Забронируйте на 3 человека.	yes. book for 3 people.	Yeah, make a reservation for three people.	['train_book-people-3']	['train_book-people-3']
QA	none	Breakdown	वारसॉ हमेशा से किस प्रकार का शहर रहा है?	What type of city has Warsaw been for as long as it's been a city?	What kind of city has Warsaw always been?	multi-cultural	multi-cultural

Figure 3.3: Examples of errors made by COMET20-DA. It struggles at identifying NE errors and paraphrases.

Task	Errors by the Extrinsic model	False Positive	False Negative
SP	25%	mistranslation (90%), omission(10%)	mistranslation (25.7%), fluency (20%), omission (5.7%), no error (48.6%)
QA	20%	mistranslation (60%), omission(8.6%), addition (5.7%), fluency (20%), undertranslation (2.9%), untranslated (2.9%)	mistranslation (18%), fluency (22%), addition (2%), no error (54%)
DST	5%	mistranslation (100%)	omission (26%), mistranslation (1%), no error (73%)

Table 3.3: The proportion of the different types of errors erroneously detected and undetected by COMET20-DA for languages mentioned in Section 3.3.2. False positives and false negatives are computed by excluding the examples where the extrinsic task model was at fault.

find the answer by focusing on the words in the context surrounding the NE rather than the error in that NE. Detecting mistranslation in NEs is crucial for both DST and SP as this error category dominates the false positives. A minor typo of *Lester* instead of *Leicester* marks the wrong location in the dialogue state which is often undetected by the metric. Addition and omission errors are undetected for SP while mistranslation of reservation times is undetected for DST.

We find that some of the erroneous predictions can be attributed to the failure of the extrinsic task model than the metric. For example, the MT model uses an alternative term of *direct* instead of *nonstop* while generating the translation for the reference "show me nonstop flights from montreal to orlando". The semantic parser fails to generalise despite being trained with mBART50 to ideally inherit some skill at disambiguiting semantically similar phrases. This error type accounts for 25% for SP, 20% for QA and 5% in DST of the total annotated errors. We give examples in Figure 3.3. Finally, we note that the qualitative evaluation was carried out when the output is in English. Future studies should focus on a thorough evaluation across multiple languages to verify if the observations hold through across those languages.

3.3.3 Finding the Threshold

Interpreting system-level scores provided by automatic metrics requires additional context such as the language pair of the machine translation model or another MT system for comparison [6]. In this classification setup, we rely on interpreting the segment-level score to determine whether the translation is suitable for the downstream task. We

[6] https://github.com/Unbabel/COMET/issues/18

Extrinsic Task	SP	QA	DST
BLEU	15.5 ± 08.8	16.1 ± 04.9	20.0 ± 0.00
CHRF	44.0 ± 13.7	53.9 ± 07.8	30.7 ± 0.45
BERTSCORE	0.50 ± 0.21	0.54 ± 0.08	0.39 ± 0.21
COMET20-DA	0.21 ± 0.35	0.30 ± 0.23	0.58 ± 0.08
COMET21-MQM	0.03 ± 0.01	0.06 ± 0.01	0.02 ± 0.00
UniTE	0.04 ± 0.22	-0.40 ± 0.38	-0.01 ± 0.29
COMET21-QE-DA	0.02 ± 0.07	0.02 ± 0.01	0.06 ± 0.01
COMET21-QE-MQM	0.11 ± 0.01	0.00 ± 0.04	0.03 ± 0.00
UniTE-QE	-0.01 ± 0.22	-0.24 ± 0.13	0.11 ± 0.18

Table 3.4: Mean and Standard Deviation of the best threshold on the development set for all the language pairs in the respective extrinsic tasks. The thresholds are inconsistent across language pairs and tasks for both bounded and unbounded metrics.

find that choosing the right threshold to identify translations requiring correction is not straightforward. Our current method to obtain a threshold relies on validating candidate thresholds on the development set and selecting an option with the best F1 score. These different thresholds are obtained by plotting a histogram of scores with ten bins per task and language pair.

We report the mean and standard deviation of best thresholds for every language pair for every metric in Table 3.4. Surprisingly, the thresholds are inconsistent and biased for bounded metrics: BLEU (0–100), CHRF (0–100), and BERTSCORE (0–1). The standard deviations across the table indicate that the threshold varies greatly across language pairs. We find that thresholds of these metrics are also not transferable across tasks. COMET metrics, except COMET20-DA, have lower standard deviations. By design, the range of COMET metrics in this work is unbounded. However, as discussed in the theoretical range of COMET metrics [7], empirically, the range for COMET21-MQM lies between -0.2 to 0.2, questioning whether the lower standard deviation is an indicator of threshold consistency. Some language pairs within the COMET metrics have negative thresholds. We also find that some of the use cases under the UniTE metrics have a mean negative threshold, indicating that good translations can have negative UniTE scores. Similar to Marie (2022), we suggest that the notion of negative scores for good translations, only for certain language pairs, is counter-intuitive as most NLP metrics tend to produce positive scores.

Thus, we find that both bounded and unbounded metrics discussed here do not provide segment-level scores whose range can be interpreted meaningfully across tasks

[7] https://unbabel.github.io/COMET/html/faqs.html

Metric	SP	QA	DST
BLEU	0.003	0.013	0.050
CHRF	0.018	0.021	0.055
BERTSCORE	0.028	0.065	0.036
COMET20-DA	0.071	0.085	0.083
COMET21-MQM	0.080	0.019	0.116
UNITE	0.225	0.056	0.193

Table 3.5: MCC scores of reference based metrics with pseudo references when gold references are unavailable at test time. Performance is worse than metrics with oracle references and reference-free metrics (Table 3.2)

and language pairs.

3.3.4 Reference-based Metrics in an Online Setting

In an online setting, we do not have access to references at test time. To test the effectiveness of reference-based methods here, we consider translating the translation back into the test language, *i.e.,* the round trip translation. For example, for an English(en) parser, the test language ti_{zh} is translated into mt_{en} and then translated back to Chinese(zh) as mt_{zh}. The metrics now consider mt_{en} as source, mt_{zh} as hypothesis, and ti_{zh} as the reference. We generate these new translations using the mBART50 translation model (Tang et al., 2021) and report the results in Table 3.5.

Compared to the results in Table 3.2, there is a further drop in performance across all the tasks and metrics. The metrics also perform worse than their reference-free counterparts. The second translation is likely to add additional errors to the existing translation. This cascading of errors confuses the metric and it can mark a perfectly useful translation as a breakdown. The only exception is that of the UniTE metric which has comparable performance (but overall poor) due to its multi-task training.

3.3.5 Towards Span-based Evaluation

MT evaluation in academic setups focuses on intrinsic task of quality of machine translation without any downstream objective. While conducting human evaluation of MT outputs, especially direct assessments, the guidelines for assigning scores in given ranges are based on use of the translated sentence as it is. Metrics that are trained on this data inherit the noise from this process. Throughout this work, we test if this generic perception of quality of translation is actually useful when MT systems are used in downstream tasks.

Our results indicate that existing metrics do not correlate translation quality with translation utility. The primary reason for this is seen in Section 3.3.2 and Section 3.3.3 that suggest interpreting the quality of the produced MT translation based on a number is unreliable and difficult. We recommend exploring segment-level MT evaluation as an error classification task instead of a regression task. Specifically, we recommend designing a setup where the words in the source/hypothesis can be tagged with explicit error labels. As seen in Section 3.3.2, different tasks have exhibit tolerance to different MT errors. Our hypothesis is that moving evaluation from scores to labels would reduce the overhead in picking examples which need post-editing in online settings. As compared to opaque scores, independent segments will have more information about the type of the error present in that translation when presented with error labels. We can then correct only the translations that contain errors critical to the chosen downstream application while allowing the pipeline to process the remaining translations. Independent of the downstream task, intrinsic evaluation of MT systems with labels will aid in explanability of the drawbacks of their proposed methods. .

We advocate using the MQM scoring scheme with expert annotators (See Section 2.1.2) for evaluating MT outputs. This is in line with Lommel et al. (2014); Freitag et al. (2021a) and the recent WMT challenges (Freitag et al., 2021b, 2022, 2023) advocate the use of labelled error spans for MT evaluation. Using a human evaluation approach similar to MQM will lead to a comprehensive collection of human assessments(Freitag et al., 2021a). With this rich repository of human evaluation, future breakdown classifiers can be only trained on a subset of human evaluation data containing errors most relevant to the downstream application.

3.4 Summary

We proposed a method for evaluating MT metrics which is reliable at the segment-level and does not depend on human judgements by using correlation MT metrics with the success of extrinsic downstream tasks. We evaluated nine different metrics on the ability to detect errors in generated translations when machine translation is used as an intermediate step for three extrinsic tasks: Semantic Parsing, Question Answering, and Dialogue State Tracking. We found that segment-level scores provided by all the metrics show negligible correlation with the success/failure outcomes of the end task across different language pairs. We attributed this result to segment scores produced by these metrics being uninformative and that different extrinsic tasks demonstrate different levels of sensitivity to different MT errors. We recommend future segment-level evaluation to focus on metrics that predict labels and mark the erroneous text instead of a single summary score.

Chapter 4

Construction of ACES and SPAN-ACES

In Chapter 3, we found that single summary scores produced by metrics at segment-level are uninformative. We also observed that different tasks exhibit varying sensitivity to different errors. The previous results describe the errors observed in those three tasks. To decide whether a particular metric is suitable for any other task or is robust to a particular error of interest, we develop a collection of contrastive challenge sets, consisting of 68 distinct MT errors. This chapter describes the development of ACES and the subsequent version with annotated error spans - SPAN-ACES. The development of ACES was first reported in Amrhein et al. (2022) while SPAN-ACES is detailed in Moghe et al. (2024).

4.1 Introduction

MT metrics are a fundamental component of the development of high-quality MT systems as most state-of-the-art models claim their effectiveness through such metrics (Kocmi et al., 2021). While human evaluation of these MT systems is ideal, it is labour-intensive, time-consuming, and expensive. Development of automatic metrics has thus received significant interest over the past years (Koehn and Monz, 2006; Freitag et al., 2023) resulting in a surge of new metrics. To systematically study the advantages and shortcomings of such metrics, and to identify broad trends in metric development, we rely on the construction of challenge sets for MT metrics.

Challenge sets are a useful tool in measuring the performance of systems or metrics on one or more specific phenomena of interest. They may be used to compare the performance of a range of *different* systems or to identify performance improvement/degradation between successive iterations of the *same* system.

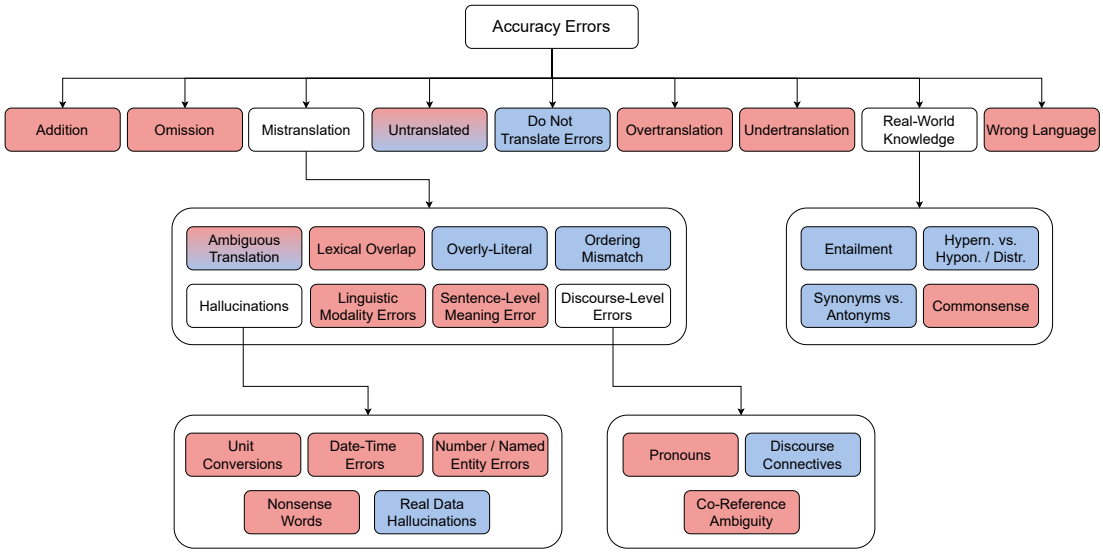

Figure 4.1: Diagram of the error categories present in ACES based on the MQM ontology. Red indicates challenge sets are created automatically, and blue means challenge sets are created manually.

The WMT 2021 Metrics shared task (Freitag et al., 2021b) introduced a shared task on constructing contrastive challenge sets for the evaluation of MT metrics. Contrastive challenge sets aim to assess how well a given metric can discriminate between a *good* and *incorrect* translation of the *source* text where the incorrect translation consists of a translation error of interest [1]. Providing a *reference* translation allows for flexibility: it may be included to assess reference-based (i.e. MT) metrics, or excluded to assess reference-free (i.e. Quality Estimation (QE)) metrics. Benchmarking metrics on such challenge sets provides insights into their strengths while simultaneously uncovering their weaknesses.

In this chapter, we describe the Translation **A**ccuracy **C**hallen**E** **S**et (ACES). The ACES dataset[2] consists of 36,476 examples covering 146 language pairs and representing challenges from 68 phenomena. Most MT metric challenge sets (Alves et al., 2022; Karpinska et al., 2022; Macketanz et al., 2018b) either focus on a small number of phenomena or a small number of languages. Our datasets are large scale in coverage of phenomena as well as language pairs providing a comprehensive challenge set for MT metrics.

We focus on translation accuracy errors because in recent years, machine translation outputs have become increasingly fluent (Bentivogli et al., 2016; Toral and Sánchez-Cartagena, 2017; Castilho et al., 2017). Further, accuracy errors can have dangerous

[1] We continue using the good/incorrect terminology to be consistent with the description in the shared task.

[2] The datasets are available at `https://huggingface.co/datasets/nikitam/ACES`

consequences in certain contexts, for example in the medical and legal domains (Vieira et al., 2021).

ACES uses the hierarchy of errors under the class *Accuracy* from the Multidimension Quality Metrics (MQM) ontology (Lommel et al., 2014) to design the ACES challenge sets. We based our challenge sets on MQM ontology as it is comprehensive in terms of errors and recent metrics have been adapting MQM like data during training (Rei et al., 2022a; Juraska et al., 2023). Additionally, it has been adopted for human evaluation in the recent WMT shared tasks (Freitag et al., 2021b, 2022, 2023). As seen from the previous chapter in Section 3.3.2, different tasks show varying sensitivity to the same MT error. Thus, we did not pre-define the error severity and left the interpretation of the severity dependent on the use case. 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. 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. A full overview of all error classes can be seen in Figure 4.1. Our challenge set consists of synthetically generated adversarial examples, examples from re-purposed contrastive MT test sets (both marked in red), and manually annotated examples (marked in blue).

A metric that can, in addition to providing scores, accurately label errors in MT output provides many clear advantages over one that only provides scores (Freitag et al., 2021a). Observations in Chapter 3 suggest that interpreting the quality of MT output based on a single summary score is both unreliable and uninformative. In Section 3.3.5, we recommend the development of metrics that predict labels for error spans in the MT output. When considering whether to deploy an MT system (or which of several systems to deploy), system developers can take into consideration the type, frequency, and severity of the errors that the system is likely to make, coupled with information about what types of errors may be tolerated/not for a given downstream task as also seen in Section 3.3.2.

With these motivations, we extend the ACES dataset into SPAN-ACES, where we include error span annotations for each example. These annotations indicate the location of error spans present in the *incorrect* translation and pertaining to the specific MT error in focus. In the case of examples that were obtained by changing some words in the alternate translations and references, it was easy to reverse-engineer the challenge

set construction process to mark the spans. For other examples, we resorted to the use of specific Python libraries and annotators to mark the errors.

This chapter discusses the construction of ACES and SPAN-ACES in detail. The next chapter (Chapter 5) includes the empirical results on these datasets. We use ACES to benchmark the metrics that participated in the WMT 2022 and 2023 metrics shared tasks. We further investigate if Large Language Models can perform MT evaluation (Kocmi and Federmann, 2023b; Xu et al., 2023) using ACES. We also benchmark GEMBA-MQM (Kocmi and Federmann, 2023b)), XCOMET-XL (Guerreiro et al., 2023), and adapted versions of COMET-22 (Rei et al., 2022a) and UniTE (Wan et al., 2022) on SPAN-ACES in Chapter 5.

4.2 Challenge Sets

Creating a contrastive challenge set for evaluating a machine translation 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. These included preliminary check with string-based metric followed by a manual inspection of a subset of the generated challenge sets. The details for the same are mentioned in the respective challenge set descriptions.

4.2.1 Datasets

The examples in ACES are based on several academic datasets designed to test particular properties in Machine Translation or other multilingual NLP tasks. 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).

FLORES-101 (Goyal et al., 2022) and **FLORES-200** (NLLB Team et al., 2022) are low resource MT evaluation benchmarks with parallel data in 101 and 200 languages respectively. **PAWS-X** (Yang et al., 2019) is a cross-lingual dataset for paraphrase identification in seven languages that consists of pairs of sentences that are labelled as true or adversarial paraphrases. **XNLI** (Conneau et al., 2018) is a multilingual Natural Language Inference (NLI) dataset consisting of premise-hypothesis pairs with their corresponding inference label for 14 languages. The other datasets used in the development of ACES serve specific challenges. **WinoMT** (Stanovsky et al., 2019), a challenge set developed for analysing gender bias in MT with examples exhibiting 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 **WMT 2018 English-German pronoun translation evaluation test suite** (Guillou et al., 2018) contains examples of the ambiguous English pronouns *it* and *they* extracted from the TED talks portion of ParCorFull (Lapshinova-Koltunski et al., 2018). 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 discourse connective annotation of their sense types. **Wino-X** (Emelin and Sennrich, 2021) is a parallel dataset of German, French, and Russian Winograd schemas, aligned with their English counterparts used to test commonsense reasoning and coreference resolution of MT models.

We will now discuss the different categories of challenge sets. We briefly explain their creation process and discuss it in detail in Appendix A.3.1. We list some examples from ACES in Table 4.1.

	Addition
	target includes content not present in the source
SRC (de):	In den letzten 20 Jahren ist die Auswahl in Uptown Charlotte exponentiell gewachsen.
REF (en):	In the past 20 years, the amount in Uptown Charlotte has grown exponentially.
✓:	Over the past 20 years, the selection in Uptown Charlotte has grown exponentially.
✗:	Over the past 20 years, the selection of **child-friendly options** in Uptown Charlotte has grown exponentially.
	Omission
	errors where content is missing from the translation that is present in the source
SRC (fr):	Une tornade est un tourbillon d'air à basse-pression en forme de colonne, l'air alentour est aspiré vers l'intérieur et le haut.
REF (en):	A tornado is a **spinning column** of very low-pressure air, which sucks the surrounding air inward and upward.
✓:	A tornado is a **column-shaped** low-pressure air turbine, the air around it is sucked inside and up.
✗:	A tornado is a low-pressure air turbine, the air around it is sucked inside and up.
	Untranslated - Word Level
	errors occurring when a text segment that was intended for translation is left untranslated in the target content
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.
	Mistranslation - Ambiguous Translation
	an unambiguous source text is translated ambiguously
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.
	Do Not Translate
	content in the source that should be copied to the output in the source language, but was mistakenly translated into the target language.
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".
	Undertranslation
	erroneous translation has a meaning that is more generic than the source
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**.
	Overtranslation
	erroneous translation has a meaning that is more specific than the source
SRC (ja):	その 40 分の映画はアノーがアラン・ゴダードと協力して脚本を書いた。
REF (en):	The 40-minute **film** was written by Annaud with Alain Godard.
✓:	The 40-minute **film** was written by Annaud along with Alain Godard.
✗:	The 40-minute **cinema verite** was written by Annaud with Alain Godard.
	Real-world Knowledge - Textual Entailment
	meaning of the source/reference is entailed by the "good" translation
SRC (de):	Ein Mann **wurde ermordet**.
REF (en):	A man **was murdered**.
✓:	A man **died**.
✗:	A man **was attacked**.
	Wrong Language
	incorrect translation is a perfect translation in a related 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.

Table 4.1: Examples from each top-level accuracy error category in ACES. An example consists of a source sentence (SRC), reference (REF), good (✓) and incorrect (✗) translations, language pair, and a phenomenon label. We also provide a description of the relevant phenomenon. en: English, de: German, fr: French, ja: Japanese, es: Spanish, ca: Catalan

4.2.2 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 using the likelihood of tokens for a given MT model. To generate examples, we use the concatenated dev and devtest sets from the FLORES-101 evaluation benchmark for 46 languages for which there exists a stanza parser[3]. We create datasets for all languages paired with English plus ten additional language pairs that we selected randomly. For translation, we use the M2M100[4] model with 1.2B parameters (Fan et al., 2021).

4.2.3 Mistranslation

The mistranslation phenomenon is broadly defined as the target translation not accurately containing the information in the source content.

4.2.3.1 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

[3] https://stanfordnlp.github.io/stanza/available_models.html
[4] https://huggingface.co/facebook/m2m100_1.2B

cues to an ambiguous reference to create two contrasting translation hypotheses. We create three separate challenge sets of this type:

Occupation Name Gender using the WinoMT dataset where the source language is a gendered language and the target language is English. The cues added to the reference to form the "good" and "incorrect" translations are "female" and "male".

Word Sense Disambiguation using the MuCoW dataset where the ambiguity lies in homographs in the target language that are unambiguous in the source sentence. The cues added to the reference to form the contrastive translations are sense-specific.

Discourse Connectives using the Europarl ConDisco corpus where the ambiguity lies in the English discourse connective "since" which can have both causal and temporal meanings.

4.2.3.2 Mistranslation - Hallucinations

In this category, we group several subcategories of mistranslation errors that happen at the word level and could occur due to hallucination by an MT model. Hallucinations are erroneous generations where the output is partially related or unrelated to the source sentence (Dale et al., 2023). These challenge sets test whether the machine translation evaluation metrics can reliably identify hallucinations when presented with a correct alternative translation.

We create five different challenge sets based on hallucination errors:

Date-Time Errors: using the FLORES-101 data where a month name in the reference (e.g. November) is replaced with a corresponding abbreviation in the "good" translation (e.g. Nov.) and a different month name in the "incorrect" translation (e.g. August).

Numbers and Named Entities: We create a challenge set for numbers and named entities where 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. 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 the 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. We use cross-lingual paraphrases from the PAWS-X dataset as a pool of alternative translations to create this challenge set. We only consider language pairs for which we can use a spacy NER model on the target side, which results in 42 language pairs.

Unit Conversion: using FLORES-101 dataset, where we replace unit mentions in the reference (e.g. 100 feet) with a different unit and corresponding amount in the "good" translation (e.g. 30.5 metres) and either the wrong amount (e.g. 100 metres) or wrong unit (30.5 feet) compared to the reference in the "incorrect" translation.

Nonsense Words: We develop a challenge set for evaluating hallucinations at subword level (Sennrich et al., 2016). 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 by using the multilingual BERT tokenizer (Devlin et al., 2019). 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.

Real Data Hallucinations: To also create a more realistic hallucination benchmark, we manually check some machine 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.

4.2.3.3 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). Similarly, existing surface-level metrics rely on n-gram matching between the hypothesis and the reference. We create this challenge set to test if metrics can reliably identify an incorrect translation especially when it shares a high degree of lexical overlap with the reference. 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.

4.2.3.4 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. We focus on the English modal auxiliary verbs: "must" (necessity), and "may", "might", "could" (possibility). 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. We use a combination of the FLORES-200 and PAWS-X datasets as the basis of the challenge sets.

4.2.3.5 Mistranslation - Overly Literal Translations

MQM defines this error type as translations that are overly literal, for example, literal translations of figurative language. We create two challenge sets based on this error type:

Idioms: We create this challenge set based on the PIE[5] 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.

Real Data Errors: For this challenge set, we manually check MT translations of the

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

FLORES-101 datasets. If we find an overly-literal translation, we manually correct it to form the good translation and use the overly-literal translation as the incorrect translation.

4.2.3.6 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, for a given premise. Thus, the premise and hypothesis have substantial overlap but they vary in meaning. We use the XNLI dataset to create such examples where there is at least a 0.5 chrF score between the English premise and hypothesis only for the neutral and contradiction examples. We use either the premise/hypothesis as the reference, an automatic translation as the "good" translation, premise/hypothesis from the remaining non-English languages, and hypothesis/premise as the "incorrect" translation.

4.2.3.7 Mistranslation - Ordering Mismatch

We also investigate the effects of changing word order in a way that changes meaning. For example, "I like apple pie and fried chicken" is changed to " I like chicken pie and fried apple" to form the incorrect translation. This challenge set is created manually by changing translations from the FLORES-101 dataset and covers de→en, en→de and fr→de.

4.2.4 Mistranslation - Discourse-level Errors

We introduce a new subclass of mistranslation errors that specifically cover discourse-level phenomena. We create several challenge sets based on discourse-level errors:
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 focus on 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*. 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.

Discourse Connectives: We leverage the Europarl ConcoDisco 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.

Commonsense Co-Reference Disambiguation: We use the English sentences in the Wino-X challenge set which 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.

4.2.5 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". We create two challenge sets based on untranslated content errors:

Word-Level: We manually annotate real errors in translations of the FLORES-101 dev and devtest sets. We count complete copies as untranslated content as well as content that comes from the source language but was only adapted to look more like the target language.

Sentence-Level: We create a challenge set for untranslated sentences 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.

4.2.6 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. 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.

4.2.7 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). A randomly selected noun from the reference translation is replaced by its corresponding hypernym or hyponym, by using Wordnet to simulate undertranslation or overtranslation errors, respectively.

4.2.8 Real-world Knowledge

We propose a new error category where translations disagree with real-world knowledge in addition to the accuracy categories in MQM. We create five challenge sets based on this error type. For the first four, we manually construct examples each for en→de and de→en. 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.

Textual Entailment: We construct examples for which the good translation entails the meaning of the original sentence (and its reference). For example, we use the entailment *was murdered* → *died* (i.e. if a person is murdered then they must have died) to construct the good translation in the example above. We construct the incorrect translation by replacing the entailed predicate (*died*) with a related but non-entailed predicate (here *was attacked*) – a person may have been murdered without being attacked like poisoned.

