SAPIENZA
UNIVERSITÀ DI ROMA

Towards Accurate and Interpretable Machine Translation Evaluation

Dipartimento di Ingegneria Informatica, Automatica, e Gestionale (DIAG)
PhD in Engineering in Computer Science (XXXVIII cycle)

Stefano Perrella
ID number 1755661

Advisor
Prof. Roberto Navigli

Academic Year 2024/2025

Thesis defended on 29 January 2026
in front of a Board of Examiners composed by:

Prof. Alessandro Raganato
Prof. Barbara Pernici
Prof. Damian Andrew Tamburri

This thesis was reviewed by two external reviewers:

Prof. Alexandra Birch
Prof. Matt Post

Towards Accurate and Interpretable Machine Translation Evaluation
PhD thesis. Sapienza University of Rome

© 2025 Stefano Perrella. All rights reserved

This thesis has been typeset by LaTeX and the Sapthesis class.

Author's email: perrella@diag.uniroma1.it

Dedicated to

my family

Abstract

Deep learning has revolutionized Artificial Intelligence (AI), with most AI applications now leveraging deep learning techniques to achieve state-of-the-art performance. One of the most prominent applications is Machine Translation (MT), which automatically translates text from one language to another. Over the past few years, the quality of automatic translation has improved dramatically, with current neural translation systems now approaching human performance on standard benchmarks. As a result, the MT field faces new challenges: As top systems produce translations that differ only in minor nuances rather than in major errors, improvements have become increasingly difficult to measure.

Unlike many other AI applications, MT is particularly challenging to evaluate. Given a source text, there is rarely a single correct translation, and many outputs may be valid yet differ in form. Similarly, many outputs can be incorrect for different reasons. Due to this intrinsic complexity, MT evaluation has emerged as a dedicated research area focused on developing better evaluation techniques. This area has also advanced significantly in recent years, primarily due to the replacement of earlier heuristic approaches with neural evaluation metrics. However, while improving evaluation accuracy, neural metrics sacrificed interpretability. Indeed, the most widely used neural metrics are black-box models that reduce translation quality to a single scalar score, making the evaluation process opaque and providing limited insights into the reasons behind a particular assessment. If MT evaluation yields unreliable assessments, it risks steering system development in the wrong direction. This highlights the importance of another research area in MT: the MT meta-evaluation, which assesses the effectiveness of MT evaluation metrics and ensures their alignment with human evaluation by measuring their correlation with human judgments.

In this dissertation, we advance the ability to measure progress in the MT field by addressing key challenges in both MT evaluation and meta-evaluation, with a particular emphasis on enhancing evaluation interpretability.

To mitigate the lack of interpretability in MT evaluation, we make two main contributions: (i) we introduce neural metrics capable of identifying error spans within translations and assigning severity levels to them, thereby enabling finer-grained and more interpretable feedback, compared to neural metrics that output scalar assessments; (ii) we propose an

interpretable meta-evaluation framework that assesses metric performance using Precision, Recall, F-score, and Re-Ranking Precision. While standard correlation with human judgments is primarily useful for ranking metrics, our meta-evaluation framework offers more intuitive insights into their absolute evaluation performance.

Our contributions to MT meta-evaluation extend beyond improving interpretability. We identify fundamental flaws in widely used meta-evaluation strategies, which can inadvertently reward metrics for the wrong reasons. To reveal and quantify these weaknesses, we introduce the concept of sentinel metrics – intentionally incomplete metrics designed to highlight shortcomings in meta-evaluation methods – and incorporate them into the meta-evaluation process.

Finally, we shift focus from improving evaluation techniques to facilitating better test data selection. As MT systems achieve increasingly higher performance, existing benchmarks have become too easy, making it difficult to distinguish among top-performing systems or identify areas requiring further improvement. We address this issue by introducing Translation Difficulty Estimation – the task of identifying texts that are difficult to translate – and show how difficulty estimators can be used to construct more challenging MT benchmarks. We also develop state-of-the-art difficulty estimators and use them to build the test set for the General Machine Translation Shared Task at the 2025 edition of the Conference on Machine Translation (WMT).

Taken together, our contributions advance the field of Machine Translation by providing interpretable evaluation metrics and meta-metrics, more reliable meta-evaluation, and more informative benchmarks – ultimately improving how progress in MT is measured and understood.

Contents

1 Introduction **1**

1.1 MT Evaluation . . . 2

1.1.1 Human Evaluation and MT Benchmarks . . . 4

1.2 Thesis Outline . . . 5

1.3 Publications and Contributions . . . 6

2 Background **11**

2.1 Machine Translation Evaluation . . . 11

2.1.1 Human Evaluation . . . 11

2.1.2 Automatic Evaluation . . . 19

2.2 Machine Translation Meta-Evaluation . . . 26

3 Interpretable Machine Translation Evaluation **31**

3.1 The MaTESe metrics . . . 31

3.1.1 Formal definition . . . 32

3.1.2 Data pre-processing . . . 35

3.2 Experimental Setup . . . 36

3.2.1 Architectures . . . 36

3.2.2 Training and validation data . . . 37

3.2.3 Evaluation metrics . . . 38

3.3 Results . . . 40

3.3.1 Architectures comparison . . . 40

3.3.2 Monolingual-multilingual comparison . . . 41

3.3.3 MaTESe-QE . . . 42

3.3.4 Correlations with Human Judgments . . . 42
3.3.5 Submissions to WMT . . . 44
3.4 Conclusions . . . 45

4 Interpretable Machine Translation Meta-Evaluation 49
4.1 Introduction . . . 49
4.2 An Interpretable Evaluation Framework for MT Metrics . . . 51
4.2.1 Metrics as Binary Classifiers for Data Filtering . . . 51
4.2.2 Metrics as Utility Functions for Translation Re-Ranking . . . 52
4.3 Experimental Setup . . . 53
4.3.1 The Data . . . 53
4.3.2 The Metrics . . . 54
4.3.3 Selecting the Thresholds τ . . . 55
4.3.4 Extracting Binary Labels from Manually-Annotated Datasets . . . 55
4.4 Results . . . 56
4.4.1 Binary Classification . . . 56
4.4.2 Translation Re-Ranking . . . 65
4.5 Conclusion . . . 67

5 Improving the MT Meta-Evaluation with Sentinel Metrics 69
5.1 The MT Meta-Evaluation . . . 69
5.2 To Group or Not to Group? . . . 71
5.2.1 The Relation Between Spurious Correlations and Grouping . . . 72
5.2.2 The Sentinel Metrics . . . 73
5.2.3 Training the Sentinel Metrics . . . 73
5.2.4 Results . . . 75
5.3 The Evaluation of Ties . . . 82
5.3.1 The Problem with Tie Calibration . . . 83
5.3.2 Two New Sentinel Metrics . . . 83
5.3.3 Results . . . 84
5.4 Conclusion . . . 86