Hypernyms and Hyponyms: We consider a translation that contains a *hypernym* of a word to be better than one that contains a *hyponym*. For example, whilst translating "Hund" ("dog") with the broader term "animal" results in some loss of information, this is preferable over hallucinating information by using a more specific term such as "labrador" (i.e. an instance of the hyponym class "dog"). We used Wordnet and WordRel.com[6] (an online dictionary of words' relations) to identify hypernyms and

[6] https://wordrel.com/

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. This category is different from the two categories in Section 4.2.7 as the good translation is still a paraphrase of the reference (no loss of information) while the incorrect translation is created by manipulating the reference. We illustrate the difference across these three categories in the following example:

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**.
Overtranslation: ✓:	Bob and Ted were brothers, and Ted is John's **son**.
✗:	Bob and Ted were brothers. Ted is John 's **male offspring**.
Undertranslation: ✓:	Bob and Ted were brothers, and Ted is John's **son**.
✗:	Bob and Ted were brothers. Ted is John 's **child**.
Hypernyms and Hyponyms ✓:	Bob and Ted were brothers. Ted is John 's **child**.
✗:	Bob and Ted were brothers. Ted is John 's **male offspring**.

Hypernyms and Distractors: Similar to above, we construct examples in which the good translation contains a hypernym (e.g. "pet") of the word in the reference (e.g. "dog"). We form the incorrect translation by replacing the original word in the source/reference with a different member from the same class (e.g. "cat"; both cats and dogs belong to the class of pets). In Section 4.2.7, we only manipulate the reference to create an incorrect translation with the respective error.

Antonyms: We also construct 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. For verbs, we manually constructed a more challenging set of examples intended to be used to assess whether the metrics can distinguish between translations that contain a synonym versus an antonym of a given word.

Commonsense: We are also interested in whether evaluation metrics prefer translations that adhere to common sense. To test this, we remove explanatory subordinate clauses from the sources and references in the dataset described in Section 4.2.4. This guarantees that when choosing between a good and incorrect translation, the metric cannot infer the correct answer from looking at the source or the reference. 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.

4.2.9 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 using the FLORES-200 dataset 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).

4.2.10 Fluency

Although the central focus is on accuracy errors, we include a small set of fluency errors for the punctuation category. A practical reason for this is due to the licence for the TED Talks dataset that does not allow the use of partial sentences from the dataset.
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 $\rightarrow$ 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.

See Tables A.11 and A.12 in Appendix A.3.1 for further information on the distribution of examples and language pairs in ACES.

4.3 Span Annotations

To support the development of Quality Estimation and MT evaluation metrics that predict error spans, we extended the dataset to include error span annotations. Specifically, we annotate all error spans of the type denoted by the phenomenon category label, ignoring the presence of errors belonging to other categories. We therefore label only errors present in the incorrect translation, which by design contains errors of the

phenomenon category denoted by the label. We annotate spans at the word/token level similar to the MQM format (Freitag et al., 2021a) and in line with recent developments in error span prediction metrics (Perrella et al., 2022; Rei et al., 2022a). Following the WMT 2022 MQM Human Evaluation span annotation format (Freitag et al., 2022), error spans are enclosed in tags (<v> error span </ v>) denoting the start and end position of the error in the incorrect translation. Note that due to the formulation of the manual annotation guidelines (see Appendix A.3.2) it is not possible for two spans to overlap. [7]

We provide annotations for all ACES examples, using a combination of automated and manual methods. The annotation methods used for each phenomenon can be found in Appendix A.3.2. For many of the phenomena categories, we were able to automatically annotate examples using rule-based methods informed by the methodology that we followed to construct the examples. For the remaining phenomena, which we could not annotate automatically due to the manual methods used to generate the good and incorrect translations, we annotated the error spans manually (see Appendix A.3.2). We also manually annotated a small number of examples (1959 from the mistranslation phenomena and 3 from the real-world knowledge phenomena) for which the automated annotation rules failed.

4.3.1 Automatic Annotations

We automatically annotate the error spans in the incorrect translations for 34514 samples out of 36476, by performing a deterministic comparison of the incorrect translation to either the good translation or the reference sentence. The automatic annotation methods mainly depend on the way the challenge sets for each phenomenon were constructed and contain only word-level annotations following the annotation guidelines. The details about the automatic annotation methods are as follows:

Annotation of addition, omission and substitutions: This method tokenises the good translation and incorrect translation, and compares the tokens to annotate word-level addition, omission and substitutions which may occur multiple times. It is only used to annotate the simpler cases of substitutions, when each word was replaced with another word.

Annotation of substitution of a variable-sized span comparing to the correct

[7] While it is common for multiple errors to coexist within a given text span, the current design of our dataset does not deal with such errors. These type of errors would require an improved annotation schema in future.

translation: This method tokenises the good translation and the incorrect translation and then finds a single word-level error span with variable size.

Annotation of substitution of a variable-sized span comparing to the reference sentence: Similar to "Annotation of substitution of a variable-sized span comparing to the correct translation", this method tokenises the reference and the incorrect translation and then finds a single word-level error span with variable size.

Annotation of the date-time translation errors: In the Hallucination - Date-Time challenge set, the incorrect translations were built by substituting a month name in the reference with another month. This method finds the month names which are different in the incorrect translations and the reference, ignoring the months replaced with their corresponding abbreviations.

Annotation of the unit-conversion translation errors In the Hallucination - Unit Conversion phenomenon, the unit mentions in the reference (e.g. 100 feet) were replaced with either the wrong amount (e.g. 100 metres) or wrong unit (30.5 feet) in the incorrect translation. Using the Python package `quantulum3`[8], we detect the amount and units used in the incorrect translation and annotate either the wrong amount or the wrong unit, according to the phenomenon category label.

Annotation of the error where two words in the good translation were swapped In ordering-mismatch challenge set, the incorrect sentence was generated by swapping the places of two words in the good translation. This method computes the annotations when two spans were swapped, and we manually annotated 4 samples which the method was not able to correctly annotate.

Annotation of the whole sentence: This method trivially annotates the whole incorrect translation as an error. For examples belonging to the following Mistranslation - Sentence-Level Meaning Error phenomena, constructed using the XNLI dataset, we automatically mark the entire sentence as an error: xnli-addition-contradiction, xnli-addition-neutral, xnli-omission-contradiction, xnli-omission-neutral. Despite some degree of lexical overlap between the good- and incorrect-translation, the incorrect-translation is drawn from either a contradiction or neutral hypothesis in the XNLI dataset, and will therefore by definition *not be a translation* of the premise (i.e. the sentence extracted as the good-translation).

[8] `https://github.com/nielstron/quantulum3`

4.3.2 Manual Annotation

Automated annotation is suitable for many of the examples, e.g. where the good and incorrect translations only exhibit differences relevant to the particular phenomenon indicated by the category label. However, it is not suitable in all cases, for example where the good and incorrect translations contain additional differences (not related to the error phenomenon), which could result in the automatic annotation method introducing annotation errors. We identified three phenomena for which automated annotation was unsuitable, and submitted all examples from these categories for manual annotation. These included Commonsense Co-Reference Disambiguation, Real Data Hallucinations, and Lexical Overlap, which were shown to the annotators as coreference, hallucination, and overlap respectively.

We extracted a total of 2,006 examples belonging to these phenomena (427 hallucination, 559 coreference, and 1020 word swap), with examples for the following languages: English (471), French (551), German (456), Japanese (322), Korean (4), Marathi (44), and Russian (158). The manual annotation of these examples was completed by a team of seven annotators (one per language), who are either professional translators or linguists. The annotators were provided with a set of general guidelines plus specific instructions for each of the different phenomena listed above. The annotation guidelines are summarised in the following sections and the complete set of guidelines given to the annotators is provided in Span Annotation Guidelines.

Automated checks were carried out over the manual annotations to provide a basic validation. These checks were used to ensure that 1) each example had been annotated, i.e. contained at least one span of text within tags, 2) all spans were marked with an open and close tag (i.e. the number of open and close tags per example, should match), 3) no changes had been made to the example text other than the addition of the tags. Examples that failed these checks were sent to the annotators for re-annotation. We also automatically identified and resolved instances where additional whitespace was introduced (in error) at the start or end of an error span, ensuring that the annotated text and original (un-annotated) text differed only in terms of the presence/absence of error tags. Note, we could not carry out inter-annotator agreement over the manually annotated examples as we only had access to a single annotator per example. We do not have details of the budget to hire more annotators as the manual annotation was done by our colleague in a corporate organization.

Addition: a text span that is not present in sentence A is included in sentence B Sentence A: The cat is a species of small carnivorous mammal. Sentence B: The cat is a <**domestic**> species of small carnivorous mammal.
Substitution: a text span in sentence A is substituted with a different text span in sentence B Sentence A: Female domestic cats can have kittens from spring to late autumn. Sentence B: Female domestic cats can have kittens from <**May**> to <**December**>.
Deletion: a text span that is present in sentence A is omitted from sentence B Sentence A: Feral cats are domestic cats that were born in or have reverted to a wild state. Sentence B: Feral cats are domestic cats <>or have reverted to a wild state.
Reordering: a text span in sentence A that appears in a different position in sentence B Sentence A: Montreal is the second most populous city in Canada and the most populous city in the province of Quebec. Sentence B: Montreal is the <>most populous city in Canada and the <**second**> most populous city in the province of Quebec.

Table 4.2: Manual annotation guidelines: Operations for general guidelines

4.3.2.1 Overview of Annotation Guidelines

We split the annotation guidelines into a) general guidelines suitable for annotating all examples, and b) error type-specific guidelines intended for annotating specific categories. The annotators are presented with an ACES phenomenon label representing the type of error present, and two sentences: A and B, where B is the incorrect translation (i.e. contains one or more errors) and A is either the good translation or the reference (depending on the phenomenon). The annotators are asked to identify and mark *all* error spans in sentence B that belong to the error type indicated by the phenomenon label. Error spans are marked with tags (<>) at the word level, i.e. in the case that the error is a *misspelling* (e.g. "combuter" instead of "computer") the complete word (i.e. "combuter") should be marked.

General guidelines. The general guidelines may be applied for the annotation of any example in ACES. We begin by defining four possible operations to mark error spans: *addition*, *substitution*, *deletion*, and *reordering* (see Table 4.2). In simple scenarios, a single operation may be sufficient to annotate an example. In more complex scenarios multiple operations may be required.

Error type-specific guidelines: Additionally, we include specific guidelines for the annotation of three phenomenon categories: *hallucination*, *coreference*, and *word swap* (see Table 4.3). The annotation of examples belonging to these categories may be achieved by marking the presence of one or more operations. For example, the

hallucination example in Table 4.3 contains both an "addition" (i.e. <**Welsh, French,**>) and a "substitution" (i.e. Gaelic → <**Garlic**>). The three categories, for which we provide *error type-specific guidelines*, cover all of the examples submitted for manual annotation.

Hallucination: text that is not present in sentence A is observed in sentence B or a word in sentence A is replaced by a more frequent or *orthographically similar* word in sentence B Sentence A: The official languages of Scotland are: English, Scots, and Scottish Gaelic. Sentence B: The official languages of Scotland are: English, <**Welsh, French,**> Scots, and Scottish <**Garlic**>.
Coreference: a pronoun in sentence A is replaced with a (potentially) inappropriate noun-phrase in sentence B Sentence A: The cat had caught the mouse and it was trying to wriggle free. Sentence B: The cat had caught the mouse and <**the cat**> was trying to wriggle free.
Word swap: the position of a word or text span in sentence A appears swapped in sentence B Sentence A: Their music is considered by many as an alternative metal with rap metal and industrial metal influences, which according to previous interviews call themselves "murder - rock". Sentence B: Their music is considered by many as <**industrial**> metal with rap metal and <**alternative**> metal influences. According to previous interviews, they consider themselves "murder rock".

Table 4.3: Manual annotation guidelines: Error type-specific guidelines

4.3.2.2 Development of Manual Annotation Guidelines

To aid in the development and refinement of the annotation guidelines, we conducted a two-phase annotation pilot. In the first phase, we drew up the set of formal guidelines, described in Section 4.3.2.1. In the second phase, we verified the guidelines and measured inter-annotator agreement. We then asked professional annotators to complete the manual annotation of the four ACES phenomena listed above, using the guidelines.

In the first pilot phase, four of the collaborators[9] manually annotated error spans for a sample of 100 examples with English as the target language, randomly selected across all phenomena in ACES. The annotators had access to the source-language sentence, the three target-language translations: good- incorrect- and reference-translation, and the phenomenon label. We considered only the target-language side and marked one or more error spans in the incorrect translation only. We then conducted an adjudication exercise in which all four annotators manually compared the four sets of annotations for each example and discussed our strategies for annotation. From this, we derived a

[9]Two annotators for the first pilot phase are native English speakers; two are fluent English speakers

set of general guidelines to accommodate the annotation of any example in ACES. We then added specific guidelines for examples belonging to the categories: *hallucination*, *coreference*, and *word swap*.

In the second pilot phase, we verified the quality of the manual annotation guidelines. To verify the general guidelines, and provide a gold standard against which to measure the automated evaluation method, the same four annotators from the first pilot phase annotated another sample of 100 examples with English as the target language, randomly selected across all ACES phenomena. To verify the quality of the span annotations, we automatically measured inter-annotator agreement. We computed the percentage of exact matches[10] as total_exact_matches divided by total_spans_marked, i.e. where all four annotators agree on the same error span, as 81.82% (examples=100, total spans=110, exact-match spans=90), indicating high agreement[11]. We also verified the type-specific guidelines for annotating *hallucination*, *coreference*, and *word swap*. As the *coreference* category requires manual annotation in German (ACES contains only en-de examples for the *coreference-based-on-commonsense* phenomenon), and examples of the other phenomena exist for English, we asked two native German / fluent English speakers[12] to annotate a randomly selected sample of 100 examples (25 examples from each of the relevant ACES phenomenon categories). We report inter-annotator agreement of 77.40% (examples=100, total spans=146, exact-match spans=113).

In addition to measuring inter-annotator agreement, we also examined the examples where two or more annotators marked different spans. We concluded that the majority of differences arose from simple human errors as opposed to differing interpretations of the guidelines. For example, annotators sometimes accidentally marked longer spans than necessary, or marked the presence of a deletion in the wrong position. We concluded that many of these mistakes could have been avoided had the annotators carefully double-checked their annotations. We therefore added a note to the guidelines to this effect, but made no further changes to the instructions. It is also worth noting that for a handful of examples, the presence of Machine Translation led to annotators struggling to agree on a correct annotation – an issue that is not easily resolved, but is infrequent in the ACES dataset.

[10]We ignore both leading and trailing whitespace when comparing spans

[11]Highest inter-annotator agreement with three annotators: 90.48% (examples=100, total spans=105, exact-match spans=95)

[12]One annotator for the second pilot phase is also an author of this work

4.4 Summary

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. The examples were obtained by synthetic perturbation of existing references or re-purposing of academic datasets or through manual collection of erroneous examples. We extended this dataset to SPAN-ACES to include annotations of the tokens containing the mentioned MT error. These annotations were either obtained automatically or added post-hoc by experts based on their construction in ACES. We propose the adoption of both ACES and SPAN-ACES by the MT community, as a benchmark for developing Machine Translation metrics. We envisage several use cases in which the challenge sets may be employed: to profile and compare metric performance across a range of error categories, and to identify improvement/degradation in performance of successive development iterations of the same metric, discussed in detail in the next chapter.

Chapter 5

Evaluation and Meta-Evaluation with ACES

In Chapter 3, we demonstrated that segment-level singular scores for MT evaluation are not helpful. We observed a disparity in the sensitivity of downstream tasks to a given type of machine translation error. To that end, we created ACES, described in Chapter 4, which consists of 68 accuracy-based MT errors to holistically evaluate MT metrics. In this chapter, we benchmark 47 metrics submitted to the WMT 2022 and WMT 2023 metric shared tasks on ACES and conduct several analyses. We extend this evaluation to include LLM-based methods. We provide baseline results for the SPAN-ACES dataset. The contents of this chapter are published in (Amrhein et al., 2022, 2023) and are currently under review (Moghe et al., 2024).

5.1 Introduction

MT evaluation has been equally important as MT system development. Especially, in recent years, where state-of-the-art metrics are used solely to establish the superiority of one MT system over another (Marie et al., 2021). Most metrics claim their effectiveness on system-level evaluation - their ability to rank different quality MT systems on standard benchmarks such as the WMT shared tasks (Koehn and Monz, 2006; Freitag et al., 2023). However, such evaluation only offers an overview of a metric's ability to distinguish if one MT system is collectively better at generating translations than the other system. Even with segment-level evaluation documented in the WMT shared tasks where these methods are evaluated on their ability to predict the quality of translations irrespective of the systems producing them, only a correlation score per language pair is

provided. There is no mention of their robustness to specific translation errors (Freitag et al., 2021b).

As seen in the previous chapters, evaluating the quality of individual translations is important as it is useful in an online setting. We observe in Section 3.3.2 that different tasks show varying sensitivity to the same MT error. Most research in MT evaluation does not consider the inclusion of robustness to specific error types while deploying a metric. Moreover, as pointed out in Marie et al. (2021), it is debatable whether the new metrics are actively contributing to the progress in machine translation. The progress in metric research is often documented in shared tasks which change their test sets every year and often base their analyses on the metrics that participated in that year. To understand if metrics are actually making progress over time, we need a fixed benchmark to document these developments comprehensively. In the previous chapter, we created ACES consisting of 36,476 examples covering 146 language pairs on 68 different phenomena. We have carefully designed the benchmark to avoid the leader board chasing and instead provide an error analysis of the metric's failure akin to a doctor's report. We anticipate that eventually this benchmark will be overfitted and would need a considerable design overhaul.

Our benchmark is a collection of challenge sets - specialised test suites created to measure the success of systems or metrics on a particular phenomenon of interest. In this thesis, our phenomenon of interest includes segment-level translation accuracy errors - the sentence meaning changes due to the presence of the mentioned error. We included 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. For brevity, we group the 68 categories into 10 major categories and provided a weighted average score for every metric, termed as ACES-SCORE

In this chapter, we evaluate the metrics submitted to the WMT 2022 and WMT 2023 metrics shared task (Freitag et al., 2022, 2023) and a range of baseline metrics on ACES. These consist of several contemporary metrics that have shown impressive results on the system-level evaluation (Freitag et al., 2022, 2023). Despite this, we find that there is no metric that is consistently reliable across all the ACES categories. Metrics with different design strategies posses distinct strengths and weaknesses. We confirm the findings from shared tasks, that neural metrics are more robust.

By excluding subjectivity and noisy judgements from human evaluation during

the creation of the challenge set, we offer a more reliable segment-level evaluation. We find our results to be reproducible and useful to document the progress of metrics with incremental changes like change of base encoder, addition of more training data. We report incremental performance changes between metrics submitted to both 2022 and 2023. While improvements are observed for some metrics, there is a degradation in performance for other metrics. However, even for those metrics for which an overall improvement was observed, this improvement is inconsistent across the top-level categories in ACES.

In addition to the comprehensive benchmarking of metrics, we conduct several meta-evaluation analyses to find that:

1. Reference-based metrics are overly reliant on the reference and tend to disregard the information present in the source sentence.

2. Reference-based neural metrics still rely on surface-level overlap with the reference.

3. Some properties of the base model in neural metrics may cause undesirable effects on evaluation.

4. Addition of metric training data improves the effectiveness of the metrics.

The recent investigations on using Large Language Models (LLMs) for MT evaluation (Kocmi and Federmann, 2023b; Xu et al., 2023) state that these models are capable of high quality system-level evaluation, claiming state-of-the-art. Fernandes et al. (2023) demonstrate their effectiveness for segment-level evaluation. We find that these results have been demonstrated for a handful of high-resource language pairs. For more careful and thorough understanding of using LLMs for evaluation, we benchmark three LLMs of varying sizes on ACES. Benchmarking these LLMs on ACES reveals that these models perform worse than the string-overlap metrics. These results degrade further in the reference-free setting where all of the LLMs have a negative correlation across all of the ACES categories.

In the previous chapter, we proposed the use of SPAN-ACES to aid in advancing the development of MT metrics which aim to provide error-span labels over MT output in addition to scores. While some currently available MT metrics are already able to mark error spans including MATESE (Perrella et al., 2022), COMET-22 (Rei et al., 2022a) that are trained on MQM (Lommel et al., 2014) and GEMBA-MQM (Kocmi and Federmann, 2023b), AUTOMQM (Fernandes et al., 2023) that prompt LLMs to

obtain the corresponding error span, we believe that error-span labelling is an important next step in MT metric evolution. Independent challenge sets such as SPAN-ACES will be essential in driving development forward. We benchmark GEMBA-MQM (Kocmi and Federmann, 2023b)), XCOMET-XL (Guerreiro et al., 2023), and adapted versions of COMET-22 (Rei et al., 2022a) and UNITE (Wan et al., 2022) on SPAN-ACES. Our results suggest that these methods show are far from perfect on the error labelling task with highest span-F1 score reaching 26.9. These results and corresponding poor results on the ACES benchmark demonstrate that the shift towards label-based metrics requires further exploration.

We shall provide a list of metrics and methods evaluated on our datasets, their results on the benchmark, and comprehensive analyses of these results in the rest of the chapter.

	supervised	surface overlap	base-embedding	LLM-based	2022	2023
BLEU		✓			✓	✓
f101spBLEU		✓			✓	
f200spBLEU		✓			✓	✓
chrF		✓			✓	✓
BERTScore			?		✓	✓
BLEURT20	WMT human eval		BERT		✓	✓
COMET-20	WMT human eval		XML-R		✓	
COMET-QE	WMT human eval		XML-R		✓	
YiSi-1			?		✓	✓
Random-sysname						✓

Table 5.1: Baseline metrics from WMT 2022 and 2023 Metrics shared tasks. ? indicates no information was made available.

5.2 Evaluation Methodology

In this section, we discuss the various metrics or derived methods for performing MT evaluation. We also discuss the evaluation of metrics on the ACES dataset using Kendall's tau-like correlation.

5.2.1 Metrics submitted to WMT shared tasks

The tables 5.1 and 5.2 list the baseline, reference-based, and reference-free metrics from WMT 2022 and 2023 that provide segment-level judgements and cover all of the

language pairs in ACES.

We indicate whether metrics are embeddings-based with a subset of metrics using the supervision signal provided by Direct Assessment (DA) judgements from WMT (Bojar et al., 2016a) or MQM (Lommel et al., 2014) annotations, LLM-based, or rely on surface-level overlap with the reference. For embedding-based metrics we indicate the base embedding: BERT (Devlin et al., 2019), XLM-R (Conneau et al., 2020), mt5 (Xue et al., 2021), InfoXLM (Chi et al., 2021), LLaMa2 (Touvron et al., 2023), or paraphrase-multilingual-mpnet-base-V2 (Reimers and Gurevych, 2019). For details about the different design paradigms and the corresponding metrics, refer to Section 2.2.

5.2.2 LLM Metrics

Following the rapid adoption of LLM-based approaches to address a range of NLP tasks, there has also been a steady increase in the use of LLMs for MT evaluation with promising results (Xu et al., 2023; Lu et al., 2023; Kocmi and Federmann, 2023b). We note that these observations are often limited to system-level evaluation and/or high-resource language pairs. We thus intend to investigate the extent to which these LLMs can (if they can) perform MT evaluation more holistically through the ACES dataset.

We consider three variants of using LLMs for evaluation. The first one is GEMBA-DA (Kocmi and Federmann, 2023b) where the model (GPT Davinci-003 a predecessor to GPT-4 model) is prompted using a zero-shot approach to produce a translation score between 0 and 100. Note that GEMBA-DA was the precursor of the GEMBA-MQM model. For the next two methods, we considered LLAMA2 (7B) (Touvron et al., 2023) and FLAN-ALPACA-XL (Chia et al., 2023) (3B) which is Flan-T5 (Chung et al., 2022) fine-tuned on the Alpaca dataset (Taori et al., 2023). We chose LLAMA2 (7B), despite it being predominantly trained in English, to see if the accidental multilingual tokens are enough to provide multilingual evaluation. We included FLAN-ALPACA-XL as it is a *smaller* LLM and the base model was trained with multilingual data [1].

For these LLMs (FLAN-ALPACA-XL and LLAMA2), we experimented with both zero-shot and five-shot prompting. In five-shot prompting, five examples of scored translations across varying scoring ranges and language pairs were provided with the prompt. However, we found that five-shot prompting performed poorly in our initial experiments and therefore we provide only the zero-shot results. We provide the prompt

[1]We also conducted experiments on BLOOM (Scao et al., 2022) but found the majority of outputs produced by the BLOOM-7B model to be unintelligible which could not be converted into scores

templates in Appendix A.4. For the postprocessing of outputs from the above LLMs, we included the first rational number that appeared in the output from the respective models as the *score* produced by that LLM. In the scenario in which no number was found, the example was given a score of 0. In such examples, the overgenerated text generally consisted of a hallucinated example of a source-reference-translation triplet.

As ACES is a contrastive dataset, we also experimented with providing a prompt that compares the two translations, labelled A and B respectively, and instructs the LLM to select the *better* translation. However, in our initial experiments, we found that the models typically produce an option followed by the generation of both of the candidate translations. This copying of translations makes it hard to identify if the generation of the option was a result of the model performing the evaluation or an artefact of the overgeneration.

5.2.3 Metrics with error spans

In addition to the above metrics, we also conduct baseline experiments for SPAN-ACES. We include recently developed metrics that directly predict error spans while generating the scores, namely XCOMET-XL (Guerreiro et al., 2023) and GEMBA-MQM (Kocmi and Federmann, 2023b). These metrics also provide severity of the error for the predicted error span - minor, major, and critical.

Additionally, we derive baselines from existing metrics that were trained to only produce scores. We re-purpose the work in Rei et al. (2023), which included the proposal of several neural explainability methods for interpreting state-of-the-art fine-tuned neural machine translation metrics such as COMET (Rei et al., 2022a) and UNITE (Wan et al., 2022). In one of these techniques, *embed–align*, they calculate the maximum cosine similarity between each translation token embedding and the reference and/or source token embeddings (Tao et al., 2022) and assign that scalar value to each translation token. According to their observations, the model has higher attention scores on the erroneous tokens. Starting from embed-align scores attributed to each translation token, we generate error spans over the translations by marking any token which has an embed-align score higher than a constant threshold. We set the threshold that yields the span predictions with the highest Recall@K score on the WMT 2021 MQM annotations development dataset[2] (Akhbardeh et al., 2021). This method produces three different types of span predictions per metric: embed–align[mt, src],

[2]threshold=0.1 for COMET-22, threshold=0.14 for UNiTE

embed–align[mt, ref] and embed–align[mt, src; ref] using the embeddings extracted from each of the COMET-22 and UNITE models [3].