6 Estimating Machine Translation Difficulty 93

6.1 Why Translation Difficulty? . . . 93

6.2 Related Work . . . 94

6.3 The Difficulty Estimation Task . . . 96

6.4 Methods for Difficulty Estimation . . . 98

6.4.1 Heuristic-based estimators . . . 98

6.4.2 Learned estimators . . . 99

6.4.3 LLM-as-a-Judge . . . 100

6.4.4 Crowd-based Estimators . . . 102

6.5 Experiments . . . 103

6.5.1 Experimental Setup . . . 103

6.5.2 Results . . . 104

6.5.3 Discussion . . . 105

6.5.4 Comparing Human and Machine Translation Difficulty . . . 105

6.6 Creating Difficult Benchmarks . . . 106

6.6.1 Setup . . . 106

6.6.2 Results . . . 108

6.7 Bias toward MT models . . . 110

6.8 Conclusion . . . 111

7 The 2025 Conference on Machine Translation 115

7.1 The Impact of Sentinel Metrics . . . 115

7.2 The Impact of Translation Difficulty Estimation . . . 116

8 Conclusions 119

8.1 Summary of Contributions . . . 119

8.2 Future Work . . . 121

Bibliography 123

Chapter 1

Introduction

Machine Translation (MT) involves automatically translating text from one language to another. Over the past few years, research in MT has brought substantial improvements to the field. Early statistical approaches have been replaced by neural ones, and more recently, Large Language Models (LLMs) have advanced translation quality further – approaching, and in some cases even surpassing, human performance on standard benchmarks (Kocmi et al., 2023, 2024a, 2025c). Similarly, automatic MT evaluation – i.e., the automatic assessment of translation quality – has improved. Early approaches to MT evaluation relied on comparing translations to manually curated references via simple heuristics. Instead, current MT evaluation techniques are based on neural models, and often assess the quality of translated text by comparing it to its source, without the need for gold references. However, despite increased evaluation accuracy, recent automatic evaluation methods lack interpretability: by being based on black-box neural models, their decision-making processes are opaque, making their assessments difficult to interpret. Furthermore, as translation quality improves, even human MT evaluation has become more challenging. Indeed, the outputs of top-performing systems tend to diverge due to minor linguistic phenomena rather than major errors, which makes it difficult for evaluators – whether human or automatic – to distinguish between higher- and lower-quality translations.

The overarching goal of this thesis is to enhance our ability to measure progress in MT, with particular attention to the interpretability of evaluation results. This chapter introduces the main research problems addressed in this work and outlines the thesis structure.

1.1 MT Evaluation

Unlike many Natural Language Processing (NLP) tasks that yield structured or categorical outputs – such as Part-of-Speech Tagging, Named Entity Recognition, or Sentiment Analysis – Machine Translation produces free-form text. This makes its evaluation substantially more challenging. For structured tasks, accuracy-based metrics can be directly computed against a gold standard, as there is typically one single correct label or structure for each input. In contrast, a translation can be accurate in multiple ways, since the same meaning may be conveyed using different word choices, phrasings, or syntactic structures. Consequently, automatic MT evaluation has become a research area of its own, with metrics being developed to approximate human judgment as closely as possible.

Early approaches to MT evaluation approximate human judgment by measuring surface-level similarity between a translation and a manually curated reference (Papineni et al., 2002; Popović, 2015). However, these methods depend on the availability of such references and fail to reliably quantify translation quality when a high-quality translation is lexically dissimilar from the reference. More recently, with the advent of deep learning, automatic evaluation has evolved from surface-level approaches to neural metrics capable of assessing translation quality at a semantic level (Rei et al., 2020; Juraska et al., 2023; Guerreiro et al., 2024). However, although neural metrics generally exhibit higher correlation with human judgments than surface-level ones (Freitag et al., 2022b), they suffer from a major drawback: they lack interpretability. These metrics operate as black-box systems whose decision-making processes are opaque, and most provide assessments only as scalar scores that are difficult to interpret. This limitation risks confounding evaluation results, as it can be challenging to understand the reasons behind the produced assessments. Moreover, it hinders the use of automatic evaluation for fine-grained analysis of translation quality beyond single-score assessments.

The lack of interpretability in MT evaluation also complicates the evaluation of the metrics themselves – a process known as meta-evaluation. Since most metrics produce only scalar scores, the MT meta-evaluation largely relies on measuring metric performance via correlation between metric outputs and human judgments (Mathur et al., 2020b; Freitag et al., 2021b, 2022b, 2023, 2024). High correlation indicates stronger alignment with human judgment and, consequently, greater evaluation accuracy. However, while useful for