5.2.4 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[4]. The Kendall's tau-like metric (see Equation (5.1)) 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). Ties are considered as discordant. Note that a higher τ indicates a better performance and that the values can range between -1 and 1.

$$\tau = \frac{concordant - discordant}{concordant + discordant} \tag{5.1}$$

In addition, we calculate the ACES-Score, a weighted combination of the top-level categories, which allows us to identify high-level performance trends of the metrics (see Equation (5.2)). The weights correspond to the values under the MQM framework (Freitag et al., 2021a) for major (weight=5), minor (weight=1) and fluency/punctuation errors (weight=0.1). We categorise untranslated, do not translate and wrong language as minor errors due to the ease with which they can be identified with automatic language detection tools or during post-editing. We also include real-world knowledge under minor errors since we do not generally expect MT evaluation metrics to have any notion of real-world knowledge and do not wish to punish them for this. Note that the ACES-Score ranges from -29.1 (all phenomena have a correlation of -1) to 29.1 (all phenomena have a correlation of +1).

[3] We use the wmt22-comet-da version for COMET-22 and SRC+REF version for UNITE

[4] Evaluation scripts are available here: `https://github.com/EdinburghNLP/ACES`

$$\text{ACES} = sum \left\{ \begin{array}{r} 5 * \tau_{\text{addition}} \\ 5 * \tau_{\text{omission}} \\ 5 * \tau_{\text{mistranslation}} \\ 1 * \tau_{\text{untranslated}} \\ 1 * \tau_{\text{do not translate}} \\ 5 * \tau_{\text{overtranslation}} \\ 5 * \tau_{\text{undertranslation}} \\ 1 * \tau_{\text{real-world knowledge}} \\ 1 * \tau_{\text{wrong language}} \\ 0.1 * \tau_{\text{punctuation}} \end{array} \right\} \tag{5.2}$$

We discuss the evaluation on SPAN-ACES closer to its results section.

5.3 Results

We first describe the results of the metrics on the WMT shared tasks, then dig deeper into mistranslation results. We also look at performance of LLMs on ACES and then report the baseline results on SPAN-ACES.

5.3.1 Phenomena-level Results

We begin by providing a broad overview of metric performance on the different phenomena categories, before conducting more detailed analyses (see Section 5.4). We restrict the overview to the metrics which provide a) segment-level scores and b) scores for all language pairs and directions in ACES. After filtering according to these criteria, 24 metrics from 2022 remain: nine baseline, eight reference-based, and seven reference-free metrics. In 2023, 33 metrics fulfil these criteria: 10 baseline, 11 reference-based, and 12 reference-free metrics.

We first calculate Kendall's tau-like correlation scores for all of the ACES examples (see Equation (5.1)). We then report the average score overall examples belonging to each of the nine top-level accuracy categories in ACES, plus the fluency category *punctuation* (see Tables 5.3 and 5.4). We then compute the ACES-SCORE per metric.

We report an overview of the results for WMT 2022 in Table 5.3 and the results for WMT 2023 in Table 5.4. Using the ACES-Score (the final column in each of the tables)

	supervised	surface overlap	base-embedding	LLM-based	2022	2023
COMET-22*†	DA+MQM				✓	✓
metricx_xl_DA_2019	DA		mt5		✓	
metricx_xl_MQM_2020	MQM		mt5		✓	
metricx_xxl_DA_2019	DA		mt5		✓	
metricx_xxl_MQM_2020	MQM		mt5		✓	
MS-COMET-22	human judgements		mt5		✓	
UniTE					✓	
UniTE-ref †					✓	
eBLEU						✓
embed_llama			Llama 2	✓		✓
MetricX-23	DA+MQM		mT5			✓
MetricX-23-b	DA+MQM		mT5			✓
MetricX-23-c	DA+MQM		mT5			✓
partokengram_F		✓				✓
tokengram_F		✓				✓
XCOMET-Ensemble	DA+MQM		XLM-R			✓
XCOMET-XL †	DA+MQM		XLM-R			✓
XCOMET-XXL	DA+MQM		XLM-R			✓
XLsim	WMT human eval		XLM-R			✓
COMETKiwi*	DA		InfoXLM		✓	✓
Cross-QE			?		✓	
HWTSC-Teacher-Sim			paraphrase-multilingual -mpnet-base-v2		✓	
HWTSC-TLM			XLM-R		✓	
KG-BERTScore					✓	✓
MATESE-QE	MQM				✓	
MS-COMET-QE-22*					✓	✓
UniTE-src					✓	
cometoid22-wmt21	?		InfoXLM			✓
cometoid22-wmt22	?		InfoXLM			✓
cometoid22-wmt23	?		InfoXLM			✓
CometKiwi-XL			XLM-R			✓
CometKiwi-XXL			XLM-R			✓
GEMBA-MQM †				✓		✓
MetricX-23-QE	DA+MQM		mT5			✓
MetricX-23-QE-b	DA+MQM		mT5			✓
MetricX-23-QE-c	DA+MQM		mT5			✓
XCOMET-QE-Ensemble	DA+MQM		XLM-R			✓
XLsimQE	WMT human eval		XLM-R			✓

Table 5.2: Reference-based (top) and reference-free (bottom) metrics from WMT 2022 and 2023 Metrics shared tasks. * denotes a participating metric from 2022 that was used as a baseline in 2023. † denotes that metrics were used as baselines for SPAN-ACES. ? indicates no information was made available.

	addition	omission	mistrans.	untranslated	do not translate	overtrans.	undertrans.	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	-2.7
f101spBLEU	0.662	0.590	-0.084	0.660	0.940	-0.738	-0.826	-0.405	0.638	0.639	-0.1
f200spBLEU	0.664	0.590	-0.082	0.687	0.920	-0.752	-0.794	-0.394	0.658	0.648	0.1
chrF	0.642	0.784	0.162	**0.781**	**0.960**	-0.696	-0.592	-0.294	**0.691**	0.743	3.7
BERTScore	**0.880**	0.750	0.320	0.767	**0.960**	-0.110	-0.190	0.031	0.563	**0.849**	10.6
BLEURT-20	0.437	0.810	0.429	0.748	0.860	0.200	0.014	0.401	0.533	0.649	12.0
COMET-20	0.437	0.808	0.378	0.748	0.900	0.314	0.112	0.267	0.033	0.706	12.2
COMET-QE	-0.538	0.397	0.378	0.135	0.120	0.622	0.442	0.322	-0.505	0.251	6.6
YiSi-1	0.770	0.866	0.356	0.730	0.920	-0.062	-0.076	0.110	0.431	0.734	11.5
COMET-22	0.333	0.806	0.566	0.536	0.900	0.690	0.538	0.574	-0.318	0.539	16.4
metricx_xl_DA_2019	0.395	0.852	0.545	0.722	0.940	0.692	0.376	**0.740**	0.521	0.670	17.2
metricx_xl_MQM_2020	-0.281	0.670	0.523	0.579	-0.740	0.718	**0.602**	0.705	-0.126	0.445	13.1
metricx_xxl_DA_2019	0.303	0.832	0.580	0.762	0.920	0.572	0.246	0.691	0.250	0.630	15.3
metricx_xxl_MQM_2020	-0.099	0.534	0.578	0.651	0.880	**0.752**	0.552	0.712	-0.321	0.369	13.5
MS-COMET-22	-0.219	0.686	0.397	0.504	0.700	0.548	0.290	0.230	0.041	0.508	10.0
UniTE	0.439	0.876	0.501	0.571	0.920	0.496	0.302	0.624	-0.337	0.793	14.9
UniTE-ref	0.359	0.868	0.535	0.412	0.840	0.640	0.398	0.585	-0.387	0.709	15.5
COMETKiwi	0.361	0.830	**0.631**	0.230	0.780	0.738	0.574	0.582	-0.359	0.490	16.9
Cross-QE	0.163	0.876	0.546	-0.094	0.320	0.726	0.506	0.446	-0.374	0.455	14.4
HWTSC-Teacher-Sim	-0.031	0.495	0.406	-0.269	0.700	0.552	0.456	0.261	-0.021	0.271	10.1
HWTSC-TLM	-0.363	0.345	0.384	0.154	-0.040	0.544	0.474	0.071	-0.168	0.634	7.0
KG-BERTScore	0.790	0.812	0.489	-0.456	0.760	0.654	0.528	0.487	0.306	0.255	**17.5**
MS-COMET-QE-22	-0.177	0.678	0.439	0.388	0.240	0.518	0.386	0.248	-0.197	0.523	9.9
UniTE-src	0.285	**0.930**	0.599	-0.615	0.860	0.698	0.540	0.537	-0.417	0.733	15.7
Average	0.290	0.713	0.389	0.404	0.735	0.312	0.167	0.282	0.075	0.578	10.9

Table 5.3: 2022 Results. 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).

	addition	omission	mistrans.	untranslated	do not translate	overtrans.	undertrans.	real-world knowledge	wrong language	punctuation	ACES-Score
Examples	*999*	*999*	*24457*	*1300*	*100*	*1000*	*1000*	*2948*	*2000*	*1673*	
BERTscore	**0.872**	0.754	0.318	0.771	0.940	-0.186	-0.288	0.030	0.551	**0.844**	9.7
BLEU	0.742	0.427	-0.227	0.353	0.580	-0.838	-0.856	-0.768	0.660	0.704	-2.8
BLEURT-20	0.435	0.812	0.427	0.743	0.860	0.202	0.014	0.388	0.536	0.708	12.0
chrF	0.644	0.784	0.162	**0.781**	**0.960**	-0.696	-0.592	-0.294	0.693	0.773	3.7
COMET-22	0.295	0.822	0.402	0.718	0.820	0.502	0.258	0.382	0.078	0.673	13.458
CometKiwi	0.536	**0.918**	0.614	-0.105	0.520	0.766	0.604	0.577	-0.307	0.765	**17.9**
f200spBLEU	0.666	0.584	-0.082	0.680	0.920	-0.752	-0.794	-0.394	0.657	0.708	0.041
MS-COMET-QE-22	-0.179	0.674	0.440	0.394	0.300	0.524	0.382	0.262	-0.195	0.632	10.0
Random-sysname	-0.117	-0.117	-0.116	-0.083	-0.100	-0.118	-0.152	-0.245	-0.113	-0.074	-3.6
YiSi-1	0.766	0.868	0.354	0.720	0.940	-0.062	-0.076	0.110	0.421	0.763	11.5
eBLEU	0.674	0.682	0.197	0.739	0.880	-0.662	-0.684	-0.042	**0.771**	0.270	3.4
embed_llama	0.211	0.457	0.016	0.503	0.400	-0.170	-0.492	-0.165	0.154	0.476	1.054
MetricX-23	-0.027	0.568	0.578	0.473	0.800	0.790	0.586	0.766	-0.486	0.636	14.1
MetricX-23-b	-0.135	0.622	0.572	0.613	0.860	0.772	0.568	0.749	-0.444	0.532	13.8
MetricX-23-c	-0.015	0.794	0.617	0.611	0.800	0.740	0.526	**0.783**	-0.629	0.527	15.0
partokengram_F	0.087	0.191	-0.034	0.310	0.140	-0.042	-0.028	0.032	0.508	0.171	1.9
tokengram_F	0.698	0.758	0.160	0.779	**0.960**	-0.732	-0.632	-0.273	0.687	0.830	3.5
XCOMET-Ensemble	0.311	0.786	0.663	0.379	0.780	**0.794**	0.612	0.708	-0.423	0.595	17.3
XCOMET-XL	0.169	0.542	0.570	0.222	0.800	0.656	0.464	0.582	-0.367	0.220	13.3
XCOMET-XXL	-0.119	0.413	0.547	0.234	0.600	0.736	0.568	0.508	-0.507	0.509	11.6
XLsim	0.429	0.618	0.153	0.643	0.820	-0.210	-0.290	-0.044	0.392	0.753	5.4
cometoid22-wmt21	-0.339	0.658	0.493	-0.076	0.280	0.670	0.566	0.362	-0.454	0.608	10.4
cometoid22-wmt22	-0.301	0.674	0.493	-0.119	0.280	0.686	0.538	0.340	-0.472	0.599	10.534
cometoid22-wmt23	-0.253	0.702	0.502	-0.046	0.420	0.750	0.590	0.362	-0.319	0.557	11.9
CometKiwi-XL	0.239	0.828	0.624	0.239	0.440	0.762	0.560	0.563	-0.380	0.630	16.0
CometKiwi-XXL	0.361	0.828	0.653	0.414	0.320	0.774	0.560	0.683	-0.537	0.503	16.8
GEMBA-MQM	0.037	0.281	0.153	0.094	0.140	0.466	0.276	0.268	-0.150	0.015	6.4
KG-BERTScore	0.538	0.912	0.585	-0.206	0.700	0.772	0.606	0.594	-0.307	0.654	**18.0**
MetricX-23-QE	0.045	0.678	0.654	0.379	0.460	0.772	0.612	0.654	-0.702	0.226	14.6
MetricX-23-QE-b	0.027	0.760	0.663	0.489	0.480	0.758	**0.620**	0.647	-0.673	0.256	15.1
MetricX-23-QE-c	-0.115	0.664	**0.721**	0.384	0.340	0.726	0.618	0.753	-0.712	0.375	13.8
XCOMET-QE-Ensemble	0.277	0.754	0.644	0.181	0.720	0.764	0.582	0.626	-0.519	0.449	16.1
XLsimQE	0.205	0.383	0.087	-0.694	0.940	0.454	0.352	0.042	0.307	0.671	8.0
Average	0.232	0.639	0.382	0.349	0.609	0.314	0.187	0.289	-0.069	0.532	10.0

Table 5.4: 2023 Results. Average Kendall's tau-like correlation results for the ACES top-level categories and ACES-Scores (final column). Metrics are grouped into baseline (top), and participating reference-based (middle) and reference-free (bottom) metrics.

we can see at a glance that the majority of the metrics submitted to the WMT 2022 shared task outperform the baseline metrics. The same is true of the WMT 2023 metrics – except for CometKiwi, a successful submission from 2022 which was used as a baseline in 2023 – the majority of the 2023 baseline metrics are outperformed by the metrics submitted by participants. Interestingly, in both years, many reference-free metrics performed on par with reference-based metrics. This is because our challenge sets are constructed to make the reference useless (ambiguous translation, discourse connectives, *etc.,*), or misleading (hallucinations, lexical overlap, sentence-level meaning error) and thus are heavily biased towards reference-free metrics. We cannot directly compare the results from 2022 and 2023 – for a small subset (2,659; approx. 7%) of the ACES examples different results were returned in 2022 and 2023 for metrics where no changes had been made (e.g. baseline metrics such as BLEU or CometKiwi, etc). A subsequent investigation suggested that differences in the pre-processing steps by the shared task organisers in 2022 and 2023 may have led to the differences; like the handling of double quotes present in some of the ACES examples.

The best-performing metric in 2022 is a reference-free metric, namely KG-BERTSCORE, closely followed by the reference-based metric METRICX_XL_DA_2019. The best-performing metrics in 2023 are COMETKIWI (a reference-free baseline metric), and KG-BERTSCORE. Perhaps unsurprisingly, BLEU is one of the worst performing metrics, underperformed only by the random baseline, RANDOM-SYSNAME, in 2023. We caution that we developed ACES to investigate strengths and weaknesses of metrics on a phenomena level – hence, we advise the reader not to draw any conclusions based solely on the ACES-Score.

Our observations regarding the metric performance were similar for both 2022 and 2023, and the following three points hold true for both years. Firstly, we observed that metric performance 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. Secondly, while 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 5.3, 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.[5]

[5] Currently, we use average performance of the metrics on a category highlight if a category is more challenging than others. This definition of challenging category can be improved per use case.

On the other hand, for omission and do not translate (with an average Kendall tau-like correlation of > 0.7 in 2022 and > 0.6 in 2023) metric performance is generally rather high.

Thirdly, we also observe variation in terms of the performance of metrics belonging to the baseline, reference-based, and reference-free groups. For example, in both years, the baseline metrics generally 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.

While there are many similarities in the performance trends observed in 2022 and 2023, there are also some differences. In 2023, we observe that the reference-free group exhibits overall stronger performance compared with the other groups, but in particular for the *mistranslation*, *overtranslation*, *undertranslation*, and *real-world knowledge* categories. We observe that (unlike in 2022) some of the 2023 metrics perform similarly to or worse than the baseline metrics. In particular, EMBED_LLAMA and GEMBA-MQM which are designed using Large Language Models (LLMs), struggle with this challenge set. We see similar poor results in Sections 5.3.3 and 5.3.4.2 by using different prompting strategies with LLMs. This suggests that LLMs in their current form are struggling with MT evaluation and specifically in a contrastive evaluation setting. We leave exploration of better design strategies for LLMs as future work.

5.3.2 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 (*discourse*, *hallucination*, and *other*) in Table 5.5 (2022 results) and Table 5.6 (2023 results). The *discourse* sub-category includes errors involving the mistranslation of discourse-level phenomena such as pronouns and discourse connectives. *Hallucination* includes errors at the word level that could occur due to hallucination by an MT model, for example, the use of wrong units, dates, times, numbers or named entities, as well as hallucinations at the subword level that result in nonsensical words. The *other* sub-category covers all other categories of mistranslation errors including overly literal translations of idioms and the introduction of ambiguities in the translation output.

As for the results overview in Section 5.3.1, we find that performance on the different

sub-categories is variable, with no clear *winner* among the metrics in either 2022 or 2023. The results from both years 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 5.4.

5.3.3 LLM Results

We report the results of the LLM experiments described in Section 5.2.2 in Table 5.7. Overall, we find MT evaluation via LLMs a hard task in the zero-shot setup. This is also evident in the results in Section 5.3.1 where we highlight the low performance of GEMBA-MQM and EMBED-LLAMA. This is contrary to findings where LLMs show promising trends for evaluation (Fernandes et al., 2023; Kocmi and Federmann, 2023b).

We find that of the three LLMs, GEMBA-DA has better (though still poor) performance. These results worsen for the reference-less setting where most of the phenomena have a negative correlation. Despite the instructions for DA scores to be assigned using a continuous scale of 0–100, we find that the LLMs tend to produce a peaked distribution. For example, GEMBA-DA produces only seven different scores for the full set of examples. This results in a higher number of ties which get penalised in Equation (5.1). Even after instructing the LLMs to output scores within the range of 0–100, we observed instances where the LLMs produced scores beyond that range.

These results suggest that while LLMs may perform well for MT evaluation under a constrained setup, their zero-shot inference abilities for MT evaluation are far from perfect. This can be attributed to a lack of multilingual training data (Kocmi and Federmann, 2023b) as well as a limited numerical understanding of LLMs (Dziri et al., 2023). We additionally express concerns over *test-data leakage* as ACES is built on several other academic datasets (see Section 5.3.1) that may have been a part of the LLM training data (Carlini et al., 2020). We also note that these models are quite slow at inference. It takes approximately six hours to make a pass over the entire dataset using FLAN-T5-XL on a 24GB GPU, while it takes five days with two 24GB GPUs for LLAMA2 on 8-bit precision.

-	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 5.5: 2022 Results. 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.

	disco.	halluci.	other
Examples	***3698***	***10270***	***10489***
BERTscore	0.563	-0.062	0.361
BLEU	-0.042	-0.418	-0.250
BLEURT-20	0.695	0.141	0.398
chrF	0.406	-0.138	0.160
COMET-22	0.657	0.113	0.383
CometKiwi	0.779	0.465	0.580
f200spBLEU	0.095	-0.190	-0.150
MS-COMET-QE-22	0.631	0.240	0.417
Random-sysname	-0.117	-0.122	-0.111
YiSi-1	0.608	0.017	0.366
eBLEU	0.374	-0.166	0.282
embed_llama	-0.089	-0.140	0.189
MetricX-23	0.757	0.663	0.393
MetricX-23-b	0.749	0.656	0.390
MetricX-23-c	0.694	**0.755**	0.477
partokengram_F	-0.062	-0.101	0.027
tokengram_F	0.396	-0.132	0.157
XCOMET-Ensemble	**0.791**	0.566	0.626
XCOMET-XL	0.706	0.482	0.521
XCOMET-XXL	0.609	0.540	0.504
XLsim	0.217	-0.066	0.236
cometoid22-wmt21	0.782	0.286	0.400
cometoid22-wmt22	0.748	0.290	0.423
cometoid22-wmt23	0.758	0.223	0.478
CometKiwi-XL	0.752	0.501	0.602
CometKiwi-XXL	0.735	0.535	0.661
GEMBA-MQM	0.076	0.291	0.127
KG-BERTScore	0.685	0.466	0.580
MetricX-23-QE	0.728	0.604	0.628
MetricX-23-QE-b	0.694	0.617	0.666
MetricX-23-QE-c	0.747	0.659	**0.739**
XCOMET-QE-Ensemble	0.702	0.558	0.651
XLsimQE	0.053	0.050	0.134
Average	0.511	0.248	0.365

Table 5.6: 2023 Results. 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.

	GEMBA-DA		LLAMA-2 (7B)		FLAN-T5-XL + Alpaca (3B)	
	REF	QE	REF	QE	REF	QE
addition	-0.235	-0.794	-0.531	-0.587	-0.834	-0.922
mistranslation	-0.031	-0.322	-0.521	-0.552	-0.656	-0.832
real-world knowledge	0.366	0.157	-0.403	-0.6	-0.280	-0.739
untranslated	-0.334	-0.606	-0.547	-0.626	-0.529	-0.631
do not translate	-0.100	-0.840	-0.460	-0.52	-0.180	-0.500
undertranslation	0.090	-0.286	-0.512	-0.602	0.016	-0.730
overtranslation	0.472	-0.034	-0.404	-0.524	0.026	-0.744
omission	-0.281	-0.568	-0.553	-0.503	-0.848	-0.854
punctuation	-0.306	-0.355	-0.479	-0.650	-0.875	-0.924
wrong language	0.026	-0.688	-0.528	-0.483	-0.632	-0.705
ACES-Score	-0.02	-12.0	-14.6	-16.1	-13.2	-23.1

Table 5.7: Results across three LLMs on the ACES dataset: GPT-4 through GEMBA-DA, LLAMA-2, and FLAN-T5-XL fine-tuned with Alpaca. REF: Reference based, QE: Quality Estimation/Reference-free. Using zero-shot prompting on LLMs for MT evaluation has results poorer than even the surface overlap baselines in Table 5.3. This result worsens when the LLMs operate in a QE setting.

5.3.4 Results on SPAN-ACES

We first discuss the evaluation for SPAN-ACES and then report the results for the baseline methods discussed in Section 5.2.3.

5.3.4.1 Metrics for SPAN-ACES

We consider two different types of evaluation for SPAN-ACES:

Span Extraction: We first measure how well the methods that produce spans perform the task of identifying erroneous span(s) in a translation. We evaluate the predicted spans for the incorrect translation against the gold annotation. We calculate sample F1 per example where a span is considered to be a true positive if the predicted span exactly matches its ground truth and average across the dataset. We denote it as *Span-F1*. We also experimented with using partial matches between the gold error span and the predicted error span. However, using standardised tokenization based on words/sub-words/characters and then developing a threshold for partial match is not trivial and results in incorrect inflation of scores.

Contrastive Evaluation: To evaluate these methods on ACES and compare their results, we obtain span predictions for the good translation as well. We use a length heuristic where we measure the number of times the metric produces fewer spans for the good translation compared with the incorrect translation (concordant) and greater than or equal to the incorrect translation (discordant) and calculate the correlation as

	COMET-22		UNITE			XCOMET-XL		GEMBA-MQM	
	src-ref	ref	src	src-ref	ref	src	length	weight	length	weight
Span Extraction Evaluation										
Span F1	26.9	26.2	4	22.7	22.7	7.3	10.6	10.6	8.67	8.67
Contrastive Evaluation										
addition	0.598	0.477	-0.177	0.522	0.475	0.317	-0.269	-0.191	-0.077	0.103
mistranslation	-0.313	-0.364	-0.482	-0.447	-0.431	-0.308	-0.222	-0.016	0.005	0.240
real-world knowledge	-0.470	-0.501	-0.417	-0.360	-0.377	-0.279	-0.202	0.088	-0.330	0.328
untranslated	-0.641	-0.056	-0.689	-0.759	0.260	-0.910	-0.239	-0.166	-0.152	0.103
do not translate	0.500	0.340	-0.380	0.460	0.520	0.380	0.060	0.100	-0.080	0.140
undertranslation	-0.192	-0.206	-0.392	0.110	0.092	-0.220	-0.066	0.250	0.162	0.368
overtranslation	-0.144	-0.174	-0.362	0.312	0.284	-0.088	0.008	0.430	0.236	0.554
omission	-0.770	-0.842	-0.838	-0.814	-0.784	-0.700	-0.381	-0.197	0.165	0.385
punctuation	-0.385	-0.479	-0.609	-0.642	-0.574	-0.624	-0.593	-0.525	0.039	0.129
wrong language	0.406	0.289	-0.212	0.484	0.387	0.285	-0.225	-0.279	-0.132	-0.047
ACES-Score	-4.3	-5.5	-13.0	-1.8	-1.1	-5.6	-5.3	1.1	1.8	8.8

Table 5.8: Results of span-based metrics on SPAN-ACES for the tasks of span extraction and then contrastive evaluation on ACES using the predicted spans as outlined in Section 5.3.4.1. Under COMET-22 and UniTE, use of src and ref denotes if these components were used to obtain attention weights which were converted to spans. Span-F1 is only calculated for the incorrect translation. For the contrastive evaluation on ACES, all the above methods consider a candidate translation to be better than the other translation if the number of predicted spans in the former translation is less than the later, denoted by "length". For the "weight" version of XCOMET-XL and GEMBA-MQM, the labels denoting error severity of the predicted spans are converted to a weighted score. We note the derived metrics - COMET-22 and UNITE have better results on the span extraction task than the metrics designed to predict the spans. This trend flips for the contrastive evaluation. Overall, all of the methods struggle on both tasks.

described in Section 5.2.4. If the severity of errors for the predicted spans is available, we use a weighted score based on the severity label.