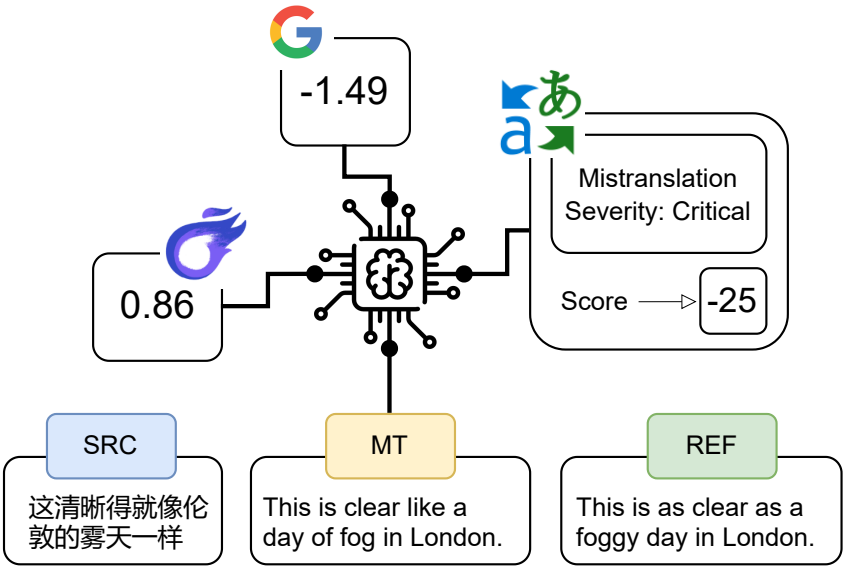


Figure 1.1. Translation quality assessments returned, from left to right, by COMET (Rei et al., 2020), MetricX-23-QE (Juraska et al., 2023), and GEMBA-MQM (Kocmi and Federmann, 2023a) for the provided machine-translated text.

comparing metrics, correlation measures can themselves be difficult to interpret, offering limited insights into the true capabilities of MT metrics.

The Interpretability Issue

In Artificial Intelligence (AI), interpretability is commonly defined as "the ability to explain or to provide the meaning in understandable terms to a human" (Doshi-Velez and Kim, 2017; Guidotti et al., 2018; Barredo Arrieta et al., 2020). This definition encompasses both the challenge of understanding AI systems' decision-making processes and the challenge of interpreting models' outputs. This thesis focuses on the latter. Accordingly, when discussing interpretability, our primary interest lies in understanding the meaning of metric scores rather than the internal mechanisms of the metrics themselves.

In light of this, we attribute MT metrics' lack of interpretability to three main factors:[1]

1. **Error attribution**: Scalar assessments do not identify specific translation errors.

2. **Range consistency**: It is unclear whether a difference in metric score has the same

[1]We wish to clarify that we identified these three factors as notably impactful, but they are non-exhaustive and may overlap.

meaning if it occurs in different regions of the score range.

3. **Performance**: Metric capabilities are typically measured through correlation with human judgment, which fails to provide users with a clear understanding of their performance and reliability.

In simpler terms, consider the example of Figure 1.1, which shows the assessments returned by three popular MT metrics. Two of these metrics – namely, COMET (Rei et al., 2020) and MetricX-23-QE (Juraska et al., 2023) – produce assessments exclusively in the form of scalar scores of translation quality, while GEMBA-MQM (Kocmi and Federmann, 2023a) additionally returns error spans annotated with a category and severity. Due to the lack of *Error attribution*, it is unclear which translation errors, if any, led COMET and MetricX-23-QE to return 0.86 and -1.49, respectively. By contrast, GEMBA-MQM allows us to easily verify whether the presented translation contains a critical mistranslation – which, in this case, it clearly does not. Moreover, COMET and MetricX-23-QE provide no *Range consistency* guarantee; for example, it is unclear whether a score of 0.86 is twice as good as 0.43. In addition, they operate on different score ranges, making it difficult to compare assessments across MT metrics. Finally, for all metrics, regardless of their output format, we lack an understanding of their *Performance* beyond correlation with human judgments, making it unclear whether we can draw conclusions from their assessments safely.

1.1.1 Human Evaluation and MT Benchmarks

Despite the advancements in automatic evaluation, human evaluation remains the gold standard for measuring progress in the field. However, the improvements in translation quality brought about by LLMs also pose new challenges for human annotators, as it has become increasingly difficult to distinguish between translations that differ only in subtle nuances. In this respect, evaluation accuracy can be improved along two dimensions: increasing the quality of human annotations or enhancing the discriminative power of the selected test data.

In recent years, substantial research has focused on refining human MT evaluation, with standard human evaluation practices undergoing considerable changes. Early editions of the Conference on Machine Translation (WMT[2]) collected annotations via crowd-sourcing,

[2]WMT is the main event for Machine Translation and related research. It was a workshop until 2016, when

tasking monolingual, non-expert annotators with comparing translations to manually curated references (Bojar et al., 2017a, 2018). Moreover, annotators were provided only with the translation and its corresponding reference translation, without any access to the full document context from which the source sentence was drawn (Läubli et al., 2018). In contrast, more recent evaluation campaigns employ professional translators and expert bilingual annotators who compare translations directly to their source texts. Annotators also have access to the full source document context, allowing for a more accurate assessment that accounts for cross-sentential phenomena (Kocmi et al., 2023, 2024a).