5.3.4.2 Results

We now report the results of different models that produce error spans (and occasionally labels) from Section 5.3.4.1 on the SPAN-ACES dataset in Table 5.8. Overall, we find that these methods perform poorly on both the error span extraction and contrastive evaluation tasks.

On the span extraction task, we find that the derived methods – COMET-22 and UNITE – i.e. using attention maps over the source/reference sentences lead to higher Span-F1 scores than either XCOMET and GEMBA-MQM which were specifically

designed to generate error spans. This adds some more evidence to the findings in Rei et al. (2023) that suggest metrics (COMET-22 and UniTE) tend to use token-level information that can be associated with tangible translation errors. Within using attention maps over the source/reference sentences for COMET-22 and UniTE, we find that the scores for the *src* only version are the worst suggesting that these metrics use very limited information from the source (as also seen in Section 5.4.1).

While using the length heuristic for the contrastive evaluation, GEMBA-MQM has better results followed by UniTE. As GEMBA-MQM and XCOMET-XL also provide labels to their predicted error spans, we also convert these labels into score based on the weights in Guerreiro et al. (2023) (critical: 10, major: 5, minor: 1), then cap the error score per sentence at 25, and finally convert the score to a value between 0 and 1. We find that weighted label scores have some improvement over the length heuristic suggesting that more sophisticated heuristics need to be developed in the future to obtain better meta-evaluation strategies. After using the label weighted score, we find that the performance for XCOMET-XL is still lower than the performance in Table 5.4, suggesting that the scores produced by the joint model may not necessarily rely on the error spans produced by that model. In contrast, GEMBA-MQM improves on its performance in Tables 5.4 and 5.8. We attribute this to either a change in the underlying model powering GPT-4 between submissions to WMT and re-running for SPAN-ACES or the use of a different weighting scheme. We also find it encouraging, that *GEMBA-MQM* improves over *GEMBA-DA*, providing us with some evidence that label-based evaluation can be helpful.

We speculate that these poor results may be attributed to (i) the unavailability of labelled MQM data during training (COMET-22 and UniTE), (ii) the availability of labelled data for only a few language pairs (XCOMET-XL), (iii) the use of proprietary models, and thus no knowledge of underlying training data (GEMBA-MQM), (iv) the fact that these metrics are the earliest designs for span-based evaluation, and (v) the fact that our annotation schemes and evaluation regimes are also the first of their kind, potentially introducing new challenges for span-based evaluation metrics. We also caution the readers that our heuristics for contrastive evaluation only offer a starting point. Future work can include model confidence, different weighting scheme, POS tags *etc.,* to compare the two translations.

5.4 Analysis

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

5.4.1 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. See an example below where the source uses a specific word *Brühe* while the reference uses an ambiguous term.

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

We present a targeted evaluation intended to provide some insights into how important the source is for different metrics. For brevity, we include the top three performing metrics in each category in 2022 and 2023, and a couple of baseline metrics. Table 5.9 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. XCOMET-ENSEMBLE performs best out of the reference-based metrics and XCOMET-QE-ENSEMBLE performs best of all reference-free metrics. It is noteworthy that there is still a considerable gap

	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-22	-0.415	0.792	**0.940**	**1.000**	-0.628	0.374	**0.558**	0.040	0.333
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
MetricX-23	-1.000	**1.000**	-0.864	-0.062	0.062	0.870	0.227	-0.222	0.001
MetricX-23-c	-0.849	0.849	-0.998	-0.581	**0.576**	**0.996**	0.150	-0.133	0.172
XCOMET-Ensemble	-0.585	0.981	0.852	0.948	0.273	0.922	0.554	**0.231**	**0.522**
Cross-QE	**0.208**	0.830	0.976	0.995	-0.337	0.364	**0.762**	0.355	0.519
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
CometKiwi	0.075	**1.000**	**0.990**	**0.998**	-0.171	0.440	0.740	0.384	0.557
KG-BERTScore	0.075	**1.000**	**0.990**	**0.998**	-0.171	0.440	0.702	0.460	0.315
MetricX-23-QE-b	-0.566	0.868	0.968	0.995	**0.722**	**0.968**	0.643	**0.490**	0.643
XCOMET-QE-Ensemble	-0.208	0.925	0.930	0.975	0.546	0.912	0.740	0.477	**0.662**

Table 5.9: 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. The best results for each phenomenon and each group of models are marked in bold and green and the average overall can be seen in the last column.

between these two models across most of the error categories, suggesting that reference-based models should pay more attention to the source when a reference is ambiguous to reach the performance of reference-free metrics.

This finding is also supported by our real-world knowledge commonsense challenge set. If we compare the scores on the examples where the subordinate clauses are missing from both the source and the reference to the ones where they are only missing from the reference, we can directly see the effect of disambiguation through the source. The corresponding correlation gains are shown in Table 5.10. 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.

Reference-based	corr-gain	Reference-free	corr-gain
BERTScore	0.002	COMET-QE	0.018
COMET-20	0.06	Cross-QE	0.292
COMET-22	0.19	HWTSC-Teacher-Sim	0.154
metricx_xxl_DA_2019	0.012	KG-BERTScore	0.154
metricx_xxl_MQM_2020	-0.016	MS-COMET-QE-22	0.196
MS-COMET-22	0.05	UniTE-src	0.216
UniTE	0.042	cometoid22-wmt23	0.138
COMET-22	0.042	CometKiwi	0.454
MetricX-23	0.004	CometKiwi-XL	0.148
MetricX-23-b	-0.002	GEMBA-MQM	**1.107**
MetricX-23-c	0.008	KG-BERTScore	0.436
XCOMET-Ensemble	**0.162**	MS-COMET-QE-22	0.198
XCOMET-XL	0.11	MetricX-23-QE-b	0.296
XCOMET-XXL	0.016	XCOMET-QE-Ensemble	0.112
		XLsimQE	0.184

Table 5.10: Results on the real-world knowledge commonsense challenge set with reference-based metrics in the left block and reference-free metrics in the right 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.

5.4.2 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 add an example about the three levels. Note that as the levels increase, the surface level similarity between the good translation and the reference decreases while the surface level overlap between the incorrect translation and the reference increases.

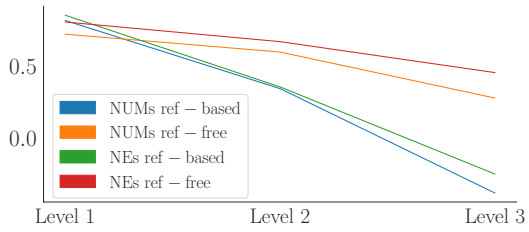

Figure 5.1: Decrease in correlation for reference-based and reference-free metrics on the named entity and number hallucination challenge sets.

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 take the average correlation for all reference-based metrics, (excluding lexical overlap metrics like BLEU) and the average correlation of all reference-free metrics that cover all languages across both the years 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 5.1.

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.

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

	reference-based	reference-free
hallucination	-0.21 $\pm$ 0.15	+0.01 $\pm$ 0.05
overly-literal	-0.32 $\pm$ 0.16	+0.07$\pm$ 0.09
untranslated	-0.43 $\pm$ 0.15	-0.00 $\pm$ 0.07

Table 5.11: Average correlation difference and standard deviation between the challenge sets with reference-copied good translations and the challenge sets with the synonymous good translations.

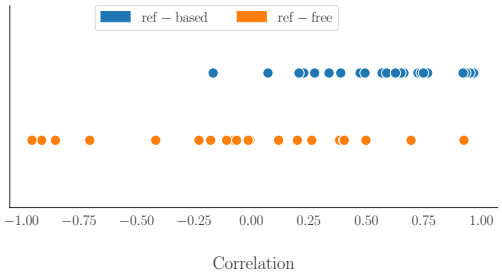

Figure 5.2: Correlation of reference-based metrics (blue) and reference-free metrics (orange) on the sentence-level untranslated test challenge set

.

but does not have the same meaning as an incorrect translation. On average, the reference-based metrics only reach a correlation of 0.05 $\pm$ 0.17 on this challenge set, whereas the reference-free metrics reach a correlation of 0.24 $\pm$ 0.17. This shows that reference-based metrics are less robust when the incorrect translation has high lexical overlap with the reference.

5.4.3 Do multilingual embeddings help design better metrics?

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

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 5.2) and wrong language phenomena (see Table 5.3).

Figure 5.2 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 - 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 5.4.2), 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 Tables 5.3, 5.6) 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 forms, 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. The human translation in a similar language (labelled as the incorrect translation) may have a closer representation to the human reference because in the MT output (labelled as the good translation) and some semantic information may be lost. We leave further investigation of this for future work.

To summarise, pre-trained multilingual models are trained without any task-specific objective. Representations from multilingual pre-trained models tend to be language agnostic causing undesirable effects on MT evaluation. This is evident especially when the translation contains words from the source sentence and when the hypothesis is predominantly in a language different than the target language.

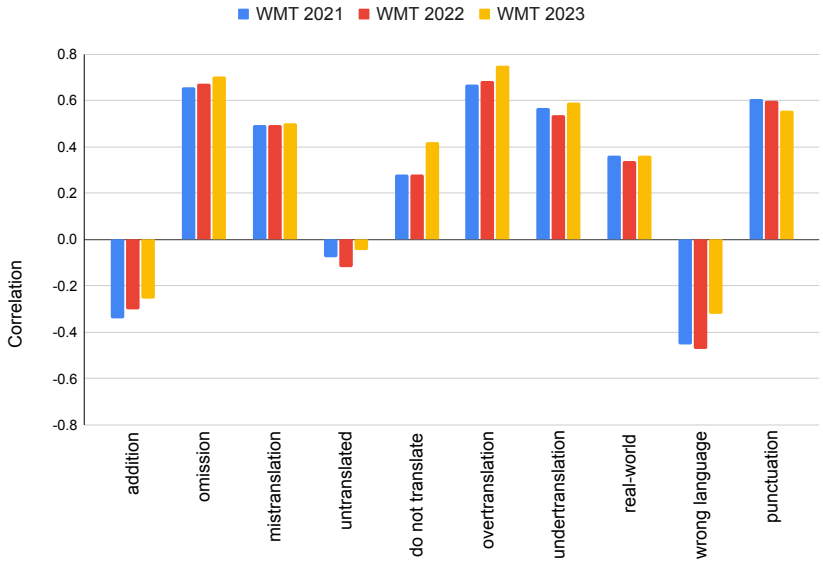

Figure 5.3: Correlations for different top-level phenomena categories with different models trained on successively more data with the COMETOID architecture. We find that adding more data helps.

5.4.4 How does metric training data size affect MT evaluation?

The COMETOID22 submission in 2023 included three different reference-free metric versions, each trained on successively more data. This allows us to investigate the effects of the metric training data size[6] on the performance on ACES. (Note that we cannot draw any conclusions about the training data size of the pretraining models that are used.) In Figure 5.3, we can see the effect of training data size on the performance on the top-level phenomena categories. COMETOID22-WMT23, the model that has seen the most data outperforms the other two metrics on almost all top-level categories. The correlation gain is especially pronounced for the *untranslated*, *do not translate* (content in the source is erroneously translated into the target language), *overtranslation* (the target translation contains more specific information than the source) and *wrong language* categories (see Table 4.1 for examples for each of the phenomena). For clearer insights as to where the performance gain comes from, we would need to analyse the training data in depth. However, it is evident from these results that more training data is beneficial for metric development. In the next section, we look at metric score changes over metric implementation cycles - where likely more than just the training data changed.

[6]Note that for COMETOID22 this is not human judgement labelled data but rather pseudo labelled data where labels come from the reference-based COMET-22 model.

5.4.5 Changes between 2022 and 2023

We compare the results of metrics submitted by the same teams in both 2022 and 2023 in Table 5.12. To make a valid comparison, we exclude the examples affected by the double quote pre-processing resulting in 33817 examples which are discussed below.

	COMETKiwi		KG-BERTScore	XCOMET		
	-XL	-XXL		-Ensemble	-XL	-XXL
addition	-0.120	-0.004	-0.251	0.595	0.455	0.142
omission	-0.004	-0.002	0.103	0.118	-0.126	-0.254
mistranslation	-0.005	0.013	0.077	0.126	0.038	0.005
untranslated	0.000	0.142	0.266	-0.181	-0.342	-0.362
do not translate	-0.395	-0.553	0.000	0.053	0.079	-0.105
overtranslation	0.027	0.035	0.119	0.073	-0.067	0.017
undertranslation	-0.019	-0.021	0.077	0.014	-0.132	-0.025
real-world knowledge	-0.020	0.100	0.107	0.003	-0.123	-0.198
wrong language	-0.014	-0.173	-0.618	-0.296	-0.232	-0.395
punctuation	-0.037	0.004	0.264	0.206	-0.144	0.006
ACES-Score	-1.04	-0.38	0.40	4.23	0.21	-1.64

Table 5.12: Comparison of average Kendall's tau-like correlation: delta calculated as 2023 score minus 2022 score.

We report changes in performance in terms of deltas, computed by subtracting the 2022 score from the 2023 score. We do this for the following pairs of metrics: KG-BERTSCORE (2022) and KG-BERTSCORE (2023); COMETKIWI (2022) paired with COMETKIWI-XL (2023) and COMETKIWI-XXL (2023); COMET-22 (2022) paired with XCOMET-ENSEMBLE (2023), XCOMET-XL (2023) and XCOMET-XXL (2023).

We observe that the performance of KG-BERTSCORE improved in 2023. From the description provided by the metric developers, the main difference is that the 2023 version of KG-BERTSCORE metric uses COMET-QE instead of BERTScore (Zhang et al., 2020) to compute the similarity between the source and the hypothesis. While we might therefore attribute the increase in performance to this change, a more systematic comparison of the two metric versions would be required to confirm whether this is the only contributing factor.

The metrics in the COMETKIWI family exhibit: a slight drop in performance (COMETKIWI-XL) and a similar performance to that of last year (COMETKIWI-XXL). The difference can be attributed to changing the underlying encoder, XLM-R XL and XLM-R XXL (Goyal et al., 2021) respectively, and the use of additional fine-tuning

data made available this year. We have seen that the addition of more training data helps in Section 5.4.4. Considering that there is no improvement in the performance, we question if an increase in the underlying model capacity of the encoder alone is useful for obtaining better MT evaluation.

Performance change for the XCOMET family is variable: there is a performance increase for XCOMET-ENSEMBLE (compared to COMET-22), for XCOMET-XL the increase is smaller, and the performance of XCOMET-XXL is degraded. The XCOMET family is designed to provide both a quality score and an error span. Considering that the metric also explains the scores without hurting the performance, this is indeed a positive change. Finally, it is worth noting that for *all* metrics in Table 5.12 a change in performance is observed for almost all ACES categories, for all metrics.

While it is not possible to draw conclusions or make predictions about the future of metric development based solely on the observations from two consecutive metrics shared tasks, we highlight several high-level changes. Firstly, we note the participation of many more COMET-based metrics in 2023, compared with 2022. This is presumably based on the success of COMET at previous shared tasks and its adoption within the MT community. We find that three metrics from 2022 were used as baseline metrics in 2023: COMET-22, COMETKIWI, and MS-COMET-QE-22. In contrast to the submissions in 2022, the 2023 submissions included some new metrics that use lexical overlap through text matching or embeddings (TOKENGRAM_F, PARTOKENGRAM_F, and EBLEU). However, their performance trend is similar to other surface overlap metrics. This year has also seen submissions based on large language models (EMBED_LLAMA and GEMBA-MQM). As seen in Section 5.3.1, their moderate performance indicates the need for more effective approaches. Additionally, we note an overall increase between 2022 and 2023 in the number of metrics submitted to WMT that a) provide segment-level scores and b) provide scores for all language pairs and directions in ACES. There were 37 segment-level metrics at WMT 2022, 24 of which covered the language pairs and directions in ACES, compared with 47 and 33, respectively in 2023. This suggests that the interest in metric development remains high, and could be increasing. In terms of metric sensitivities to error types, computed by subtracting the 2022 score from the 2023 score mostly supports the observations about the changes in the metric performances evaluated using Kendall's tau-like correlation.

In our overview analysis in Section 5.3.1 we highlight many similarities between metric performance trends in 2022 and 2023. However, there are a few differences. In 2023 we observe that the reference-free metric group performs strongly overall,

compared with the baseline and reference-free groups. This could be attributed to the increase in metrics based on the COMET architecture in 2023. WMT 2023 also saw the submission of two LLM-based metrics: EMBED_LLAMA and GEMBA-MQM. Despite the success of LLMs across various tasks (Brown et al., 2020), the performance of both EMBED_LLAMA and GEMBA-MQM highlights that leveraging LLMs to evaluate translated outputs on segment-level still requires some improved design strategies. All of these observations suggest that evaluating MT outputs is indeed a hard problem (Neubig, 2022). While we do have a good suite of metrics to provide a proxy for evaluation, there are indeed several interesting challenges that need to be tackled before we find an ideal evaluation regime. And even then, we need to continuously monitor this to ensure that we do not optimise towards metric weaknesses that we have not yet discovered.

5.5 Summary

In this chapter, we identify and address some of the shortcomings of MT metrics. A single segment-level (or system-level) score for a metric does not provide an overview of that metric's strengths and weaknesses. We demonstrated that ACES can be used to provide a profile of metric performance over a range of phenomena and to measure incremental performance between multiple versions of the same metric. We used ACES to evaluate the baseline and submitted metrics from the WMT 2022 and 2023 metrics shared tasks, to measure the incremental performance of those metrics submitted in both, and to provide fine-grained analyses of metric performance.

Our overview of metric performance at the phenomena and language levels in Section 5.3 reveals that there is no single best-performing metric. The more fine-grained analyses in Section 5.4 highlight that 1) many reference-based metrics that take the source as input do not give it sufficient prominence, 2) most reference-based metric scores are still considerably influenced by surface overlap with the reference, 3) the use of multilingual embeddings can have undesirable effects on MT evaluation and 4) the addition of metric-specific data improves the quality of the metric. We find that LLM-based evaluation methods have mediocre results and in some cases even worse than the surface overlap-based metrics. We provided baseline results on SPAN-ACES and find that MT metrics producing error spans still need considerable improvement in predicting MT errors.

Chapter 6

Conclusion and Recommendations

The chapters 3 to 5 outlined the core contributions of this thesis. We explored segment-level meta evaluation through the use of metrics in extrinsic tasks and challenge sets. We identified potential weaknesses in the design of these metrics. In this chapter, we summarise our contributions and provide recommendations for metric developers. We list some interesting directions for future work.

6.1 Summary of Contributions

The progress of machine translation research is dependent on the underlying evaluation regimes. As human evaluation is tedious, time-consuming, and expensive, the development of automatic metrics has gained a lot of traction.

These metrics often claim their effectiveness by their ability to rank MT systems according to their quality by comparing with the human evaluation of those systems. The state-of-the-art metrics often show a high correlation for such system-level evaluation on standard benchmarks like WMT for a specific set of language pairs. This has led to several recent works in MT system development claiming the superiority of their method solely through these automatic metrics (Marie et al., 2021; Kocmi et al., 2021). However, this over-reliance on automatic metrics without their thorough evaluation can lead to the development of lower-quality MT systems (Kocmi et al., 2021).

The system-level evaluation only offers an overview of the metric's ability. Even with segment-level evaluation, where metrics are assessed for their capability to predict the quality of translations from different machine translation systems, the outcomes are commonly presented as individual scores for each language pair. In both these setups, there is no fine-grained information about its robustness to different MT errors. Further,

the human evaluation used as the gold standard is not perfect due to the subjectivity of the evaluators.

This thesis focused on developing interesting meta-evaluation regimes to uncover the strengths and weaknesses of MT metrics, especially for segment-level evaluation. Throughout the thesis, we found that single summary scores produced by the metrics are unreliable and uninformative. We propose some recommendations to improve MT evaluation. We list our contributions as follows:

Task-based evaluation for MT metrics: We proposed a new setup that evaluates MT metrics in an online setting. In Chapter 3, we introduced *breakdown detection* task for evaluating MT metrics. It uses a correlation between scores produced by MT metrics with the success of extrinsic downstream tasks. The method does not depend on human judgements, making it easy to use and is reproducible. We evaluated nine different metrics on their ability to detect if a translation produced by an external MT system is useful for the downstream task. These tasks included Semantic Parsing, Question Answering, and Dialogue State Tracking. Despite their state-of-the-art performance at the system-level, we found that the segment-level scores provided by all the metrics show negligible correlation with the success/failure outcomes of the end task across different language pairs. This outcome suggested that scores produced by these metrics are uninformative, in large part due to having undefined ranges. We cannot consistently rely on the judgement produced by an MT metric for individual translations. We also found that different extrinsic tasks demonstrate varying levels of sensitivity to diverse MT errors; given that most MT metrics produce scores, it is hard to categorise the type of error just based on the predicted error. We recommend moving towards label-based evaluation for machine translation. This recommendation is in line largely with the efforts in QE literature where models are expected to produce OK/BAD tags per word and/or type of the error present in the span (Callison-Burch et al., 2012; Specia et al., 2021; Zhao et al., 2024).

Development of fine-grained evaluation benchmark: The experiments in Chapter 3 showed that a single summary score is neither informative nor indicative of the type of MT error present in the translation. We wanted to understand if metrics are robust to a wider set of accuracy errors to gauge their effectiveness in future downstream tasks. To that end, in Chapter 4, we curated 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. ACES is a contrastive dataset which evaluates a metric's ability to distinguish a good translation from an

incorrect translation. The errors range from simple perturbation of the source sentence or reference to more complex ones based on discourse and real-world knowledge. It is one of the first challenge sets for meta-evaluation. Most of the work on challenge sets for MT metrics either focuses on a handful of phenomena or a handful of language pairs that are typically high-resource. With an extensive and diverse catalogue of errors, spread across different resource language pairs, it makes a good benchmark for metric developers for identifying the properties of their metrics. When used for evaluating multiple metrics, it provides a good overview of the trends in metric design.

Meta-evaluation through ACES: We evaluated ACES on 47 metrics across different design paradigms, allowing us to conduct a comprehensive analysis. We aggregated the 68 phenomena under 10 categories providing a diagnostic report on accuracy errors for every metric. We reported this in Chapter 5. This large-scale evaluation enabled us to uncover general trends in the metric development across the years. We found that (i) metrics from distinct design families possess distinct strengths, making it clear that there is no single metric that performs the best for all the error categories, (ii) reference-based neural metrics tend to disregard the information in the source, (iii) reference-based neural metrics still rely on surface level overlap and finally, (iv) the characteristics of base pre-trained models can influence resulting metrics, occasionally leading to undesirable outcomes.

We investigated if LLMs can act as MT evaluators by benchmarking three LLM-based evaluation regimes on ACES, offering a broader evaluation than recent work which has looked at primarily high-resource language pairs. We found that LLM-based evaluation methods have mediocre results and in some cases perform even worse than the surface overlap-based metrics. We attributed these results to poor multilingual representations and a lack of numeracy skills in LLMs. Our observations are likely to spark new discussions in the overall abilities of LLMs and hopefully better design strategies to leverage LLMs for evaluation.

Moving towards label-based evaluation: We found that segment-level evaluation based on a single summary score can be unreliable and uninformative. ACES provides a set of scores per metric for different accuracy phenomena. As downstream tasks utilising machine translation show varying sensitivity to the same MT errors, metrics that can identify the relevant type of error will prove beneficial. Thus, we recommend the shift towards label-based evaluation to make evaluation more useful. Essentially, a metric should identify the erroneous spans in the translation and describe the type of error for that span. To support such development, we developed SPAN-ACES that

contains annotations of errors in the "incorrect-translation" portion of the ACES dataset. Using spans for evaluation has gained traction only in the past couple of years. We evaluated SPAN-ACES on a couple of metrics that produced spans and developed some of baselines by repurposing existing metrics that produce scores. Our results suggested that existing methods are almost completely unable to predict the relevant erroneous spans.

6.2 Recommendations

This thesis focused on identifying the strengths and weaknesses of contemporary and traditional metrics at the segment-level. Based on the observations from our work, we list some recommendations for future metric developers.

6.2.1 Using Labels for Evaluation

The observations from Section 3.3.2 and Section 3.3.3 suggested that interpreting the quality of the produced MT translation based on a single score is unreliable and difficult. Further, in Section 3.3.2, different tasks have varying tolerance to different MT errors. We recommend a departure from the direct assessment method - i,e. scoring translations on a continuous range.

Prefer MQM for Human Evaluation of MT outputs: We support using the MQM scoring scheme with expert annotators for evaluating MT outputs in line with Lommel and Melby (2018); Freitag et al. (2021a). This scoring scheme will allow the detailed breakdown of errors reducing the subjectivity in human evaluation. With explicit errors marked per MT output, future classifiers can be trained on a subset of human evaluation data containing errors most relevant to the downstream application.

MT Metrics Could Produce Labels over Scores: We recommend exploring segment-level MT evaluation as an error classification task instead of regression. Specifically, the words in the source/hypothesis should be tagged with explicit error labels. Resorting to MQM-like human evaluation will result in a rich repository of human evaluation based on an ontology of errors and erroneous spans marked across the source and hypothesis (Freitag et al., 2021a). Similarly, the post-editing datasets (Scarton et al. (2019); Fomicheva et al. (2022), *inter alia*) can act as additional training data for error spans. Recent exploration in this direction are the works by Guerreiro et al. (2023); Perrella et al. (2022) that treat MT evaluation as a sequence-tagging problem

by labelling the errors in an example. Similarly, Fernandes et al. (2023); Kocmi and Federmann (2023a) prompt an LLM to obtain the error labels. By attributing weights to the labels, these methods can generate a weighted score, offering a valuable option for numerically evaluating translations when required.

6.2.2 Metric Development

We list some recommendations for general metric design irrespective of the type of outputs (labels or scores) produced by such methods.

Pay attention to the source: Our analysis in Section 5.4.1 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.

Add diverse references during training: In Section 5.4.2, 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. Even from Section 3.3.2, we find that both the neural metric and the task-specific model are not robust to paraphrases. We expect future metrics to use more lexically diverse references through automatic paraphrasing (Bawden et al., 2020) or data augmentation during the training of neural metrics.