Instead, less effort has been devoted to ensuring that the test data remain sufficiently challenging for current systems. In this direction, recent research works and WMT editions have evaluated MT systems and metrics focusing on specialized domains, multimodal settings, and long-text translation (Zouhar et al., 2024; Kocmi et al., 2024a; Wu et al., 2025). Nonetheless, at WMT24, many translations still received perfect scores from human annotators – particularly in high-resource translation directions such as English to Spanish – indicating that such a benchmark may remain too easy for current MT systems and LLMs. This limits our ability to reliably measure progress in the field, as benchmarks that are either too easy or too difficult may fail to effectively distinguish between systems (Lalor et al., 2018; Rodriguez et al., 2021).

1.2 Thesis Outline

The work discussed in this dissertation aims to address the aforementioned issues in MT evaluation with the goal of making evaluation processes more accurate and interpretable.

We begin by tackling the lack of interpretability in neural MT evaluation techniques. As previously mentioned, most neural metrics are trained to predict translation quality by producing single-score assessments, which lack *Error Attribution*. To address this issue, Chapter 3 introduces the MaTESe metrics: neural MT metrics trained to identify error spans in translations via span tagging.

In Chapter 4, we address another factor contributing to the lack of interpretability in MT evaluation: the *Performance* issue. To provide more intuitive insights into metric

it became a conference.

capabilities beyond correlation with human judgment, we evaluate MT metrics in two practical scenarios that simulate common applications of evaluation metrics, namely data filtering and translation re-ranking.

Our contributions to the evaluation of the evaluation metrics continue in Chapter 5, where we uncover flawed meta-evaluation practices. We demonstrate that some widely used meta-evaluation strategies are confounded by external factors – specifically, spurious correlations between text features and human judgments – which reward metrics for the wrong reasons. Furthermore, we find that acc_{eq}, a meta-evaluation measure used at WMT 2023 and WMT 2024 (Freitag et al., 2023, 2024), systematically favors metrics that output continuous scores over those that produce discrete assessments. Such inaccuracies in evaluation risk steering the research community toward optimizing translation systems and evaluation metrics for misleading objectives. Indeed, in Section 5.2.4, we show that current metrics may be learning to evaluate based on the same spurious correlations that compromise the meta-evaluation.

While the earlier chapters address MT evaluation issues by focusing on evaluation and meta-evaluation techniques, the later chapters shift attention to test data. Specifically, Chapter 6 explores test data difficulty, introducing Translation Difficulty Estimation as the task of automatically identifying texts that are challenging for MT systems to translate. We also train, evaluate, and release the first difficulty estimators, thereby establishing the initial state of the art for the task. Finally, Chapter 7 discusses our impact on the 2025th edition of the Conference on Machine Translation (WMT). We place particular emphasis on our contribution to the General Machine Translation Shared Task (Kocmi et al., 2025c), where – building on our work in Translation Difficulty Estimation – we employ our novel, state-of-the-art difficulty estimator, sentinel-src-25, to sample difficult-to-translate texts for inclusion in the shared task's test set.

1.3 Publications and Contributions

The research discussed in this dissertation was carried out while I was a PhD student in Sapienza NLP at Sapienza University, under the supervision of Professor Roberto Navigli. During this period, I also visited the University of Zurich (UZH), where I conducted research

under the supervision of Professor Rico Sennrich, and completed a 6-month internship as an Applied Scientist at Amazon Madrid.

In this thesis, we discuss the following publications:

1. Perrella et al. (2022): **Stefano Perrella**, Lorenzo Proietti, Alessandro Scirè, Niccolò Campolungo, and Roberto Navigli. 2022. MaTESe: Machine Translation Evaluation as a Sequence Tagging Problem. In Proceedings of the Seventh Conference on Machine Translation (WMT), pages 569–577, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

2. Perrella et al. (2024a): **Stefano Perrella**, Lorenzo Proietti, Pere-Lluís Huguet Cabot, Edoardo Barba, and Roberto Navigli. 2024. Beyond Correlation: Interpretable Evaluation of Machine Translation Metrics. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 20689–20714, Miami, Florida, USA. Association for Computational Linguistics.

3. Perrella et al. (2024b): **Stefano Perrella**, Lorenzo Proietti, Alessandro Scirè, Edoardo Barba, and Roberto Navigli. 2024. Guardians of the Machine Translation Meta-Evaluation: Sentinel Metrics Fall In!. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 16216–16244, Bangkok, Thailand. Association for Computational Linguistics.

4. Proietti et al. (2025b): Lorenzo Proietti, **Stefano Perrella**, Vilém Zouhar, Roberto Navigli, and Tom Kocmi. 2025. Estimating Machine Translation Difficulty. In Findings of the Association for Computational Linguistics: EMNLP 2025, pages 24261–24285, Suzhou, China. Association for Computational Linguistics.

Furthermore, the work accomplished throughout the Ph.D. resulted in additional publications that are not explicitly discussed in this dissertation:

1. Proietti et al. (2025a): Lorenzo Proietti, **Stefano Perrella**, and Roberto Navigli. 2025. Has Machine Translation Evaluation Achieved Human Parity? The Human Reference and the Limits of Progress. In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 790–813, Vienna, Austria. Association for Computational Linguistics.

2. Martelli et al. (2024): Federico Martelli, **Stefano Perrella**, Niccolò Campolungo, Tina Munda, Svetla Koeva, Carole Tiberius, and Roberto Navigli. 2025. DiBiMT: A Gold Evaluation Benchmark for Studying Lexical Ambiguity in Machine Translation. Computational Linguistics, 51:343–413.

3. Proietti et al. (2024): Lorenzo Proietti, **Stefano Perrella**, Simone Tedeschi, Giulia Vulpis, Leonardo Lavalle, Andrea Sanchietti, Andrea Ferrari, and Roberto Navigli. 2024. Analyzing Homonymy Disambiguation Capabilities of Pretrained Language Models. In Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), pages 924–938, Torino, Italia. ELRA and ICCL.

4. Ahmadi et al. (2025): Sina Ahmadi, Micha David Hess, Elena Álvarez-Mellado, Alessia Battisti, Cui Ding, Anne Göhring, Yingqiang Gao, Zifan Jiang, Andrianos Michail, Peshmerge Morad, Joel Niklaus, Maria Christina Panagiotopoulou, **Stefano Perrella**, Juri Opitz, Anastassia Shaitarova, and Rico Sennrich. 2025. ConLoan: A Contrastive Multilingual Dataset for Evaluating Loanwords. In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 30070–30090, Vienna, Austria. Association for Computational Linguistics.

5. Kocmi et al. (2025c): Tom Kocmi, Ekaterina Artemova, Eleftherios Avramidis, Rachel Bawden, Ondrej Bojar, Konstantin Dranch, Anton Dvorkovich, Sergey Dukanov, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Howard Lakougna, Jessica M. Lundin, Christof Monz, Kenton Murray, Masaaki Nagata, **Stefano Perrella**, Lorenzo Proietti, Martin Popel, Maja Popovic, Parker Riley, Mariya Shmatova, Steinþór Steingrímsson, Lisa Yankovskaya, and Vilém Zouhar. Findings of the WMT25 General Machine Translation Shared Task: Time to Stop Evaluating on Easy Test Sets. In Proceedings of the Tenth Conference on Machine Translation, pages 355–413, Suzhou, China. Association for Computational Linguistics.

6. Kocmi et al. (2025a): Tom Kocmi, Sweta Agrawal, Ekaterina Artemova, Eleftherios Avramidis, Eleftheria Briakou, Pinzhen Chen, Marzieh Fadaee, Markus Freitag,

Roman Grundkiewicz, Yupeng Hou, Philipp Koehn, Julia Kreutzer, Saab Mansour, **Stefano Perrella**, Lorenzo Proietti, Parker Riley, Eduardo Sánchez, Patricia Schmidtova, Mariya Shmatova, and Vilém Zouhar. Findings of the WMT25 Multilingual Instruction Shared Task: Persistent Hurdles in Reasoning, Generation, and Evaluation. In Proceedings of the Tenth Conference on Machine Translation, pages 462–483, Suzhou, China. Association for Computational Linguistics.