Check the base model: Some properties of multilingual representations, especially, being language-agnostic can result in undesirable effects on MT evaluation (Section 5.4.3) like unable to distinguish if the translation contains several untranslated terms from the source sentence. We find that LLMs are not effective segment-level MT evaluators (see Section 5.2.2), due to their lack of numeracy and multilingual abilities. Thus, we advise to carefully select the base model when developing new methods. Simple strategies to model language-specific information in the metrics could improve the robustness of the metrics to adversarial language pair attacks.

Build metric ensembles: The evaluation from Section 3.3 and section 5.3 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. We recommend innovation in the

ensemble building as simple strategies like majority voting do not lead to significant improvement as seen in Section 3.3. The recent submissions to the metrics shared task already suggest ongoing efforts in that direction as some groups submitted metrics that combined ensembles of models or multiple components (COMET-22, XCOMET MEE*).

While this thesis does not focus on analysing the efficiency of metrics in terms of memory and time, efforts in building smaller-yet-effective metrics will be appreciated.

6.3 Future Work

Our recommendations and ideas for future work are limited to the segment-level evaluation for machine translation. The field of machine translation and evaluation in general encompasses a broader scope. We now list extensions of our work to other subfields within MT and NLP.

6.3.1 Context-level MT

By segment-level MT, the work in this thesis largely focuses on sentence-level evaluation (except dialogue and paragraphs in Chapter 3). We believe that addition of contextual information during the translation and evaluation stages improves the outcomes of translation and adds rigour to the evaluation process. As seen in Läubli et al. (2018), document-level evaluation is more informative than isolated evaluation of sentences within a paragraph or document. Similarly, Post and Junczys-Dowmunt (2023) advocates the shift towards the development of document-level machine translation to improve user experience. One of the central challenges in deploying such systems is lack of systematic evaluation. Current MT metrics at document-level re-purpose existing segment-level evaluation architectures by averaging segment-level scores at the lexical or embedding level. (Vernikos et al., 2022; Deutsch et al., 2023). Their weaknesses are relatively understudied, thus making meta-evaluation at document-level a crucial problem for documenting the progress in document-level MT systems.

6.3.2 LLMs and evaluation

Our results in Section 5.3.3 suggest that LLMs are far from perfect in a contrastive setup for segment-level evaluation. The main reasons for this failure are limited numeracy skills and problematic multilingual representations. Naturally, improving the numeracy

skills of LLMs (Dziri et al., 2023) can aid in the development of score-based MT evaluators based on LLMs. As these models demonstrate better generative abilities than understanding (West et al., 2024), LLMs can be used to generate synthetic data that can be used for fine-tuning smaller or traditional MT metrics (Fernandes et al., 2023).

The effectiveness of evaluation for a particular language pair is also dependent on the representations learnt by the base models. Most LLMs are English-dominant (Touvron et al., 2023; Gao et al., 2020) leading to more faithful evaluation largely for pairs that use English in either direction (Dreano et al., 2023a). Based on evidences in Muennighoff et al. (2023); Scao et al. (2022), development of LLMs supporting richer multilingual representations leads to an improvement in several multilingual downstream tasks. We hypothesize that such development can also benefit MT evaluation reducing the over-reliance on the accidental multilingual properties of monolingual LLMs to facilitate evaluation.

Development of LLMs supporting richer multilingual representations can improve MT evaluation.

Lastly, if using LLMs for evaluation, efforts towards improving efficiency in terms of memory as well as time should be considered. Their effectiveness in terms of the number of parameters, quantization, distillation *etc* remains relatively underexplored. There have been few works for non-LLM neural metrics that focus on the development of efficient metrics (Rei et al., 2022b). Future LLM-based metrics can draw inspiration from these related works of literature to build faster MT metrics.

6.3.3 Extension to other NLG tasks

This thesis solely focuses on the evaluation of MT metrics. The development of several NLG metrics borrows inspiration from developments in MT evaluation (Graham, 2015; Krubiński and Pecina, 2022). Hence, replicating these experiments on other NLG tasks is a natural extension of this thesis.

The breakdown detection task can be useful for the meta evaluation of metrics of other NLG tasks as well. For example, a poor response from a dialogue system will not illicit any conversation from the user leading to a breakdown (Martinovski and Traum, 2003). Any metric that predicts the quality of the response can then be correlated with such a breakdown. Similarly, if the generated summary is used as an answer in a question answering setup, the metric can be correlated with the accuracy of the answer.

The recommendation of locating error spans and labelling them is applicable to

the NLG tasks as well. It will offer better explainability (Reiter, 2019) as well as the flexibility to interpret the severity of errors depending on the downstream task.

The challenge sets developed in Chapter 4 can be used to evaluate MT systems by calculating the perplexity of the two competing translations. These challenge sets can be adapted to other tasks. Hallucinations, and the use of commonsense/real-world knowledge are central to summarisation, dialogue response generation, and abstractive question answering (Dziri et al., 2021; Zhao et al., 2020; Sadat et al., 2023).

At the same time, advancements in evaluation in other NLG tasks can also benefit MT evaluation. BERTSCORE (Zhang et al., 2020)/ BLEURT (Sellam et al., 2020a) are general purpose evaluation metrics that were then appropriately modified for MT evaluation. Sai et al. (2021); Nimah et al. (2023) recommend NLG checklists during the development of metrics, similar to the use of MQM.

This thesis offers insight into the current state of segment-level evaluation for machine translation. As existing metrics are far from perfect, we hope this thesis encourages the community to introspect the drawbacks of existing metrics and move towards improved evaluation practices.

Appendix A

Additional Details

A.1 Language Codes

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

Table A.1: ISO 2-Letter language codes of the languages included in the challenge set

A.2 Appendix for Chapter 3

A.2.1 Results on the Extrinsic Tasks

We first report the results for the extrinsic task for the monolingual as well as the translate-test setup on the entire test set without any alterations. This is to demonstrate the capability of original task models. The results for Semantic Parsing with denotation

		Test Language					
Task Language	**Monolingual**	**en**	**fr**	**pt**	**es**	**de**	**zh**
en	75.6	—	59.6	38.0	58.9	61.0	55.1
fr	72.7	46.5	—	54.9	39.0	68.1	13.6
pt	72.8	54.9	59.8	—	28.1	50.9	2.4
es	63.3	51.8	50.4	42.4	—	49.3	4.7
de	75.6	50.7	67.1	47.4	43.4	—	40.8
zh	72.1	44.6	37.1	24.8	12.0	32.9	—

Table A.2: Translate-Test performance of different languages for the task of semantic parsing reported with denotation accuracy. The task languages are present per row and their respective test languages are present per column. We report the monolingual performance of the task language just before the translate-test performance. All languages have a strong monolingual performance but a considerable drop happens in the translate-test setup.

Monolingual	**Translate-Test**									
en	**ar**	**de**	**el**	**es**	**hi**	**ru**	**th**	**tr**	**vi**	**zh**
Exact Match	73.4	57.4	63.9	64.4	64.8	58.4	63.6	51.5	44.1	57.8	58.6

Table A.3: Exact Match performance of the question answering system on the XQuAD dataset. We report the monolingual task performance first followed by translate-test results on various languages.

accuracy are reported in table A.2, exact match for QA in table A.3, and exact match for DST in table A.4.

A.2.2 Fine-grained Meta Evaluation Results

We now report the per language results for the three tasks - semantic parsing, extractive question answering, dialogue state tracking in the following tables A.5 to A.9.

A.3 Appendix for Chapter 4

A.3.1 More details on challenge set construction

A.3.1.1 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

		Exact Match
Monolingual	**en**	44.2
Translate-Test	**ar**	25.9
	de	32.3
	ru	22.4
	zh	22.1

Table A.4: Results of the extrinsic task of dialogue state tracking reported with Exact Match of the predicted dialogue states. The task language performance is denoted by Monolingual followed by the respective Translate-Test results.

src	tgt	Random	BLEU	chrF	BERTScore	COMET-DA	COMET-MQM	UniTE	COMET-QE-DA	COMET-QE-MQM	UniTE-QE
	de	0.465	0.492	0.500	0.45	0.436	0.465	0.469	0.511	0.474	0.481
	fr	0.440	0.487	0.519	0.467	0.473	0.491	0.525	0.489	0.525	0.509
en	pt	0.466	0.676	0.659	0.614	0.555	0.609	0.4525	0.527	0.500	0.588
	es	0.463	0.599	0.566	0.564	0.630	0.614	0.626	0.546	0.535	0.574
	zh	0.429	0.574	0.570	0.582	0.590	0.577	0.586	0.516	0.513	0.490
	en	0.490	0.611	0.598	0.623	0.624	0.637	0.629	0.556	0.620	0.673
	fr	0.409	0.523	0.539	0.515	0.595	0.613	0.608	0.592	0.522	0.536
de	pt	0.462	0.592	0.641	0.638	0.684	0.683	0.619	0.645	0.619	0.580
	es	0.479	0.605	0.621	0.569	0.666	0.631	0.684	0.596	0.576	0.621
	zh	0.468	0.614	0.670	0.571	0.614	0.553	0.581	0.524	0.532	0.554
	en	0.489	0.595	0.590	0.607	0.630	0.606	0.628	0.597	0.574	0.588
	de	0.385	0.518	0.616	0.587	0.541	0.570	0.546	0.503	0.476	0.542
fr	pt	0.472	0.620	0.620	0.565	0.543	0.583	0.538	0.549	0.534	0.520
	es	0.492	0.462	0.613	0.512	0.627	0.648	0.574	0.594	0.568	0.573
	zh	0.384	0.641	0.702	0.666	0.667	0.658	0.661	0.521	0.502	0.575
	en	0.476	0.629	0.676	0.681	0.685	0.655	0.705	0.695	0.654	0.526
	de	0.438	0.550	0.575	0.577	0.586	0.594	0.481	0.608	0.569	0.501
pt	fr	0.458	0.546	0.603	0.488	0.599	0.495	0.574	0.574	0.545	0.645
	es	0.491	0.640	0.646	0.634	0.639	0.639	0.459	0.562	0.586	0.509
	zh	0.403	0.610	0.690	0.551	0.580	0.511	0.621	0.621	0.492	0.591
	en	0.455	0.530	0.561	0.566	0.605	0.601	0.600	0.544	0.564	0.529
	de	0.455	0.530	0.546	0.587	0.540	0.521	0.584	0.49	0.486	0.513
es	fr	0.453	0.542	0.531	0.606	0.564	0.568	0.584	0.569	0.560	0.556
	pt	0.500	0.506	0.561	0.579	0.554	0.564	0.529	0.561	0.566	0.581
	zh	0.374	0.562	0.644	0.562	0.627	0.587	0.687	0.524	0.478	0.662
	en	0.455	0.530	0.561	0.566	0.605	0.601	0.600	0.544	0.564	0.529
	de	0.455	0.530	0.546	0.587	0.540	0.521	0.584	0.490	0.486	0.513
es	fr	0.453	0.542	0.531	0.606	0.564	0.568	0.584	0.569	0.560	0.556
	pt	0.500	0.506	0.561	0.579	0.554	0.564	0.529	0.561	0.566	0.581
	zh	0.374	0.562	0.644	0.562	0.627	0.587	0.687	0.524	0.478	0.662

Table A.5: MT Metric performance on F1 for extrinsic semantic parsing (MultiATIS++SQL) with the parser trained in src language.

src	tgt	Random	BLEU	chrF	BERTScore	COMET-DA	COMET-MQM	UniTE	COMET-QE-DA	COMET-QE-MQM	UniTE-QE
	de	0.012	0.008	0.016	-0.096	-0.122	-0.000	-0.06	0.025	-0.021	-0.027
	fr	-0.043	-0.024	0.039	-0.066	-0.020	-0.001	0.050	-0.021	-0.021	0.017
en	pt	-0.067	0.353	0.328	0.231	0.201	0.228	0.114	0.089	0.209	0.187
	es	0.002	0.203	0.133	0.152	0.279	0.229	0.252	0.110	0.107	0.166
	zh	-0.090	0.152	0.146	0.173	0.187	0.188	0.172	0.060	0.035	0.078
	en	-0.003	0.226	0.210	0.251	0.263	0.328	0.303	0.161	0.250	0.349
	fr	-0.007	0.046	0.078	0.033	0.196	0.226	0.243	0.185	0.044	0.078
de	pt	-0.070	0.184	0.300	0.312	0.394	0.406	0.302	0.331	0.295	0.206
	es	-0.035	0.230	0.242	0.200	0.332	0.264	0.370	0.206	0.181	0.256
	zh	-0.063	0.241	0.340	0.150	0.242	0.124	0.258	0.054	0.088	0.112
	en	0.006	0.194	0.182	0.220	0.269	0.229	0.262	0.195	0.148	0.178
	de	-0.087	0.099	0.237	0.180	0.105	0.155	0.125	0.026	-0.043	0.086
fr	pt	-0.023	0.242	0.240	0.177	0.133	0.170	0.117	0.100	0.115	0.106
	es	-0.015	0.053	0.233	0.118	0.283	0.300	0.151	0.229	0.177	0.153
	zh	-0.116	0.311	0.413	0.373	0.365	0.347	0.390	0.143	0.051	0.248
	en	0.013	0.315	0.365	0.378	0.372	0.320	0.414	0.402	0.310	0.175
	de	-0.093	0.112	0.181	0.159	0.188	0.190	0.216	0.183	0.150	0.007
pt	fr	0.013	0.100	0.222	0.061	0.218	0.030	0.155	0.053	0.090	0.291
	es	0.009	0.286	0.293	0.278	0.278	0.288	0.142	0.076	0.243	0.025
	zh	0.061	0.221	0.449	0.253	0.161	0.048	0.242	0.000	-0.011	0.212
	en	-0.063	0.080	0.179	0.136	0.214	0.208	0.200	0.095	0.128	0.058
	de	-0.075	0.092	0.169	0.175	0.082	0.047	0.186	-0.013	-0.024	0.033
es	fr	-0.065	0.140	0.118	0.214	0.129	0.140	0.196	0.150	0.124	0.112
	pt	0.014	0.012	0.144	0.169	0.148	0.143	0.110	0.160	0.133	0.166
	zh	-0.005	0.148	0.289	0.154	0.254	0.173	0.393	0.102	0.000	0.363
	en	-0.034	0.283	0.218	0.252	0.302	0.290	0.333	0.264	0.324	0.232
	de	0.008	0.260	0.274	0.302	0.314	0.347	0.273	0.139	0.199	0.169
zh	fr	-0.045	0.204	0.238	0.343	0.330	0.247	0.328	0.222	0.259	0.287
	pt	-0.130	0.264	0.357	0.430	0.327	0.295	0.307	0.171	0.205	0.134
	es	-0.015	0.340	0.375	0.446	0.407	0.417	0.213	0.139	0.229	0.211

Table A.6: MT Metric performance on MCC for the classification task with extrinsic semantic parsing (MultiATIS++SQL) with the parser trained in src language.

Method	ar	de	el	es	hi	ru	th	tr	vi	zh
Good / Bad	592 / 264	696 / 169	701/ 170	721 / 152	631 / 241	701 / 173	539 / 323	443 / 389	616 / 251	606 / 266
Random	0.508	0.525	0.512	0.492	0.489	0.505	0.490	0.468	0.473	0.498
BLEU	0.549	0.515	0.564	0.543	0.571	0.562	0.556	0.487	0.549	0.585
chrF	0.579	0.541	0.575	0.546	0.595	0.545	0.567	0.480	0.557	0.554
BERTScore	0.569	0.538	0.586	0.523	0.604	0.528	0.561	0.523	0.580	0.535
COMET-DA	0.596	0.560	0.571	0.543	0.593	0.543	0.561	0.549	0.562	0.540
COMET-MQM	0.535	0.351	0.307	0.225	0.361	0.365	0.330	0.429	0.509	0.453
UniTE	0.370	0.479	0.343	0.314	0.308	0.519	0.366	0.438	0.282	0.326
COMET-QE-DA	0.575	0.534	0.559	0.530	0.550	0.544	0.532	0.474	0.530	0.495
COMET-QE-MQM	0.549	0.510	0.416	0.473	0.420	0.384	0.356	0.459	0.509	0.492
UniTE-QE	0.356	0.217	0.344	0.363	0.322	0.534	0.525	0.416	0.281	0.523

Table A.7: macro F1 scores for different metrics for extrinsic task of Extractive Question Answering (XQuAD dataset) where the model is trained on English. Good/Bad are the number of examples in the respective labels (Not breakdown/Breakdown) for the classification task.

Language	ar	de	el	es	hi	ru	th	tr	vi	zh
Good / Bad	592 / 264	696 / 169	701/ 170	721 / 152	631 / 241	701 / 173	539 / 323	443 / 389	616 / 251	606 / 266
Random	0.023	-0.002	-0.002	0.017	0.001	-0.002	-0.002	0.028	-0.051	-0.045
BLEU	0.135	0.048	0.142	0.098	0.162	0.125	0.128	0.097	0.108	0.171
chrF	0.160	0.083	0.172	0.092	0.202	0.106	0.162	0.000	0.173	0.119
BERTScore	0.139	0.076	0.173	0.051	0.209	0.131	0.121	0.046	0.173	0.148
COMET-DA	0.193	0.122	0.194	0.086	0.187	0.111	0.125	0.108	0.124	0.120
COMET-MQM	0.096	0.011	0.025	0.017	0.062	-0.023	-0.001	-0.050	0.079	0.054
UniTE	0.068	-0.031	-0.002	-0.014	0.043	0.047	-0.006	0.056	-0.017	-0.023
COMET-QE-DA	0.178	0.084	0.142	0.068	0.125	0.115	0.066	0.049	0.063	0.110
COMET-QE-MQM	0.099	0.050	-0.013	0.025	0.090	-0.025	0.041	-0.077	0.068	0.070
UniTE-QE	0.065	-0.031	0.012	-0.008	0.035	0.069	0.073	0.056	-0.009	-0.069

Table A.8: MCC values for different metrics for extrinsic task of Extractive Question Answering (XQuaD dataset) where the model is trained on English. Good/Bad are the number of examples in the respective labels (Not breakdown/Breakdown) for the classification task. Metrics have poor performance on the classification task as a majority report MCC ¡ 0.3

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[1].

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.

[1] `https://cloud.google.com/translate`

Language	zh		de		ar		ru	
Good / Bad	1465 / 1796		2162 / 1099		1744 / 1517		1517 / 1744	
Method	F1	MCC	F1	MCC	F1	MCC	F1	MCC
Random	0.449	-0.013	0.417	0.018	0.429	-0.018	0.454	0.004
BLEU	0.511	0.079	0.541	0.091	0.540	0.083	0.527	0.076
chrF	0.518	0.078	0.496	0.033	0.499	0.071	0.52	0.086
BERTScore	0.438	0.000	0.519	0.068	0.546	0.136	0.518	0.080
COMET-DA	0.611	0.248	0.581	0.181	0.664	0.328	0.579	0.220
COMET-MQM	0.594	0.201	0.574	0.165	0.625	0.255	0.598	0.196
UniTE	0.642	0.285	0.572	0.164	0.653	0.346	0.614	0.255
COMET-QE-DA	0.558	0.119	0.489	0.03	0.569	0.141	0.476	0.088
COMET-QE-MQM	0.545	0.132	0.552	0.106	0.574	0.195	0.574	0.148
UniTE-QE	0.566	0.183	0.552	0.114	0.628	0.258	0.603	0.215

Table A.9: MT metrics for extrinsic Dialogue State Tracking (Multi^{2}WoZ) using an English-trained state tracker. Good/Bad are the number of examples in the respective labels (Not breakdown/Breakdown) for the classification task. Reported Macro F1 scores and MCC scores quantify if the metric detects a breakdown for the extrinsic task. Metrics have negligible correlation with the outcomes of the end task.

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.

A.3.1.2 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 for which there exists a stanza parser[2] 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[3] 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

[2] https://stanfordnlp.github.io/stanza/available_models.html
[3] https://huggingface.co/facebook/m2m100_1.2B

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.

A.3.1.3 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.

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[4] because we find that this system produces gendered occupation 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.

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

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

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.

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.

A.3.1.4 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.

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 A.3.1.2. We choose all pairs with target languages for which we know the abbreviations for months[5] 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.

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

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

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[6] at the character-level and use spacy[7] for identifying named entities of type "person". To substitute whole named entities, we make use of the `names`[8] 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.

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

[6] `https://github.com/life4/textdistance`
[7] `https://spacy.io/`
[8] `https://github.com/treyhunner/names`

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 A.3.1.2. 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`[9] 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`[10] to convert unit mentions in the translation into different units. The permitted conversions are listed in Table A.10.

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.

The unit conversions permitted for the *Hallucination - Unit Conversion* challenge set are listed in Table A.10.

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

[9] `https://github.com/nielstron/quantulum3`
[10] `https://github.com/hgrecco/pint`

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

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

Area:
- square kilometres → square miles

Volume:
- barrels → gallons
- barrels → litres
- gallons → barrels
- gallons → litres

Weight:
- kilograms → grams
- kilograms → pounds
- grams → ounces
- ounces → grams

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

Table A.10: Permitted Unit Conversions

"##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.

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 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.

A.3.1.5 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).

A.3.1.6 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.

A.3.1.7 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.

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[11] 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.

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.

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

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

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 effects of entailment in Section A.3.1.13. 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.

A.3.1.8 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.

A.3.1.9 Mistranslation - Discourse-level Errors

We introduce a new subclass of mistranslation errors that specifically cover discourse-level phenomena. **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" $\rightarrow$ "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

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 A.3.1.9, 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.

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.

A.3.1.10 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.

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.

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.[12] 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.

A.3.1.11 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:

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

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.

A.3.1.12 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.

A.3.1.13 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.

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* $\rightarrow$ *died* (i.e. if a person is murdered then they must have died) to construct the good translation in the example above. We construct the incorrect translation by replacing the entailed predicate (*died*) with a related but non-entailed predicate (here *was attacked*) – a person may have been murdered without being attacked, i.e. by being poisoned for example. 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**.

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[13] (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.

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. Note the techniques in Section A.3.1.12 manipulate only the reference only to create an incorrect translation with the respective error.

Real-world Knowledge - Antonyms Similar to the generation of over- and under-translations, 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:

[13] https://wordrel.com/

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.

Real-world Knowledge - Commonsense

We are also interested in whether evaluation metrics prefer translations that adhere to common sense. To test this, we remove explanatory subordinate clauses from the sources and references in the dataset described in Section A.3.1.9. 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 A.3.1.3 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.

A.3.1.14 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".

We shall present some statistics on the ACES dataset.

Table A.11 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. Table A.12 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.

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																																					
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																																					
pt										89																																					
ro										91																																					
ru								106		472	87				42																																
sk										54																																					
sl										69																																					
sr										64																																					
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 A.11: 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-* hallucination-number-level-*	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-* xnli-omission-*	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 A.12: Collection of list of languages per phenomena

A.3.2 Span Annotation Guidelines

1. General guidelines

Your task is to annotate spans of translation errors that match a specific error type: e.g. “word swap”, or “overtranslation”. You are presented with two sentences (A and B) as well as a label denoting the error type that you should look for. You should compare translations A and B and mark any error spans of the specified type that occur in sentence B.

Please note that:

- You should annotate at the word level, not at the character level. I.e. in the case that the error is a misspelling (e.g. “combuter” instead of “computer”) the complete word (“combuter”) should be marked.
- You should ***only*** mark errors of the type specified by the error type label, and no other errors that may be present in sentence B.
- You are ***not*** required to mark any errors that may be present in sentence A.
- Whilst the majority of sentences you will encounter will be fluent, some machine-generated sentences will contain disfluencies.
- In the examples in this document, errors are highlighted in bold text to help make the examples clearer. You do ***not*** need to bold the error spans in your annotations.
- This document is intended to be comprehensive and cover the cases assigned across multiple annotators. As such, a batch that is assigned to you may contain only a subset of the error types listed in the *Error type-specific* section (below).
- You should only mark punctuation as part of error spans if it is part of the error (e.g. added as part of an addition operation or changed as part of a substitution operation).

Please read the guidelines thoroughly before you start the annotation task. Once you have finished, please make a second pass to identify and correct any mistakes that you may have made. Please also make a note of any examples that you were unsure how to annotate e.g. the example ID and a brief note.

All error spans should be marked with open and closing tags (e.g. <error span>). Errors of specific types may be formed by addition, substitution, deletion or reordering operations. For deletion operations, you should insert an empty pair of tags <> where content is missing in sentence B.

Whitespace: Error tags should ***not*** contain leading (e.g. < **error span**>) or trailing (e.g. <**error span** >) whitespace.

Addition: a text span that is not present in sentence A is included in sentence B.

> Sentence A: The cat is a species of small carnivorous mammal.
> Sentence B: The cat is a <**domestic**> species of small carnivorous mammal.

Substitution: a text span in sentence A is substituted with a different text span in sentence B.

> Sentence A: Female domestic cats can have kittens from spring to late autumn.
> Sentence B: Female domestic cats can have kittens from <**May**> to <**December**>.

Deletion: a text span that is present in sentence A is omitted from sentence B. Note that when marking a deletion, care should be taken to ensure that no extra whitespace is inserted into the sentence. Tags marking the deletion should be inserted after the space separating the two words where the deletion occurred.

> Sentence A: Feral cats are domestic cats that were born in or have reverted to a wild state.
> Sentence B: Feral cats are domestic cats <>or have reverted to a wild state.

Reordering: a text span in sentence A that appears in a different position in sentence B, as though the sentence has been reordered.

> Sentence A: Montreal is the second most populous city in Canada and the most populous city in the province of Quebec.
> Sentence B: Montreal is the <>most populous city in Canada and the <**second**> most populous city in the province of Quebec.

Note: reordering operations can be viewed as a combination of a *deletion* and an *addition* operation to change the order of elements of a sentence.

Example 1: Marking a single error span of a specified error type; ignoring other error types

In this example, the aim is to mark "overtranslation" type errors, i.e. where translation B is more specific than translation A:

> Sentence A: The festival in Houston took place in the summer.
> Sentence B: The festival in took place in August.

The error span is "August", which is more specific than "the summer" - the information that the event took place in August has been "hallucinated".

> Annotated B: The Republican National Convention in was in <**August**>.

Note that the missing information in sentence B ("Houston") can be ignored because it is an "omission" error not an "overtranslation" error. Other examples of errors that can be ignored include e.g. agreement errors in German.

Example 2: Marking multiple error spans in the same example

If there are multiple errors of the specified type present in sentence B, you should mark each error span individually. For example, if the error label is "omission" you should mark the two spans of omitted text in sentence B:

> Sentence A: Like the other planets in the Solar System, Mars was formed 4.5 billion years ago.
> Sentence B: Like the other planets, Mars was formed 4.5 years ago.
>
> Annotated B: Like the other planets <>, Mars was formed 4.5 <>years ago.

2. Error type-specific guidelines

In your annotations, you will only encounter three specific error types. Additional guidelines are provided below for these error types - hallucination, word swap and coreference.

Hallucination

In a *hallucination* example, text that is not present in sentence A is observed in sentence B or word in sentence A is replaced by a more frequent or orthographically similar word in sentence B. I.e. hallucination can be an "addition" or a "substitution" case. This may result in a change of meaning in sentence B. You should mark the "hallucinated" text in sentence B.

> Sentence A: The official languages of Scotland are: English, Scots, and Scottish Gaelic.
> Sentence B: The official languages of Scotland are: English, Welsh, French, Scots, and Scottish Garlic.

The information that Welsh and French are official languages of Scotland has been hallucinated and inserted into sentence B. Additionally, "Gaelic" has been hallucinated as "Garlic". This should be annotated as:

> Annotated B: The official languages of Scotland are: English, <**Welsh, French,**> Scots, and Scottish <**Garlic**>.

Word Swap

In a *word swap* example the position of a word or a span of text in sentence A appears swapped in sentence B. This may result in sentence B being factually incorrect. You should mark (in sentence B) the spans of text that have been swapped.

> Sentence A: Their music is considered by many as an alternative metal with rap metal and industrial metal influences, which according to previous interviews call themselves "murder - rock".
> Sentence B: Their music is considered by many as industrial metal with rap metal and alternative metal influences. According to previous interviews, they consider themselves "murder rock".

The position of the words "alternative" and "industrial" is different in sentence A, compared with sentence B and should be annotated as follows:

> Annotated B: Their music is considered by many as <**industrial**> metal with rap metal and <**alternative**> metal influences. According to previous interviews, they consider themselves "murder rock".

Coreference

In a *coreference* example a pronoun in sentence A is replaced with a (potentially) inappropriate noun-phrase in sentence B. You should mark the relevant noun-phrase in

sentence B.

Example:

> Sentence A: The cat had caught the mouse and it was trying to wriggle free.
> Sentence B: The cat had caught the mouse and the cat was trying to wriggle free.

The pronoun "it" has been replaced with the noun-phrase "the cat", resulting in a change in meaning. This should be annotated as:

> Annotated B: The cat had caught the mouse and <**the cat**> was trying to wriggle free.

The methods used to annotate the error spans for each of the phenomena in SPAN-ACES are listed in Table A.13.

A.4 Appendix for Chapter 5

For reference-based evaluation, we used the following prompt in Appendix A.4. For reference-free evaluation, we excluded the "with respect to human reference" and "Human Reference" from the prompt.:

```
Score the following translation with respect to human reference on a continuous scale of 0 to 100 where
score of zero means "no meaning preserved" and score of one hundred means "perfect meaning and grammar".
Only output an integer between 0 to 100.

Source: <source sentence here>
Human Reference: <reference sentence here>
Translation: <candidate translation>
```

Figure A.1: The prompt for using LLMs as MT evaluators for the experiments in Section 5.3.3

Phenomenon	Annotation Method
addition	addition/omissions
mistranslation: ambiguous-translation	word-lvl-compare-to-good
mistranslation: discourse, pronouns	addition/omissions
antonym-replacement	word-lvl-compare-to-ref
commonsense-only-ref-ambiguous	word-lvl-compare-to-good
commonsense-src-and-ref-ambiguous	word-lvl-compare-to-good
untranslated-sent	whole-sentence
coreference-based-on-commonsense	manual
do-not-translate	word-lvl-compare-to-good
hallucination-date-time	date-time
hallucination-named-entity-level-*	word-lvl-compare-to-good
hallucination-number-level-*	word-lvl-compare-to-good
hallucination-real-data-vs-ref-word	manual
hallucination-real-data-vs-synonym	manual
hallucination-unit-conversion-amount-matches-ref	unit-conversion
hallucination-unit-conversion-unit-matches-ref	unit-conversion
undertranslation	word-lvl-compare-to-ref
overtranslation	word-lvl-compare-to-ref
lexical-overlap	manual
modal_verb:deletion	addition/omissions
modal_verb:substitution	word-lvl-compare-to-good
nonsense	word-lvl-compare-to-ref
omission	addition/omissions
ordering-mismatch	word-swap
overly-literal-vs-correct-idiom	word-lvl-compare-to-good
overly-literal-vs-explanation	word-lvl-compare-to-good
overly-literal-vs-ref-word	word-lvl-compare-to-good
overly-literal-vs-synonym	word-lvl-compare-to-good
pleonastic_it:deletion	addition/omissions
pleonastic_it:substitution	addition/omissions
punctuation:deletion_all	addition/omissions
punctuation:deletion_commas	addition/omissions
punctuation:deletion_quotes	addition/omissions
punctuation:statement-to-question	addition/omissions
real-world-knowledge-entailment	word-lvl-compare-to-good
real-world-knowledge-hypernym-vs-distractor	word-lvl-compare-to-good
real-world-knowledge-hypernym-vs-hyponym	word-lvl-compare-to-good
real-world-knowledge-synonym-vs-antonym	word-lvl-compare-to-good
wrong-language	whole-sentence
untranslated-vs-ref-word	word-lvl-compare-to-good
untranslated-vs-synonym	word-lvl-compare-to-good
xnli-addition-contradiction	whole-sentence
xnli-addition-neutral	whole-sentence
xnli-omission-contradiction	whole-sentence
xnli-omission-neutral	whole-sentence

Table A.13: Methods used to annotate the error spans for each of the phenomena in SPAN-ACES

Bibliography

Agarwal, M., Agrawal, S., Anastasopoulos, A., Bentivogli, L., Bojar, O., Borg, C., Carpuat, M., Cattoni, R., Cettolo, M., Chen, M., Chen, W., Choukri, K., Chronopoulou, A., Currey, A., Declerck, T., Dong, Q., Duh, K., Estève, Y., Federico, M., Gahbiche, S., Haddow, B., Hsu, B., Mon Htut, P., Inaguma, H., Javorský, D., Judge, J., Kano, Y., Ko, T., Kumar, R., Li, P., Ma, X., Mathur, P., Matusov, E., McNamee, P., P. McCrae, J., Murray, K., Nadejde, M., Nakamura, S., Negri, M., Nguyen, H., Niehues, J., Niu, X., Kr. Ojha, A., E. Ortega, J., Pal, P., Pino, J., van der Plas, L., Polák, P., Rippeth, E., Salesky, E., Shi, J., Sperber, M., Stüker, S., Sudoh, K., Tang, Y., Thompson, B., Tran, K., Turchi, M., Waibel, A., Wang, M., Watanabe, S., and Zevallos, R. (2023). FINDINGS OF THE IWSLT 2023 EVALUATION CAMPAIGN. In Salesky, E., Federico, M., and Carpuat, M., editors, *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, pages 1–61, Toronto, Canada (in-person and online). Association for Computational Linguistics.

Akhbardeh, F., Arkhangorodsky, A., Biesialska, M., Bojar, O., Chatterjee, R., Chaudhary, V., Costa-jussa, M. R., España-Bonet, C., Fan, A., Federmann, C., Freitag, M., Graham, Y., Grundkiewicz, R., Haddow, B., Harter, L., Heafield, K., Homan, C., Huck, M., Amponsah-Kaakyire, K., Kasai, J., Khashabi, D., Knight, K., Kocmi, T., Koehn, P., Lourie, N., Monz, C., Morishita, M., Nagata, M., Nagesh, A., Nakazawa, T., Negri, M., Pal, S., Tapo, A. A., Turchi, M., Vydrin, V., and Zampieri, M. (2021). Findings of the 2021 conference on machine translation (WMT21). In Barrault, L., Bojar, O., Bougares, F., Chatterjee, R., Costa-jussa, M. R., Federmann, C., Fishel, M., Fraser, A., Freitag, M., Graham, Y., Grundkiewicz, R., Guzman, P., Haddow, B., Huck, M., Yepes, A. J., Koehn, P., Kocmi, T., Martins, A., Morishita, M., and Monz, C., editors, *Proceedings of the Sixth Conference on Machine Translation*, pages 1–88, Online. Association for Computational Linguistics.

Alves, D., Rei, R., Farinha, A. C., C. de Souza, J. G., and Martins, A. F. T. (2022). Robust MT evaluation with sentence-level multilingual augmentation. In Koehn, P., Barrault, L., Bojar, O., Bougares, F., Chatterjee, R., Costa-jussà, M. R., Federmann, C., Fishel, M., Fraser, A., Freitag, M., Graham, Y., Grundkiewicz, R., Guzman, P., Haddow, B., Huck, M., Jimeno Yepes, A., Kocmi, T., Martins, A., Morishita, M., Monz, C., Nagata, M., Nakazawa, T., Negri, M., Névéol, A., Neves, M., Popel, M., Turchi, M., and Zampieri, M., editors, *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 469–478, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Amrhein, C., Moghe, N., and Guillou, L. (2022). ACES: Translation accuracy challenge sets for evaluating machine translation metrics. In Koehn, P., Barrault, L., Bojar, O., Bougares, F., Chatterjee, R., Costa-jussà, M. R., Federmann, C., Fishel, M., Fraser, A., Freitag, M., Graham, Y., Grundkiewicz, R., Guzman, P., Haddow, B., Huck, M., Jimeno Yepes, A., Kocmi, T., Martins, A., Morishita, M., Monz, C., Nagata, M., Nakazawa, T., Negri, M., Névéol, A., Neves, M., Popel, M., Turchi, M., and Zampieri, M., editors, *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 479–513, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Amrhein, C., Moghe, N., and Guillou, L. (2023). ACES: Translation accuracy challenge sets at WMT 2023. In Koehn, P., Haddow, B., Kocmi, T., and Monz, C., editors, *Proceedings of the Eighth Conference on Machine Translation*, pages 695–712, Singapore. Association for Computational Linguistics.

Amrhein, C. and Sennrich, R. (2022). Identifying weaknesses in machine translation metrics through minimum Bayes risk decoding: A case study for COMET. In He, Y., Ji, H., Li, S., Liu, Y., and Chang, C.-H., editors, *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1125–1141, Online only. Association for Computational Linguistics.

Artetxe, M., Ruder, S., and Yogatama, D. (2020). On the cross-lingual transferability of monolingual representations. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637, Online. Association for Computational Linguistics.

Avramidis, E. and Macketanz, V. (2022). Linguistically motivated evaluation of machine translation metrics based on a challenge set. In Koehn, P., Barrault, L., Bojar, O., Bougares, F., Chatterjee, R., Costa-jussà, M. R., Federmann, C., Fishel, M., Fraser, A., Freitag, M., Graham, Y., Grundkiewicz, R., Guzman, P., Haddow, B., Huck, M., Jimeno Yepes, A., Kocmi, T., Martins, A., Morishita, M., Monz, C., Nagata, M., Nakazawa, T., Negri, M., Névéol, A., Neves, M., Popel, M., Turchi, M., and Zampieri, M., editors, *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 514–529, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Avramidis, E., Manakhimova, S., Macketanz, V., and Möller, S. (2023). Challenging the state-of-the-art machine translation metrics from a linguistic perspective. In Koehn, P., Haddow, B., Kocmi, T., and Monz, C., editors, *Proceedings of the Eighth Conference on Machine Translation*, pages 713–729, Singapore. Association for Computational Linguistics.

Bansal, S. (2019). *Low-resource speech translation*. PhD thesis, University of Edinburgh, UK.

Bawden, R., Zhang, B., Yankovskaya, L., Tättar, A., and Post, M. (2020). A study in improving BLEU reference coverage with diverse automatic paraphrasing. In Cohn, T., He, Y., and Liu, Y., editors, *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 918–932, Online. Association for Computational Linguistics.

Bentivogli, L., Bisazza, A., Cettolo, M., and Federico, M. (2016). Neural versus phrase-based machine translation quality: a case study. In Su, J., Duh, K., and Carreras, X., editors, *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 257–267, Austin, Texas. Association for Computational Linguistics.

Berka, J., Cerný, M., and Bojar, O. (2011). Quiz-based evaluation of machine translation. *Prague Bull. Math. Linguistics*, 95:77–86.

Blain, F., Zerva, C., Ribeiro, R., Guerreiro, N. M., Kanojia, D., C. de Souza, J. G., Silva, B., Vaz, T., Jingxuan, Y., Azadi, F., Orasan, C., and Martins, A. (2023). Findings of the WMT 2023 shared task on quality estimation. In Koehn, P., Haddow, B.,

Kocmi, T., and Monz, C., editors, *Proceedings of the Eighth Conference on Machine Translation*, pages 629–653, Singapore. Association for Computational Linguistics.

Blatz, J., Fitzgerald, E., Foster, G., Gandrabur, S., Goutte, C., Kulesza, A., Sanchis, A., and Ueffing, N. (2004). Confidence estimation for machine translation. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, pages 315–321, Geneva, Switzerland. COLING.

Bojar, O., Chatterjee, R., Federmann, C., Graham, Y., Haddow, B., Huck, M., Jimeno Yepes, A., Koehn, P., Logacheva, V., Monz, C., Negri, M., Névéol, A., Neves, M., Popel, M., Post, M., Rubino, R., Scarton, C., Specia, L., Turchi, M., Verspoor, K., and Zampieri, M. (2016a). Findings of the 2016 conference on machine translation. In Bojar, O., Buck, C., Chatterjee, R., Federmann, C., Guillou, L., Haddow, B., Huck, M., Yepes, A. J., Névéol, A., Neves, M., Pecina, P., Popel, M., Koehn, P., Monz, C., Negri, M., Post, M., Specia, L., Verspoor, K., Tiedemann, J., and Turchi, M., editors, *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 131–198, Berlin, Germany. Association for Computational Linguistics.

Bojar, O., Graham, Y., and Kamran, A. (2017). Results of the WMT17 metrics shared task. In Bojar, O., Buck, C., Chatterjee, R., Federmann, C., Graham, Y., Haddow, B., Huck, M., Yepes, A. J., Koehn, P., and Kreutzer, J., editors, *Proceedings of the Second Conference on Machine Translation*, pages 489–513, Copenhagen, Denmark. Association for Computational Linguistics.

Bojar, O., Graham, Y., Kamran, A., and Stanojević, M. (2016b). Results of the WMT16 metrics shared task. In Bojar, O., Buck, C., Chatterjee, R., Federmann, C., Guillou, L., Haddow, B., Huck, M., Yepes, A. J., Névéol, A., Neves, M., Pecina, P., Popel, M., Koehn, P., Monz, C., Negri, M., Post, M., Specia, L., Verspoor, K., Tiedemann, J., and Turchi, M., editors, *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 199–231, Berlin, Germany. Association for Computational Linguistics.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). Language models are few-shot learners. *Computing Research Repository*, arXiv:2005.14165.

Callison-Burch, C., Fordyce, C., Koehn, P., Monz, C., and Schroeder, J. (2007). (meta-) evaluation of machine translation. In Callison-Burch, C., Koehn, P., Fordyce, C. S., and Monz, C., editors, *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 136–158, Prague, Czech Republic. Association for Computational Linguistics.

Callison-Burch, C., Fordyce, C., Koehn, P., Monz, C., and Schroeder, J. (2008). Further meta-evaluation of machine translation. In Callison-Burch, C., Koehn, P., Monz, C., Schroeder, J., and Fordyce, C. S., editors, *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 70–106, Columbus, Ohio. Association for Computational Linguistics.

Callison-Burch, C., Koehn, P., Monz, C., Post, M., Soricut, R., and Specia, L. (2012). Findings of the 2012 workshop on statistical machine translation. In Callison-Burch, C., Koehn, P., Monz, C., Post, M., Soricut, R., and Specia, L., editors, *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 10–51, Montréal, Canada. Association for Computational Linguistics.

Callison-Burch, C., Osborne, M., and Koehn, P. (2006). Re-evaluating the role of Bleu in machine translation research. In McCarthy, D. and Wintner, S., editors, *11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 249–256, Trento, Italy. Association for Computational Linguistics.

Carlini, N., Tramèr, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T. B., Song, D. X., Erlingsson, Ú., Oprea, A., and Raffel, C. (2020). Extracting training data from large language models. In *USENIX Security Symposium*.

Carroll, J. B. (1966). An experiment in evaluating the quality of translations. *Mech. Transl. Comput. Linguistics*, 9(3-4):55–66.

Castilho, S. (2020). On the same page? comparing inter-annotator agreement in sentence and document level human machine translation evaluation. In Barrault, L., Bojar, O., Bougares, F., Chatterjee, R., Costa-jussà, M. R., Federmann, C., Fishel, M., Fraser, A., Graham, Y., Guzman, P., Haddow, B., Huck, M., Yepes, A. J., Koehn, P., Martins, A., Morishita, M., Monz, C., Nagata, M., Nakazawa, T., and Negri, M., editors, *Proceedings of the Fifth Conference on Machine Translation*, pages 1150–1159, Online. Association for Computational Linguistics.

Castilho, S., Moorkens, J., Gaspari, F., Calixto, I., Tinsley, J., and Way, A. (2017). Is neural machine translation the new state of the art? *The Prague Bulletin of Mathematical Linguistics*, 108:109–120.

Castilho, S., O'Brien, S., Alves, F., and O'Brien, M. (2014). Does post-editing increase usability? a study with Brazilian Portuguese as target language. In Cettolo, M., Federico, M., Specia, L., and Way, A., editors, *Proceedings of the 17th Annual Conference of the European Association for Machine Translation*, pages 183–190, Dubrovnik, Croatia. European Association for Machine Translation.

Chen, X., Wei, D., Shang, H., Li, Z., Wu, Z., Yu, Z., Zhu, T., Zhu, M., Xie, N., Lei, L., Tao, S., Yang, H., and Qin, Y. (2022). Exploring robustness of machine translation metrics: A study of twenty-two automatic metrics in the WMT22 metric task. In Koehn, P., Barrault, L., Bojar, O., Bougares, F., Chatterjee, R., Costa-jussà, M. R., Federmann, C., Fishel, M., Fraser, A., Freitag, M., Graham, Y., Grundkiewicz, R., Guzman, P., Haddow, B., Huck, M., Jimeno Yepes, A., Kocmi, T., Martins, A., Morishita, M., Monz, C., Nagata, M., Nakazawa, T., Negri, M., Névéol, A., Neves, M., Popel, M., Turchi, M., and Zampieri, M., editors, *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 530–540, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Chi, Z., Dong, L., Wei, F., Yang, N., Singhal, S., Wang, W., Song, X., Mao, X.-L., Huang, H., and Zhou, M. (2021). InfoXLM: An information-theoretic framework for cross-lingual language model pre-training. In Toutanova, K., Rumshisky, A., Zettlemoyer, L., Hakkani-Tur, D., Beltagy, I., Bethard, S., Cotterell, R., Chakraborty, T., and Zhou, Y., editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3576–3588, Online. Association for Computational Linguistics.

Chia, Y. K., Hong, P., Bing, L., and Poria, S. (2023). Instructeval: Towards holistic evaluation of instruction-tuned large language models. *Computing Research Repository*, arXiv:2306.04757.

Chung, H. W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, Y., Wang, X., Dehghani, M., Brahma, S., Webson, A., Gu, S. S., Dai, Z., Suzgun, M., Chen, X., Chowdhery, A., Castro-Ros, A., Pellat, M., Robinson, K., Valter, D., Narang, S., Mishra, G., Yu, A., Zhao, V., Huang, Y., Dai, A., Yu, H., Petrov, S., Chi, E. H., Dean,

J., Devlin, J., Roberts, A., Zhou, D., Le, Q. V., and Wei, J. (2022). Scaling instruction-finetuned language models. *Computing Research Repository*, arXiv:2210.11416.

Colman, T., Fonteyne, M., Daems, J., Dirix, N., and Macken, L. (2022). GECO-MT: The ghent eye-tracking corpus of machine translation. In Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Odijk, J., and Piperidis, S., editors, *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 29–38, Marseille, France. European Language Resources Association.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Conneau, A., Rinott, R., Lample, G., Williams, A., Bowman, S., Schwenk, H., and Stoyanov, V. (2018). XNLI: Evaluating cross-lingual sentence representations. In Riloff, E., Chiang, D., Hockenmaier, J., and Tsujii, J., editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Dale, D., Voita, E., Barrault, L., and Costa-jussà, M. R. (2023). Detecting and mitigating hallucinations in machine translation: Model internal workings alone do well, sentence similarity Even better. In Rogers, A., Boyd-Graber, J., and Okazaki, N., editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 36–50, Toronto, Canada. Association for Computational Linguistics.

Dankers, V., Lucas, C., and Titov, I. (2022). Can transformer be too compositional? analysing idiom processing in neural machine translation. In Muresan, S., Nakov, P., and Villavicencio, A., editors, *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.

Del Gaudio, R., Burchardt, A., and Lommel, A. (2015). Evaluating a machine translation system in a technical support scenario. In Hajič, J. and Branco, A., editors,

Proceedings of the 1st Deep Machine Translation Workshop, pages 39–47, Praha, Czechia. ÚFAL MFF UK.

Deutsch, D., Juraska, J., Finkelstein, M., and Freitag, M. (2023). Training and meta-evaluating machine translation evaluation metrics at the paragraph level. In Koehn, P., Haddow, B., Kocmi, T., and Monz, C., editors, *Proceedings of the Eighth Conference on Machine Translation*, pages 996–1013, Singapore. Association for Computational Linguistics.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Burstein, J., Doran, C., and Solorio, T., editors, *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.

Doherty, S. and O'Brien, S. (2009). Can MT output be evaluated through eye tracking? In *Proceedings of Machine Translation Summit XII: Posters*, Ottawa, Canada.

Dong, Q., Li, L., Dai, D., Zheng, C., Wu, Z., Chang, B., Sun, X., Xu, J., Li, L., and Sui, Z. (2023). A survey for in-context learning. *Computing Research Repository*, arXiv:2301.00234.

Doyon, J. B., Taylor, K. B., and White, J. S. (1999). Task-based evaluation for machine translation. In *Proceedings of Machine Translation Summit VII*, pages 574–578, Singapore, Singapore.

Dreano, S., Molloy, D., and Murphy, N. (2023a). Embed_Llama: Using LLM embeddings for the metrics shared task. In Koehn, P., Haddow, B., Kocmi, T., and Monz, C., editors, *Proceedings of the Eighth Conference on Machine Translation*, pages 738–745, Singapore. Association for Computational Linguistics.

Dreano, S., Molloy, D., and Murphy, N. (2023b). Tokengram_F, a fast and accurate token-based chrF++ derivative. In Koehn, P., Haddow, B., Kocmi, T., and Monz, C., editors, *Proceedings of the Eighth Conference on Machine Translation*, pages 730–737, Singapore. Association for Computational Linguistics.

Dziri, N., Lu, X., Sclar, M., Li, X. L., Jiang, L., Lin, B. Y., Welleck, S., West, P., Bhagavatula, C., Bras, R. L., Hwang, J. D., Sanyal, S., Ren, X., Ettinger, A., Harchaoui, Z.,

and Choi, Y. (2023). Faith and fate: Limits of transformers on compositionality. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Dziri, N., Madotto, A., Zaïane, O., and Bose, A. J. (2021). Neural path hunter: Reducing hallucination in dialogue systems via path grounding. In Moens, M.-F., Huang, X., Specia, L., and Yih, S. W.-t., editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2197–2214, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

ElNokrashy, M. and Kocmi, T. (2023). eBLEU: Unexpectedly good machine translation evaluation using simple word embeddings. In Koehn, P., Haddow, B., Kocmi, T., and Monz, C., editors, *Proceedings of the Eighth Conference on Machine Translation*, pages 746–750, Singapore. Association for Computational Linguistics.

Emelin, D. and Sennrich, R. (2021). Wino-X: Multilingual Winograd schemas for commonsense reasoning and coreference resolution. In Moens, M.-F., Huang, X., Specia, L., and Yih, S. W.-t., editors, *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.

Eric, M., Goel, R., Paul, S., Sethi, A., Agarwal, S., Gao, S., Kumar, A., Goyal, A., Ku, P., and Hakkani-Tur, D. (2020). MultiWOZ 2.1: A consolidated multi-domain dialogue dataset with state corrections and state tracking baselines. In Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 422–428, Marseille, France. European Language Resources Association.

Fan, A., Bhosale, S., Schwenk, H., Ma, Z., El-Kishky, A., Goyal, S., Baines, M., Celebi, O., Wenzek, G., Chaudhary, V., Goyal, N., Birch, T., Liptchinsky, V., Edunov, S., Auli, M., and Joulin, A. (2021). Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research*, 22(107):1–48.

Fernandes, P., Deutsch, D., Finkelstein, M., Riley, P., Martins, A., Neubig, G., Garg, A., Clark, J., Freitag, M., and Firat, O. (2023). The devil is in the errors: Leveraging large language models for fine-grained machine translation evaluation. In Koehn, P., Haddow, B., Kocmi, T., and Monz, C., editors, *Proceedings of the Eighth*

Conference on Machine Translation, pages 1066–1083, Singapore. Association for Computational Linguistics.

Fomicheva, M. and Specia, L. (2019). Taking MT evaluation metrics to extremes: Beyond correlation with human judgments. *Computational Linguistics*, 45(3):515–558.

Fomicheva, M., Sun, S., Fonseca, E., Zerva, C., Blain, F., Chaudhary, V., Guzmán, F., Lopatina, N., Specia, L., and Martins, A. F. T. (2022). MLQE-PE: A multilingual quality estimation and post-editing dataset. In Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Odijk, J., and Piperidis, S., editors, *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4963–4974, Marseille, France. European Language Resources Association.

Fonseca, E., Yankovskaya, L., Martins, A. F. T., Fishel, M., and Federmann, C. (2019). Findings of the WMT 2019 shared tasks on quality estimation. In Bojar, O., Chatterjee, R., Federmann, C., Fishel, M., Graham, Y., Haddow, B., Huck, M., Yepes, A. J., Koehn, P., Martins, A., Monz, C., Negri, M., Névéol, A., Neves, M., Post, M., Turchi, M., and Verspoor, K., editors, *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 1–10, Florence, Italy. Association for Computational Linguistics.

Forcada, M. L., Scarton, C., Specia, L., Haddow, B., and Birch, A. (2018). Exploring gap filling as a cheaper alternative to reading comprehension questionnaires when evaluating machine translation for gisting. In Bojar, O., Chatterjee, R., Federmann, C., Fishel, M., Graham, Y., Haddow, B., Huck, M., Yepes, A. J., Koehn, P., Monz, C., Negri, M., Névéol, A., Neves, M., Post, M., Specia, L., Turchi, M., and Verspoor, K., editors, *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 192–203, Brussels, Belgium. Association for Computational Linguistics.

Freitag, M., Foster, G., Grangier, D., Ratnakar, V., Tan, Q., and Macherey, W. (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.

Freitag, M., Mathur, N., Lo, C.-k., Avramidis, E., Rei, R., Thompson, B., Kocmi, T.,

Blain, F., Deutsch, D., Stewart, C., Zerva, C., Castilho, S., Lavie, A., and Foster, G. (2023). Results of WMT23 metrics shared task: Metrics might be guilty but references are not innocent. In Koehn, P., Haddow, B., Kocmi, T., and Monz, C., editors, *Proceedings of the Eighth Conference on Machine Translation*, pages 578–628, Singapore. Association for Computational Linguistics.

Freitag, M., Rei, R., Mathur, N., Lo, C.-k., Stewart, C., Avramidis, E., Kocmi, T., Foster, G., Lavie, A., and Martins, A. F. T. (2022). Results of WMT22 metrics shared task: Stop using BLEU – neural metrics are better and more robust. In Koehn, P., Barrault, L., Bojar, O., Bougares, F., Chatterjee, R., Costa-jussà, M. R., Federmann, C., Fishel, M., Fraser, A., Freitag, M., Graham, Y., Grundkiewicz, R., Guzman, P., Haddow, B., Huck, M., Jimeno Yepes, A., Kocmi, T., Martins, A., Morishita, M., Monz, C., Nagata, M., Nakazawa, T., Negri, M., Névéol, A., Neves, M., Popel, M., Turchi, M., and Zampieri, M., editors, *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 46–68, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Freitag, M., Rei, R., Mathur, N., Lo, C.-k., Stewart, C., Foster, G., Lavie, A., and Bojar, O. (2021b). Results of the WMT21 metrics shared task: Evaluating metrics with expert-based human evaluations on TED and news domain. In Barrault, L., Bojar, O., Bougares, F., Chatterjee, R., Costa-jussa, M. R., Federmann, C., Fishel, M., Fraser, A., Freitag, M., Graham, Y., Grundkiewicz, R., Guzman, P., Haddow, B., Huck, M., Yepes, A. J., Koehn, P., Kocmi, T., Martins, A., Morishita, M., and Monz, C., editors, *Proceedings of the Sixth Conference on Machine Translation*, pages 733–774, Online. Association for Computational Linguistics.

Gal, Y. and Ghahramani, Z. (2016). Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In Balcan, M. F. and Weinberger, K. Q., editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1050–1059, New York, New York, USA. PMLR.

Gandrabur, S. and Foster, G. (2003). Confidence estimation for translation prediction. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 95–102.

Gao, L., Biderman, S., Black, S., Golding, L., Hoppe, T., Foster, C., Phang, J., He, H.,

Thite, A., Nabeshima, N., et al. (2020). The Pile: An 800GB dataset of diverse text for language modeling. *Computing Research Repository*, arXiv:2101.00027.

Gerz, D., Su, P.-H., Kusztos, R., Mondal, A., Lis, M., Singhal, E., Mrkšić, N., Wen, T.-H., and Vulić, I. (2021). Multilingual and cross-lingual intent detection from spoken data. In Moens, M.-F., Huang, X., Specia, L., and Yih, S. W.-t., editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7468–7475, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Gladkoff, S. and Han, L. (2022). HOPE: A task-oriented and human-centric evaluation framework using professional post-editing towards more effective MT evaluation. In Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Odijk, J., and Piperidis, S., editors, *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 13–21, Marseille, France. European Language Resources Association.

Gowda, T., Kocmi, T., and Junczys-Dowmunt, M. (2023). Cometoid: Distilling strong reference-based machine translation metrics into Even stronger quality estimation metrics. In Koehn, P., Haddow, B., Kocmi, T., and Monz, C., editors, *Proceedings of the Eighth Conference on Machine Translation*, pages 751–755, Singapore. Association for Computational Linguistics.

Goyal, N., Du, J., Ott, M., Anantharaman, G., and Conneau, A. (2021). Larger-scale transformers for multilingual masked language modeling. *Computing Research Repository*, arXiv:2105.00572.

Goyal, N., Gao, C., Chaudhary, V., Chen, P.-J., Wenzek, G., Ju, D., Krishnan, S., Ranzato, M., Guzmán, F., and Fan, A. (2022). The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Graham, Y. (2015). Re-evaluating automatic summarization with BLEU and 192 shades of ROUGE. In Màrquez, L., Callison-Burch, C., and Su, J., editors, *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 128–137, Lisbon, Portugal. Association for Computational Linguistics.

Graham, Y., Baldwin, T., and Mathur, N. (2015). Accurate evaluation of segment-level machine translation metrics. In Mihalcea, R., Chai, J., and Sarkar, A., editors,

Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 1183–1191, Denver, Colorado. Association for Computational Linguistics.

Graham, Y., Baldwin, T., Moffat, A., and Zobel, J. (2013). Continuous measurement scales in human evaluation of machine translation. In Pareja-Lora, A., Liakata, M., and Dipper, S., editors, *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 33–41, Sofia, Bulgaria. Association for Computational Linguistics.

Guerreiro, N. M., Rei, R., van Stigt, D., Coheur, L., Colombo, P., and Martins, A. F. T. (2023). xcomet: Transparent machine translation evaluation through fine-grained error detection. *Computing Research Repository*, arXiv:2310.10482.

Guillou, L., Hardmeier, C., Lapshinova-Koltunski, E., and Loáiciga, S. (2018). A pronoun test suite evaluation of the English–German MT systems at WMT 2018. In Bojar, O., Chatterjee, R., Federmann, C., Fishel, M., Graham, Y., Haddow, B., Huck, M., Yepes, A. J., Koehn, P., Monz, C., Negri, M., Névéol, A., Neves, M., Post, M., Specia, L., Turchi, M., and Verspoor, K., editors, *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 570–577, Belgium, Brussels. Association for Computational Linguistics.

Gupta, R., Orăsan, C., and van Genabith, J. (2015). ReVal: A simple and effective machine translation evaluation metric based on recurrent neural networks. In Màrquez, L., Callison-Burch, C., and Su, J., editors, *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1066–1072, Lisbon, Portugal. Association for Computational Linguistics.

Han, H., Carpuat, M., and Boyd-Graber, J. (2022). SimQA: Detecting simultaneous MT errors through word-by-word question answering. In Goldberg, Y., Kozareva, Z., and Zhang, Y., editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5598–5616, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Hanna, M. and Bojar, O. (2021). A fine-grained analysis of BERTScore. In Barrault, L., Bojar, O., Bougares, F., Chatterjee, R., Costa-jussa, M. R., Federmann, C., Fishel, M., Fraser, A., Freitag, M., Graham, Y., Grundkiewicz, R., Guzman, P., Haddow, B., Huck, M., Yepes, A. J., Koehn, P., Kocmi, T., Martins, A., Morishita, M., and

Monz, C., editors, *Proceedings of the Sixth Conference on Machine Translation*, pages 507–517, Online. Association for Computational Linguistics.

Hemphill, C. T., Godfrey, J. J., and Doddington, G. R. (1990). The ATIS spoken language systems pilot corpus. In *Speech and Natural Language: Proceedings of a Workshop Held at Hidden Valley, Pennsylvania, June 24-27,1990*.

Hu, J., Ruder, S., Siddhant, A., Neubig, G., Firat, O., and Johnson, M. (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.

Hung, C.-C., Lauscher, A., Vulić, I., Ponzetto, S., and Glavaš, G. (2022). Multi2WOZ: A robust multilingual dataset and conversational pretraining for task-oriented dialog. In Carpuat, M., de Marneffe, M.-C., and Meza Ruiz, I. V., editors, *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3687–3703, Seattle, United States. Association for Computational Linguistics.

Ive, J., Blain, F., and Specia, L. (2018). deepQuest: A framework for neural-based quality estimation. In Bender, E. M., Derczynski, L., and Isabelle, P., editors, *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3146–3157, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Iyer, S., Konstas, I., Cheung, A., Krishnamurthy, J., and Zettlemoyer, L. (2017). Learning a neural semantic parser from user feedback. In Barzilay, R. and Kan, M.-Y., editors, *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 963–973, Vancouver, Canada. Association for Computational Linguistics.

Joshi, P., Santy, S., Budhiraja, A., Bali, K., and Choudhury, M. (2020). The state and fate of linguistic diversity and inclusion in the NLP world. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Juraska, J., Finkelstein, M., Deutsch, D., Siddhant, A., Mirzazadeh, M., and Freitag, M. (2023). MetricX-23: The Google submission to the WMT 2023 metrics shared task. In Koehn, P., Haddow, B., Kocmi, T., and Monz, C., editors, *Proceedings of the Eighth Conference on Machine Translation*, pages 756–767, Singapore. Association for Computational Linguistics.

Karpinska, M., Raj, N., Thai, K., Song, Y., Gupta, A., and Iyyer, M. (2022). DEMETR: Diagnosing evaluation metrics for translation. In Goldberg, Y., Kozareva, Z., and Zhang, Y., editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9540–9561, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Kim, H., Lee, J.-H., and Na, S.-H. (2017). Predictor-estimator using multilevel task learning with stack propagation for neural quality estimation. In Bojar, O., Buck, C., Chatterjee, R., Federmann, C., Graham, Y., Haddow, B., Huck, M., Yepes, A. J., Koehn, P., and Kreutzer, J., editors, *Proceedings of the Second Conference on Machine Translation*, pages 562–568, Copenhagen, Denmark. Association for Computational Linguistics.

Kocmi, T. and Federmann, C. (2023a). GEMBA-MQM: Detecting translation quality error spans with GPT-4. In Koehn, P., Haddow, B., Kocmi, T., and Monz, C., editors, *Proceedings of the Eighth Conference on Machine Translation*, pages 768–775, Singapore. Association for Computational Linguistics.

Kocmi, T. and Federmann, C. (2023b). Large language models are state-of-the-art evaluators of translation quality. In Nurminen, M., Brenner, J., Koponen, M., Latomaa, S., Mikhailov, M., Schierl, F., Ranasinghe, T., Vanmassenhove, E., Vidal, S. A., Aranberri, N., Nunziatini, M., Escartín, C. P., Forcada, M., Popovic, M., Scarton, C., and Moniz, H., editors, *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 193–203, Tampere, Finland. European Association for Machine Translation.

Kocmi, T., Federmann, C., Grundkiewicz, R., Junczys-Dowmunt, M., Matsushita, H., and Menezes, A. (2021). To ship or not to ship: An extensive evaluation of automatic metrics for machine translation. In Barrault, L., Bojar, O., Bougares, F., Chatterjee, R., Costa-jussa, M. R., Federmann, C., Fishel, M., Fraser, A., Freitag, M., Graham, Y., Grundkiewicz, R., Guzman, P., Haddow, B., Huck, M., Yepes, A. J., Koehn, P., Kocmi, T., Martins, A., Morishita, M., and Monz, C., editors, *Proceedings of the*

Sixth Conference on Machine Translation, pages 478–494, Online. Association for Computational Linguistics.

Kocmi, T., Matsushita, H., and Federmann, C. (2022). MS-COMET: More and better human judgements improve metric performance. In Koehn, P., Barrault, L., Bojar, O., Bougares, F., Chatterjee, R., Costa-jussà, M. R., Federmann, C., Fishel, M., Fraser, A., Freitag, M., Graham, Y., Grundkiewicz, R., Guzman, P., Haddow, B., Huck, M., Jimeno Yepes, A., Kocmi, T., Martins, A., Morishita, M., Monz, C., Nagata, M., Nakazawa, T., Negri, M., Névéol, A., Neves, M., Popel, M., Turchi, M., and Zampieri, M., editors, *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 541–548, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

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

Koehn, P. and Monz, C. (2006). Manual and automatic evaluation of machine translation between European languages. In Koehn, P. and Monz, C., editors, *Proceedings on the Workshop on Statistical Machine Translation*, pages 102–121, New York City. Association for Computational Linguistics.

Koehn, P., Och, F. J., and Marcu, D. (2003). Statistical phrase-based translation. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 127–133.

Krubiński, M. and Pecina, P. (2022). From COMET to COMES – can summary evaluation benefit from translation evaluation? In Deutsch, D., Udomcharoenchaikit, C., Opitz, J., Gao, Y., Fomicheva, M., and Eger, S., editors, *Proceedings of the 3rd Workshop on Evaluation and Comparison of NLP Systems*, pages 21–31, Online. Association for Computational Linguistics.

Kudo, T. and Richardson, J. (2018). SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In Blanco, E. and Lu, W., editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Laali, M. and Kosseim, L. (2017). Improving discourse relation projection to build discourse annotated corpora. In Mitkov, R. and Angelova, G., editors, *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 407–416, Varna, Bulgaria. INCOMA Ltd.

Laoudi, J., Tate, C. R., and Voss, C. R. (2006). Task-based MT evaluation: From who/when/where extraction to event understanding. In Calzolari, N., Choukri, K., Gangemi, A., Maegaard, B., Mariani, J., Odijk, J., and Tapias, D., editors, *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA).

Lapshinova-Koltunski, E., Hardmeier, C., and Krielke, P. (2018). ParCorFull: a parallel corpus annotated with full coreference. In Calzolari, N., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Hasida, K., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S., and Tokunaga, T., editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Läubli, S., Sennrich, R., and Volk, M. (2018). Has machine translation achieved human parity? a case for document-level evaluation. In Riloff, E., Chiang, D., Hockenmaier, J., and Tsujii, J., editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4791–4796, Brussels, Belgium. Association for Computational Linguistics.

Lee, H., Lee, J., and Kim, T.-Y. (2019). SUMBT: Slot-utterance matching for universal and scalable belief tracking. In Korhonen, A., Traum, D., and Màrquez, L., editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5478–5483, Florence, Italy. Association for Computational Linguistics.

Lee, S., Lee, J., Moon, H., Park, C., Seo, J., Eo, S., Koo, S., and Lim, H.-J. (2023). A survey on evaluation metrics for machine translation. *Mathematics*.

Lin, Z., Liu, Z., Winata, G. I., Cahyawijaya, S., Madotto, A., Bang, Y., Ishii, E., and Fung, P. (2021). XPersona: Evaluating multilingual personalized chatbot. In Papangelis, A., Budzianowski, P., Liu, B., Nouri, E., Rastogi, A., and Chen, Y.-N., editors, *Proceedings of the 3rd Workshop on Natural Language Processing for Conversational AI*, pages 102–112, Online. Association for Computational Linguistics.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. *ArXiv preprint*, abs/1907.11692.

Liu, Y., Qiao, X., Wu, Z., Chang, S., Zhang, M., Zhao, Y., Peng, S., Tao, S., Yang, H., Qin, Y., Guo, J., Wang, M., Li, Y., Li, P., and Zhao, X. (2022). Partial could be better than whole. HW-TSC 2022 submission for the metrics shared task. In Koehn, P., Barrault, L., Bojar, O., Bougares, F., Chatterjee, R., Costa-jussà, M. R., Federmann, C., Fishel, M., Fraser, A., Freitag, M., Graham, Y., Grundkiewicz, R., Guzman, P., Haddow, B., Huck, M., Jimeno Yepes, A., Kocmi, T., Martins, A., Morishita, M., Monz, C., Nagata, M., Nakazawa, T., Negri, M., Névéol, A., Neves, M., Popel, M., Turchi, M., and Zampieri, M., editors, *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 549–557, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Lo, C.-k. (2019). YiSi - a unified semantic MT quality evaluation and estimation metric for languages with different levels of available resources. In Bojar, O., Chatterjee, R., Federmann, C., Fishel, M., Graham, Y., Haddow, B., Huck, M., Yepes, A. J., Koehn, P., Martins, A., Monz, C., Negri, M., Névéol, A., Neves, M., Post, M., Turchi, M., and Verspoor, K., editors, *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 507–513, Florence, Italy. Association for Computational Linguistics.

Lo, C.-k., Larkin, S., and Knowles, R. (2023). Metric score landscape challenge (MSLC23): Understanding metrics' performance on a wider landscape of translation quality. In Koehn, P., Haddow, B., Kocmi, T., and Monz, C., editors, *Proceedings of the Eighth Conference on Machine Translation*, pages 776–799, Singapore. Association for Computational Linguistics.

Logacheva, V., Blain, F., and Specia, L. (2016). USFD's phrase-level quality estimation systems. In Bojar, O., Buck, C., Chatterjee, R., Federmann, C., Guillou, L., Haddow, B., Huck, M., Yepes, A. J., Névéol, A., Neves, M., Pecina, P., Popel, M., Koehn, P., Monz, C., Negri, M., Post, M., Specia, L., Verspoor, K., Tiedemann, J., and Turchi, M., editors, *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 800–805, Berlin, Germany. Association for Computational Linguistics.

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

Lommel, A. and Melby, A. (2018). Tutorial: MQM-DQF: A good marriage (translation quality for the 21st century). In Campbell, J., Yanishevsky, A., Doyon, J., and Jones, D., editors, *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 2: User Track)*, Boston, MA. Association for Machine Translation in the Americas.

Lopez, A. (2012). Putting human assessments of machine translation systems in order. In Callison-Burch, C., Koehn, P., Monz, C., Post, M., Soricut, R., and Specia, L., editors, *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 1–9, Montréal, Canada. Association for Computational Linguistics.

Lu, Q., Qiu, B., Ding, L., Zhang, K., Kocmi, T., and Tao, D. (2023). Error analysis prompting enables human-like translation evaluation in large language models: A case study on chatgpt. *Computing Research Repository*, arXiv:2303.13809.

Ma, Q., Wei, J., Bojar, O., and Graham, Y. (2019). Results of the WMT19 metrics shared task: Segment-level and strong MT systems pose big challenges. In Bojar, O., Chatterjee, R., Federmann, C., Fishel, M., Graham, Y., Haddow, B., Huck, M., Yepes, A. J., Koehn, P., Martins, A., Monz, C., Negri, M., Névéol, A., Neves, M., Post, M., Turchi, M., and Verspoor, K., editors, *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 62–90, Florence, Italy. Association for Computational Linguistics.

Macháček, M. and Bojar, O. (2014). Results of the WMT14 metrics shared task. In Bojar, O., Buck, C., Federmann, C., Haddow, B., Koehn, P., Monz, C., Post, M., and Specia, L., editors, *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 293–301, Baltimore, Maryland, USA. Association for Computational Linguistics.

Macketanz, V., Ai, R., Burchardt, A., and Uszkoreit, H. (2018a). TQ-AutoTest – an automated test suite for (machine) translation quality. In Calzolari, N., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Hasida, K., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S., and Tokunaga, T., editors, *Proceedings*

of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan. European Language Resources Association (ELRA).

Macketanz, V., Avramidis, E., Burchardt, A., and Uszkoreit, H. (2018b). Fine-grained evaluation of German-English machine translation based on a test suite. In Bojar, O., Chatterjee, R., Federmann, C., Fishel, M., Graham, Y., Haddow, B., Huck, M., Yepes, A. J., Koehn, P., Monz, C., Negri, M., Névéol, A., Neves, M., Post, M., Specia, L., Turchi, M., and Verspoor, K., editors, *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 578–587, Belgium, Brussels. Association for Computational Linguistics.

Marie, B. (2022). An Automatic Evaluation of the WMT22 General Machine Translation Task. *Computing Research Repository*, arXiv:2209.14172.

Marie, B., Fujita, A., and Rubino, R. (2021). Scientific credibility of machine translation research: A meta-evaluation of 769 papers. In Zong, C., Xia, F., Li, W., and Navigli, R., editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7297–7306, Online. Association for Computational Linguistics.

Martinovski, B. and Traum, D. (2003). The Error Is the Clue: Breakdown In Human-Machine Interaction. In *Proceedings of ISCA Tutorial and Research Workshop International Speech Communication Association*, Switzerland.

Mathur, N. (2021). *Robustness in Machine Translation Evaluation*. PhD thesis, The University of Melbourne, Australia.

Mathur, N., Baldwin, T., and Cohn, T. (2020a). Tangled up in BLEU: Reevaluating the evaluation of automatic machine translation evaluation metrics. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4984–4997, Online. Association for Computational Linguistics.

Mathur, N., Wei, J., Freitag, M., Ma, Q., and Bojar, O. (2020b). Results of the WMT20 metrics shared task. In Barrault, L., Bojar, O., Bougares, F., Chatterjee, R., Costa-jussà, M. R., Federmann, C., Fishel, M., Fraser, A., Graham, Y., Guzman, P., Haddow, B., Huck, M., Yepes, A. J., Koehn, P., Martins, A., Morishita, M., Monz, C., Nagata, M., Nakazawa, T., and Negri, M., editors, *Proceedings of the Fifth Conference*

on Machine Translation, pages 688–725, Online. Association for Computational Linguistics.

Matthews, B. W. (1975). Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochimica et Biophysica Acta (BBA) - Protein Structure*, 405(2):442–451.

Mehandru, N., Agrawal, S., Xiao, Y., Gao, G., Khoong, E., Carpuat, M., and Salehi, N. (2023). Physician detection of clinical harm in machine translation: Quality estimation aids in reliance and backtranslation identifies critical errors. In Bouamor, H., Pino, J., and Bali, K., editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11633–11647, Singapore. Association for Computational Linguistics.

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

Moghe, N., Fazla, A., Amrhein, C., Kocmi, T., Steedman, M., Birch, A., Sennrich, R., and Guillou, L. (2024). Machine translation meta evaluation through translation accuracy challenge sets. *Computing Research Repository*, arXiv:2401.16313.

Moghe, N., Hardmeier, C., and Bawden, R. (2020). The University of Edinburgh-Uppsala University's submission to the WMT 2020 chat translation task. In Barrault, L., Bojar, O., Bougares, F., Chatterjee, R., Costa-jussà, M. R., Federmann, C., Fishel, M., Fraser, A., Graham, Y., Guzman, P., Haddow, B., Huck, M., Yepes, A. J., Koehn, P., Martins, A., Morishita, M., Monz, C., Nagata, M., Nakazawa, T., and Negri, M., editors, *Proceedings of the Fifth Conference on Machine Translation*, pages 473–478, Online. Association for Computational Linguistics.

Moghe, N., Razumovskaia, E., Guillou, L., Vulić, I., Korhonen, A., and Birch, A. (2023a). Multi3NLU++: A multilingual, multi-intent, multi-domain dataset for natural language understanding in task-oriented dialogue. In Rogers, A., Boyd-Graber, J., and Okazaki, N., editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3732–3755, Toronto, Canada. Association for Computational Linguistics.

Moghe, N., Sherborne, T., Steedman, M., and Birch, A. (2023b). Extrinsic evaluation of machine translation metrics. In Rogers, A., Boyd-Graber, J., and Okazaki, N.,

editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13060–13078, Toronto, Canada. Association for Computational Linguistics.

Moghe, N., Steedman, M., and Birch, A. (2021). Cross-lingual intermediate fine-tuning improves dialogue state tracking. In Moens, M.-F., Huang, X., Specia, L., and Yih, S. W.-t., editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1137–1150, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Muennighoff, N., Wang, T., Sutawika, L., Roberts, A., Biderman, S., Le Scao, T., Bari, M. S., Shen, S., Yong, Z. X., Schoelkopf, H., Tang, X., Radev, D., Aji, A. F., Almubarak, K., Albanie, S., Alyafeai, Z., Webson, A., Raff, E., and Raffel, C. (2023). Crosslingual generalization through multitask finetuning. In Rogers, A., Boyd-Graber, J., and Okazaki, N., editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15991–16111, Toronto, Canada. Association for Computational Linguistics.

Mukherjee, A. and Shrivastava, M. (2023). MEE4 and XLsim : IIIT HYD's submissions' for WMT23 metrics shared task. In Koehn, P., Haddow, B., Kocmi, T., and Monz, C., editors, *Proceedings of the Eighth Conference on Machine Translation*, pages 800–805, Singapore. Association for Computational Linguistics.

Navigli, R. and Ponzetto, S. P. (2012). Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.

Neubig, G. (2022). Is my nlp model working? the answer is harder than you think.

Nimah, I., Fang, M., Menkovski, V., and Pechenizkiy, M. (2023). NLG evaluation metrics beyond correlation analysis: An empirical metric preference checklist. In Rogers, A., Boyd-Graber, J., and Okazaki, N., editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1240–1266, Toronto, Canada. Association for Computational Linguistics.

NLLB Team, Costa-jussà, M. R., Cross, J., Çelebi, O., Elbayad, M., Heafield, K., Heffernan, K., Kalbassi, E., Lam, J., Licht, D., Maillard, J., Sun, A., Wang, S., Wenzek, G., Youngblood, A., Akula, B., Barrault, L., Gonzalez, G. M., Hansanti, P., Hoffman, J., Jarrett, S., Sadagopan, K. R., Rowe, D., Spruit, S., Tran, C., Andrews,

P., Ayan, N. F., Bhosale, S., Edunov, S., Fan, A., Gao, C., Goswami, V., Guzmán, F., Koehn, P., Mourachko, A., Ropers, C., Saleem, S., Schwenk, H., and Wang, J. (2022). No language left behind: Scaling human-centered machine translation. *Computing Research Repository*, arXiv:2207.04672.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In Isabelle, P., Charniak, E., and Lin, D., editors, *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Patel, R. N. and M, S. (2016). Translation quality estimation using recurrent neural network. In Bojar, O., Buck, C., Chatterjee, R., Federmann, C., Guillou, L., Haddow, B., Huck, M., Yepes, A. J., Névéol, A., Neves, M., Pecina, P., Popel, M., Koehn, P., Monz, C., Negri, M., Post, M., Specia, L., Verspoor, K., Tiedemann, J., and Turchi, M., editors, *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 819–824, Berlin, Germany. Association for Computational Linguistics.

Perrella, S., Proietti, L., Scirè, A., Campolungo, N., and Navigli, R. (2022). MaTESe: Machine translation evaluation as a sequence tagging problem. In Koehn, P., Barrault, L., Bojar, O., Bougares, F., Chatterjee, R., Costa-jussà, M. R., Federmann, C., Fishel, M., Fraser, A., Freitag, M., Graham, Y., Grundkiewicz, R., Guzman, P., Haddow, B., Huck, M., Jimeno Yepes, A., Kocmi, T., Martins, A., Morishita, M., Monz, C., Nagata, M., Nakazawa, T., Negri, M., Névéol, A., Neves, M., Popel, M., Turchi, M., and Zampieri, M., editors, *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 569–577, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Popel, M. and Bojar, O. (2018). Training tips for the transformer model. *Prague Bull. Math. Linguistics*, 110:43–70.

Popović, M. (2017). chrF++: words helping character n-grams. In Bojar, O., Buck, C., Chatterjee, R., Federmann, C., Graham, Y., Haddow, B., Huck, M., Yepes, A. J., Koehn, P., and Kreutzer, J., editors, *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

Popović, M. (2021). Agree to disagree: Analysis of inter-annotator disagreements in human evaluation of machine translation output. In Bisazza, A. and Abend, O., editors, *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 234–243, Online. Association for Computational Linguistics.

Post, M. and Junczys-Dowmunt, M. (2023). Escaping the sentence-level paradigm in machine translation. *Computing Research Repository*, arXiv:2304.12959.

Raganato, A., Scherrer, Y., and Tiedemann, J. (2019). The MuCoW test suite at WMT 2019: Automatically harvested multilingual contrastive word sense disambiguation test sets for machine translation. In Bojar, O., Chatterjee, R., Federmann, C., Fishel, M., Graham, Y., Haddow, B., Huck, M., Yepes, A. J., Koehn, P., Martins, A., Monz, C., Negri, M., Névéol, A., Neves, M., Post, M., Turchi, M., and Verspoor, K., editors, *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 470–480, Florence, Italy. Association for Computational Linguistics.

Rajpurkar, P., Zhang, J., Lopyrev, K., and Liang, P. (2016). SQuAD: 100,000+ questions for machine comprehension of text. In Su, J., Duh, K., and Carreras, X., editors, *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Rei, R., C. de Souza, J. G., Alves, D., Zerva, C., Farinha, A. C., Glushkova, T., Lavie, A., Coheur, L., and Martins, A. F. T. (2022a). COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In Koehn, P., Barrault, L., Bojar, O., Bougares, F., Chatterjee, R., Costa-jussà, M. R., Federmann, C., Fishel, M., Fraser, A., Freitag, M., Graham, Y., Grundkiewicz, R., Guzman, P., Haddow, B., Huck, M., Jimeno Yepes, A., Kocmi, T., Martins, A., Morishita, M., Monz, C., Nagata, M., Nakazawa, T., Negri, M., Névéol, A., Neves, M., Popel, M., Turchi, M., and Zampieri, M., editors, *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Rei, R., Farinha, A. C., de Souza, J. G., Ramos, P. G., Martins, A. F., Coheur, L., and Lavie, A. (2022b). Searching for COMETINHO: The little metric that could. In Moniz, H., Macken, L., Rufener, A., Barrault, L., Costa-jussà, M. R., Declercq, C., Koponen, M., Kemp, E., Pilos, S., Forcada, M. L., Scarton, C., Van den Bogaert, J.,

Daems, J., Tezcan, A., Vanroy, B., and Fonteyne, M., editors, *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, pages 61–70, Ghent, Belgium. European Association for Machine Translation.

Rei, R., Farinha, A. C., Zerva, C., van Stigt, D., Stewart, C., Ramos, P., Glushkova, T., Martins, A. F. T., and Lavie, A. (2021). Are references really needed? unbabel-IST 2021 submission for the metrics shared task. In Barrault, L., Bojar, O., Bougares, F., Chatterjee, R., Costa-jussa, M. R., Federmann, C., Fishel, M., Fraser, A., Freitag, M., Graham, Y., Grundkiewicz, R., Guzman, P., Haddow, B., Huck, M., Yepes, A. J., Koehn, P., Kocmi, T., Martins, A., Morishita, M., and Monz, C., editors, *Proceedings of the Sixth Conference on Machine Translation*, pages 1030–1040, Online. Association for Computational Linguistics.

Rei, R., Guerreiro, N. M., Treviso, M., Coheur, L., Lavie, A., and Martins, A. (2023). The inside story: Towards better understanding of machine translation neural evaluation metrics. In Rogers, A., Boyd-Graber, J., and Okazaki, N., editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1089–1105, Toronto, Canada. Association for Computational Linguistics.

Rei, R., Stewart, C., Farinha, A. C., and Lavie, A. (2020). COMET: A neural framework for MT evaluation. In Webber, B., Cohn, T., He, Y., and Liu, Y., editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Rei, R., Treviso, M., Guerreiro, N. M., Zerva, C., Farinha, A. C., Maroti, C., C. de Souza, J. G., Glushkova, T., Alves, D., Coheur, L., Lavie, A., and Martins, A. F. T. (2022c). CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task. In Koehn, P., Barrault, L., Bojar, O., Bougares, F., Chatterjee, R., Costa-jussà, M. R., Federmann, C., Fishel, M., Fraser, A., Freitag, M., Graham, Y., Grundkiewicz, R., Guzman, P., Haddow, B., Huck, M., Jimeno Yepes, A., Kocmi, T., Martins, A., Morishita, M., Monz, C., Nagata, M., Nakazawa, T., Negri, M., Névéol, A., Neves, M., Popel, M., Turchi, M., and Zampieri, M., editors, *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Reimers, N. and Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In Inui, K., Jiang, J., Ng, V., and Wan, X., editors,

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.

Reiter, E. (2019). Natural language generation challenges for explainable AI. In Alonso, J. M. and Catala, A., editors, *Proceedings of the 1st Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence (NL4XAI 2019)*, pages 3–7. Association for Computational Linguistics.

Rudinger, R., May, C., and Van Durme, B. (2017). Social bias in elicited natural language inferences. In Hovy, D., Spruit, S., Mitchell, M., Bender, E. M., Strube, M., and Wallach, H., editors, *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 74–79, Valencia, Spain. Association for Computational Linguistics.

Sadat, M., Zhou, Z., Lange, L., Araki, J., Gundroo, A., Wang, B., Menon, R., Parvez, M., and Feng, Z. (2023). DelucionQA: Detecting hallucinations in domain-specific question answering. In Bouamor, H., Pino, J., and Bali, K., editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 822–835, Singapore. Association for Computational Linguistics.

Sai, A. B., Dixit, T., Sheth, D. Y., Mohan, S., and Khapra, M. M. (2021). Perturbation CheckLists for evaluating NLG evaluation metrics. In Moens, M.-F., Huang, X., Specia, L., and Yih, S. W.-t., editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7219–7234, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Sai B, A., Dixit, T., Nagarajan, V., Kunchukuttan, A., Kumar, P., Khapra, M. M., and Dabre, R. (2023). IndicMT eval: A dataset to meta-evaluate machine translation metrics for Indian languages. In Rogers, A., Boyd-Graber, J., and Okazaki, N., editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14210–14228, Toronto, Canada. Association for Computational Linguistics.

Scao, T. L., Fan, A., Akiki, C., Pavlick, E., Ilic, S., Hesslow, D., Castagné, R., Luccioni, A. S., Yvon, F., Gallé, M., Tow, J., Rush, A. M., Biderman, S., Webson, A., Ammanamanchi, P. S., Wang, T., Sagot, B., Muennighoff, N., del Moral, A. V.,

Ruwase, O., Bawden, R., Bekman, S., McMillan-Major, A., Beltagy, I., Nguyen, H., Saulnier, L., Tan, S., Suarez, P. O., Sanh, V., Laurençon, H., Jernite, Y., Launay, J., Mitchell, M., Raffel, C., Gokaslan, A., Simhi, A., Soroa, A., Aji, A. F., Alfassy, A., Rogers, A., Nitzav, A. K., Xu, C., Mou, C., Emezue, C., Klamm, C., Leong, C., van Strien, D., Adelani, D. I., and et al. (2022). BLOOM: A 176b-parameter open-access multilingual language model. *Computing Research Repository*, arXiv:2211.05100.

Scarton, C. (2016). *Document-Level Machine Translation Quality Estimation*. PhD thesis, The University of Sheffield, United Kingdom.

Scarton, C., Forcada, M. L., Esplà-Gomis, M., and Specia, L. (2019). Estimating post-editing effort: a study on human judgements, task-based and reference-based metrics of MT quality. In Niehues, J., Cattoni, R., Stüker, S., Negri, M., Turchi, M., Ha, T.-L., Salesky, E., Sanabria, R., Barrault, L., Specia, L., and Federico, M., editors, *Proceedings of the 16th International Conference on Spoken Language Translation*, Hong Kong. Association for Computational Linguistics.

Sellam, T., Das, D., and Parikh, A. (2020a). BLEURT: Learning robust metrics for text generation. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Sellam, T., Pu, A., Chung, H. W., Gehrmann, S., Tan, Q., Freitag, M., Das, D., and Parikh, A. (2020b). Learning to evaluate translation beyond English: BLEURT submissions to the WMT metrics 2020 shared task. In Barrault, L., Bojar, O., Bougares, F., Chatterjee, R., Costa-jussà, M. R., Federmann, C., Fishel, M., Fraser, A., Graham, Y., Guzman, P., Haddow, B., Huck, M., Yepes, A. J., Koehn, P., Martins, A., Morishita, M., Monz, C., Nagata, M., Nakazawa, T., and Negri, M., editors, *Proceedings of the Fifth Conference on Machine Translation*, pages 921–927, Online. Association for Computational Linguistics.

Sennrich, R., Haddow, B., and Birch, A. (2016). Neural machine translation of rare words with subword units. In Erk, K. and Smith, N. A., editors, *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.

Sherborne, T. and Lapata, M. (2022). Zero-shot cross-lingual semantic parsing. In

Muresan, S., Nakov, P., and Villavicencio, A., editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4134–4153, Dublin, Ireland. Association for Computational Linguistics.

Sherborne, T. and Lapata, M. (2023). Meta-Learning a Cross-lingual Manifold for Semantic Parsing. *Transactions of the Association for Computational Linguistics*, 11:49–67.

Shimanaka, H., Kajiwara, T., and Komachi, M. (2018). RUSE: Regressor using sentence embeddings for automatic machine translation evaluation. In Bojar, O., Chatterjee, R., Federmann, C., Fishel, M., Graham, Y., Haddow, B., Huck, M., Yepes, A. J., Koehn, P., Monz, C., Negri, M., Névéol, A., Neves, M., Post, M., Specia, L., Turchi, M., and Verspoor, K., editors, *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 751–758, Belgium, Brussels. Association for Computational Linguistics.

Sinha, K., Jia, R., Hupkes, D., Pineau, J., Williams, A., and Kiela, D. (2021). Masked language modeling and the distributional hypothesis: Order word matters pre-training for little. In Moens, M.-F., Huang, X., Specia, L., and Yih, S. W.-t., editors, *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.

Song, X. and Cohn, T. (2011). Regression and ranking based optimisation for sentence level MT evaluation. In Callison-Burch, C., Koehn, P., Monz, C., and Zaidan, O. F., editors, *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 123–129, Edinburgh, Scotland. Association for Computational Linguistics.

Soricut, R. and Echihabi, A. (2010). TrustRank: Inducing trust in automatic translations via ranking. In Hajič, J., Carberry, S., Clark, S., and Nivre, J., editors, *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 612–621, Uppsala, Sweden. Association for Computational Linguistics.

Specia, L., Blain, F., Fomicheva, M., Zerva, C., Li, Z., Chaudhary, V., and Martins, A. F. T. (2021). Findings of the WMT 2021 shared task on quality estimation. In Barrault, L., Bojar, O., Bougares, F., Chatterjee, R., Costa-jussa, M. R., Federmann, C., Fishel, M., Fraser, A., Freitag, M., Graham, Y., Grundkiewicz, R., Guzman, P., Haddow, B., Huck, M., Yepes, A. J., Koehn, P., Kocmi, T., Martins, A., Morishita, M.,

and Monz, C., editors, *Proceedings of the Sixth Conference on Machine Translation*, pages 684–725, Online. Association for Computational Linguistics.

Specia, L., Blain, F., Logacheva, V., F. Astudillo, R., and Martins, A. F. T. (2018). Findings of the WMT 2018 shared task on quality estimation. In Bojar, O., Chatterjee, R., Federmann, C., Fishel, M., Graham, Y., Haddow, B., Huck, M., Yepes, A. J., Koehn, P., Monz, C., Negri, M., Névéol, A., Neves, M., Post, M., Specia, L., Turchi, M., and Verspoor, K., editors, *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 689–709, Belgium, Brussels. Association for Computational Linguistics.

Specia, L. and Farzindar, A. (2010). Estimating machine translation post-editing effort with HTER. In Zhechev, V., editor, *Proceedings of the Second Joint EM+/CNGL Workshop: Bringing MT to the User: Research on Integrating MT in the Translation Industry*, pages 33–43, Denver, Colorado, USA. Association for Machine Translation in the Americas.

Stanojević, M., Kamran, A., Koehn, P., and Bojar, O. (2015). Results of the WMT15 metrics shared task. In Bojar, O., Chatterjee, R., Federmann, C., Haddow, B., Hokamp, C., Huck, M., Logacheva, V., and Pecina, P., editors, *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 256–273, Lisbon, Portugal. Association for Computational Linguistics.

Stanovsky, G., Smith, N. A., and Zettlemoyer, L. (2019). Evaluating gender bias in machine translation. In Korhonen, A., Traum, D., and Màrquez, L., editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1679–1684, Florence, Italy. Association for Computational Linguistics.

Stymne, S., Danielsson, H., Bremin, S., Hu, H., Karlsson, J., Lillkull, A. P., and Wester, M. (2012). Eye tracking as a tool for machine translation error analysis. In Calzolari, N., Choukri, K., Declerck, T., Doğan, M. U., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 1121–1126, Istanbul, Turkey. European Language Resources Association (ELRA).

Sun, T., He, J., Qiu, X., and Huang, X. (2022). BERTScore is unfair: On social bias in language model-based metrics for text generation. In Goldberg, Y., Kozareva, Z., and Zhang, Y., editors, *Proceedings of the 2022 Conference on Empirical Methods in*

Natural Language Processing, pages 3726–3739, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Tang, Y., Tran, C., Li, X., Chen, P.-J., Goyal, N., Chaudhary, V., Gu, J., and Fan, A. (2021). Multilingual translation from denoising pre-training. In Zong, C., Xia, F., Li, W., and Navigli, R., editors, *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3450–3466, Online. Association for Computational Linguistics.

Tao, S., Chang, S., Miaomiao, M., Yang, H., Geng, X., Huang, S., Zhang, M., Guo, J., Wang, M., and Li, Y. (2022). CrossQE: HW-TSC 2022 submission for the quality estimation shared task. In Koehn, P., Barrault, L., Bojar, O., Bougares, F., Chatterjee, R., Costa-jussà, M. R., Federmann, C., Fishel, M., Fraser, A., Freitag, M., Graham, Y., Grundkiewicz, R., Guzman, P., Haddow, B., Huck, M., Jimeno Yepes, A., Kocmi, T., Martins, A., Morishita, M., Monz, C., Nagata, M., Nakazawa, T., Negri, M., Névéol, A., Neves, M., Popel, M., Turchi, M., and Zampieri, M., editors, *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 646–652, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Taori, R., Gulrajani, I., Zhang, T., Dubois, Y., Li, X., Guestrin, C., Liang, P., and Hashimoto, T. B. (2023). Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`.

Tatsumi, M. (2009). Correlation between automatic evaluation metric scores, post-editing speed, and some other factors. In *Proceedings of Machine Translation Summit XII: Posters*, Ottawa, Canada.

Temnikova, I. (2010). Cognitive evaluation approach for a controlled language post-editing experiment. In Calzolari, N., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., Rosner, M., and Tapias, D., editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Tiedemann, J. and Thottingal, S. (2020). OPUS-MT – building open translation services for the world. In Martins, A., Moniz, H., Fumega, S., Martins, B., Batista, F., Coheur, L., Parra, C., Trancoso, I., Turchi, M., Bisazza, A., Moorkens, J., Guerberof, A., Nurminen, M., Marg, L., and Forcada, M. L., editors, *Proceedings of the 22nd Annual*

Conference of the European Association for Machine Translation, pages 479–480, Lisboa, Portugal. European Association for Machine Translation.

Tomita, M., Shirai, M., Tsutsumi, J., Matsumura, M., and Yuki (1993). Evaluation of MT systems by TOEFL. In *Proceedings of the Fifth Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages*, Kyoto, Japan.

Toral, A. and Sánchez-Cartagena, V. M. (2017). A multifaceted evaluation of neural versus phrase-based machine translation for 9 language directions. In Lapata, M., Blunsom, P., and Koller, A., editors, *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.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C. C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., Fuller, B., Gao, C., Goswami, V., Goyal, N., Hartshorn, A., Hosseini, S., Hou, R., Inan, H., Kardas, M., Kerkez, V., Khabsa, M., Kloumann, I., Korenev, A., Koura, P. S., Lachaux, M.-A., Lavril, T., Lee, J., Liskovich, D., Lu, Y., Mao, Y., Martinet, X., Mihaylov, T., Mishra, P., Molybog, I., Nie, Y., Poulton, A., Reizenstein, J., Rungta, R., Saladi, K., Schelten, A., Silva, R., Smith, E. M., Subramanian, R., Tan, X. E., Tang, B., Taylor, R., Williams, A., Kuan, J. X., Xu, P., Yan, Z., Zarov, I., Zhang, Y., Fan, A., Kambadur, M., Narang, S., Rodriguez, A., Stojnic, R., Edunov, S., and Scialom, T. (2023). Llama 2: Open foundation and fine-tuned chat models. *Computing Research Repository*, arXiv:2307.09288.

Ueffing, N. and Ney, H. (2005). Word-level confidence estimation for machine translation using phrase-based translation models. In Mooney, R., Brew, C., Chien, L.-F., and Kirchhoff, K., editors, *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 763–770, Vancouver, British Columbia, Canada. Association for Computational Linguistics.

Vamvas, J. and Sennrich, R. (2021). Contrastive conditioning for assessing disambiguation in MT: A case study of distilled bias. In Moens, M.-F., Huang, X., Specia, L., and Yih, S. W.-t., editors, *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.

Vamvas, J. and Sennrich, R. (2022). As little as possible, as much as necessary: Detecting over- and undertranslations with contrastive conditioning. In Muresan, S., Nakov, P., and Villavicencio, A., editors, *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.

Vernikos, G., Thompson, B., Mathur, P., and Federico, M. (2022). Embarrassingly easy document-level MT metrics: How to convert any pretrained metric into a document-level metric. In Koehn, P., Barrault, L., Bojar, O., Bougares, F., Chatterjee, R., Costa-jussà, M. R., Federmann, C., Fishel, M., Fraser, A., Freitag, M., Graham, Y., Grundkiewicz, R., Guzman, P., Haddow, B., Huck, M., Jimeno Yepes, A., Kocmi, T., Martins, A., Morishita, M., Monz, C., Nagata, M., Nakazawa, T., Negri, M., Névéol, A., Neves, M., Popel, M., Turchi, M., and Zampieri, M., editors, *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 118–128, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Vieira, L. N., O'Hagan, M., and O'Sullivan, C. (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.

Wan, Y., Liu, D., Yang, B., Zhang, H., Chen, B., Wong, D., and Chao, L. (2022). UniTE: Unified translation evaluation. In Muresan, S., Nakov, P., and Villavicencio, A., editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8117–8127, Dublin, Ireland. Association for Computational Linguistics.

Wang, J., Adelani, D. I., Agrawal, S., Rei, R., Briakou, E., Carpuat, M., Masiak, M., He, X., Bourhim, S., Bukula, A., Mohamed, M., Olatoye, T., Mokayede, H., Mwase, C., Kimotho, W., Yuehgoh, F., Aremu, A., Ojo, J., Muhammad, S. H., Osei, S., Omotayo, A.-H., Chukwuneke, C., Ogayo, P., Hourrane, O., Anigri, S. E., Ndolela, L., Mangwana, T., Mohamed, S. A., Hassan, A., Awoyomi, O. O., Alkhaled, L., Al-Azzawi, S., Etori, N. A., Ochieng, M., Siro, C., Njoroge, S., Muchiri, E., Kimotho, W., Momo, L. N. W., Abolade, D., Ajao, S., Adewumi, T., Shode, I., Macharm, R., Iro, R. N., Abdullahi, S. S., Moore, S. E., Opoku, B., Akinjobi, Z., Afolabi, A., Obiefuna, N., Ogbu, O. R., Brian, S., Otiende, V. A., Mbonu, C. E., Sari,

S. T., and Stenetorp, P. (2023). AfriMTE and AfriCOMET: Empowering COMET to embrace under-resourced African languages. *Computing Research Repository*, arXiv:2311.09828.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E. H., Le, Q. V., and Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A., editors, *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

West, P., Lu, X., Dziri, N., Brahman, F., Li, L., Hwang, J. D., Jiang, L., Fisher, J., Ravichander, A., Chandu, K., Newman, B., Koh, P. W., Ettinger, A., and Choi, Y. (2024). The generative AI paradox: "what it can create, it may not understand". In *The Twelfth International Conference on Learning Representations*.

White, J. S., O'Connell, T. A., and O'Mara, F. E. (1994). The ARPA MT evaluation methodologies: Evolution, lessons, and future approaches. In *Proceedings of the First Conference of the Association for Machine Translation in the Americas*, Columbia, Maryland, USA.

Wu, S. and Dredze, M. (2019). Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. In Inui, K., Jiang, J., Ng, V., and Wan, X., editors, *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.

Xu, W., Haider, B., and Mansour, S. (2020). End-to-end slot alignment and recognition for cross-lingual NLU. In Webber, B., Cohn, T., He, Y., and Liu, Y., editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5052–5063, Online. Association for Computational Linguistics.

Xu, W., Wang, D., Pan, L., Song, Z., Freitag, M., Wang, W., and Li, L. (2023). INSTRUCTSCORE: Towards explainable text generation evaluation with automatic feedback. In Bouamor, H., Pino, J., and Bali, K., editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5967–5994, Singapore. Association for Computational Linguistics.

Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., and Raffel, C. (2021). mT5: A massively multilingual pre-trained text-to-text transformer. In Toutanova, K., Rumshisky, A., Zettlemoyer, L., Hakkani-Tur, D., Beltagy, I., Bethard, S., Cotterell, R., Chakraborty, T., and Zhou, Y., editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Yan, Y., Wang, T., Zhao, C., Huang, S., Chen, J., and Wang, M. (2023). BLEURT has universal translations: An analysis of automatic metrics by minimum risk training. In Rogers, A., Boyd-Graber, J., and Okazaki, N., editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5428–5443, Toronto, Canada. Association for Computational Linguistics.

Yang, Y., Zhang, Y., Tar, C., and Baldridge, J. (2019). PAWS-X: A cross-lingual adversarial dataset for paraphrase identification. In Inui, K., Jiang, J., Ng, V., and Wan, X., editors, *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.

Yu, H., Ma, Q., Wu, X., and Liu, Q. (2015). CASICT-DCU participation in WMT2015 metrics task. In Bojar, O., Chatterjee, R., Federmann, C., Haddow, B., Hokamp, C., Huck, M., Logacheva, V., and Pecina, P., editors, *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 417–421, Lisbon, Portugal. Association for Computational Linguistics.

Yuan, W., Neubig, G., and Liu, P. (2021). BARTScore: Evaluating generated text as text generation. In Ranzato, M., Beygelzimer, A., Dauphin, Y. N., Liang, P., and Vaughan, J. W., editors, *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 27263–27277.

Zerva, C., Blain, F., Rei, R., Lertvittayakumjorn, P., C. de Souza, J. G., Eger, S., Kanojia, D., Alves, D., Orăsan, C., Fomicheva, M., Martins, A. F. T., and Specia, L. (2022). Findings of the WMT 2022 shared task on quality estimation. In Koehn, P., Barrault, L., Bojar, O., Bougares, F., Chatterjee, R., Costa-jussà, M. R., Federmann, C., Fishel, M., Fraser, A., Freitag, M., Graham, Y., Grundkiewicz, R., Guzman, P.,

Haddow, B., Huck, M., Jimeno Yepes, A., Kocmi, T., Martins, A., Morishita, M., Monz, C., Nagata, M., Nakazawa, T., Negri, M., Névéol, A., Neves, M., Popel, M., Turchi, M., and Zampieri, M., editors, *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 69–99, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Zhang, H., Tan, L., and Misra, A. (2022). Evaluating machine translation in cross-lingual E-commerce search. In Duh, K. and Guzmán, F., editors, *Proceedings of the 15th biennial conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 322–334, Orlando, USA. Association for Machine Translation in the Americas.

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., and Artzi, Y. (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.

Zhang, Y., Baldridge, J., and He, L. (2019). PAWS: Paraphrase adversaries from word scrambling. In Burstein, J., Doran, C., and Solorio, T., editors, *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.

Zhao, H., Liu, Y., Tao, S., Meng, W., Chen, Y., Geng, X., Su, C., Zhang, M., and Yang, H. (2024). From handcrafted features to llms: A brief survey for machine translation quality estimation. In *International Joint Conference on Neural Networks, IJCNN 2024*. IEEE.

Zhao, J., Wang, T., Yatskar, M., Ordonez, V., and Chang, K.-W. (2018). Gender bias in coreference resolution: Evaluation and debiasing methods. In Walker, M., Ji, H., and Stent, A., editors, *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.

Zhao, Z., Cohen, S. B., and Webber, B. (2020). Reducing quantity hallucinations in abstractive summarization. In Cohn, T., He, Y., and Liu, Y., editors, *Findings of the*

Association for Computational Linguistics: EMNLP 2020, pages 2237–2249, Online. Association for Computational Linguistics.

Zhou, J., Gong, H., and Bhat, S. (2021). PIE: A parallel idiomatic expression corpus for idiomatic sentence generation and paraphrasing. In Cook, P., Mitrović, J., Escartín, C. P., Vaidya, A., Osenova, P., Taslimipoor, S., and Ramisch, C., editors, *Proceedings of the 17th Workshop on Multiword Expressions (MWE 2021)*, pages 33–48, Online. Association for Computational Linguistics.