Finally, I wish to highlight that most of the research discussed in this dissertation was conducted in collaboration with other PhD students and researchers. In particular, I gratefully acknowledge my collaboration with Lorenzo Proietti – from the Sapienza NLP group – who contributed substantially to a significant portion of the presented work. To clarify my individual role, I summarize my contributions as follows:

1. **Chapter 3.** I designed the MaTESe metrics, trained and evaluated them, and conducted all the experiments.

2. **Chapter 4.** I designed the complete meta-evaluation framework and conducted all the additional analyses.

3. **Chapter 5.** I proposed the core idea of studying the impact of spurious correlations in machine translation evaluation; I designed the first three sentinel metrics, trained and evaluated them, and conducted the associated experiments.

4. **Chapter 6.** I proposed the idea of using difficulty estimators to construct challenging test sets for machine translation; I performed preliminary analyses, designed the translation difficulty estimation task, and implemented the artificial crowd difficulty estimators.

Chapter 2

Background

This chapter provides the background knowledge for the thesis's contributions. We begin with an overview of Machine Translation (MT) evaluation, covering the most popular human and automatic approaches. Subsequently, we discuss the MT meta-evaluation, that is, the process of evaluating the evaluation metrics.

2.1 Machine Translation Evaluation

MT evaluation involves assessing the quality of translated text. Given a source text, its translation, and, optionally, a reference translation, human or automatic evaluators assess the quality of the translation by comparing it with its source or reference, or both. Within this thesis, we use the term "metrics" to refer to automatic evaluators, while human evaluators may also be frequently referred to as "human judges".

2.1.1 Human Evaluation

Human MT evaluation involves human annotators assessing the quality of translated text. In this section, we illustrate the most popular human evaluation protocols, from the earliest ranking-based approaches to the more expensive and accurate span-based evaluation protocols.

Relative Ranking

The Relative Ranking (RR) evaluation protocol tasks human annotators to compare translations of the same source text (Bojar et al., 2013, 2014, 2015, 2016a). At the 2013th edition of WMT, RR annotations were collected partly by researchers (participants to the WMT shared tasks) and partly via crowd-sourcing non-expert annotators using Amazon Mechanical Turk.[1] Instead, WMT editions from 2014 to 2016 employed researchers only, based on the findings of Bojar et al. (2013) that annotations collected by researchers displayed a higher inter-annotator agreement.

In RR, annotators are presented with a source sentence, a reference translation, and the outputs of up to five anonymized and randomly ordered MT systems. They are asked to rank such outputs according to their quality, allowing for ties. Each ranking yields up to $\binom{5}{2} = 10$ pairwise comparisons per group of translations. Each pairwise comparison can result in one of three outcomes: system 1 wins, system 2 wins, or the two systems tie. Figure 2.1 shows the interface used by human annotators to perform RR for MT evaluation (Bojar et al., 2014).

Sentence-level pairwise comparisons are then used to derive system-level scores, i.e., scores assigned to MT systems to determine which system is superior. To this end, Bojar et al. (2013) computes system-level scores by measuring the proportion of times a system wins against others (excluding ties) and averaging results across the systems that participate in the comparisons. Instead, subsequent WMT editions Bojar et al. (2014, 2015, 2016a) adopt the TrueSkill algorithm – an approach originally developed to model the relative skills of players in ongoing competitions (e.g., Microsoft's Xbox Live), and later adapted by Sakaguchi et al. (2014) to MT evaluation.

However, since RR relies solely on pairwise comparisons, it measures relative difference in performance and cannot infer the absolute quality of the MT systems.

Direct Assessment

Direct Assessment (DA, Graham et al., 2013) asks human annotators to rate translation quality on a slider, with scores from 0 to 100. By relying on absolute quality scores, this methodology addresses the main limitation of RR, which provides only relative comparisons

[1] `https://www.mturk.com/`.

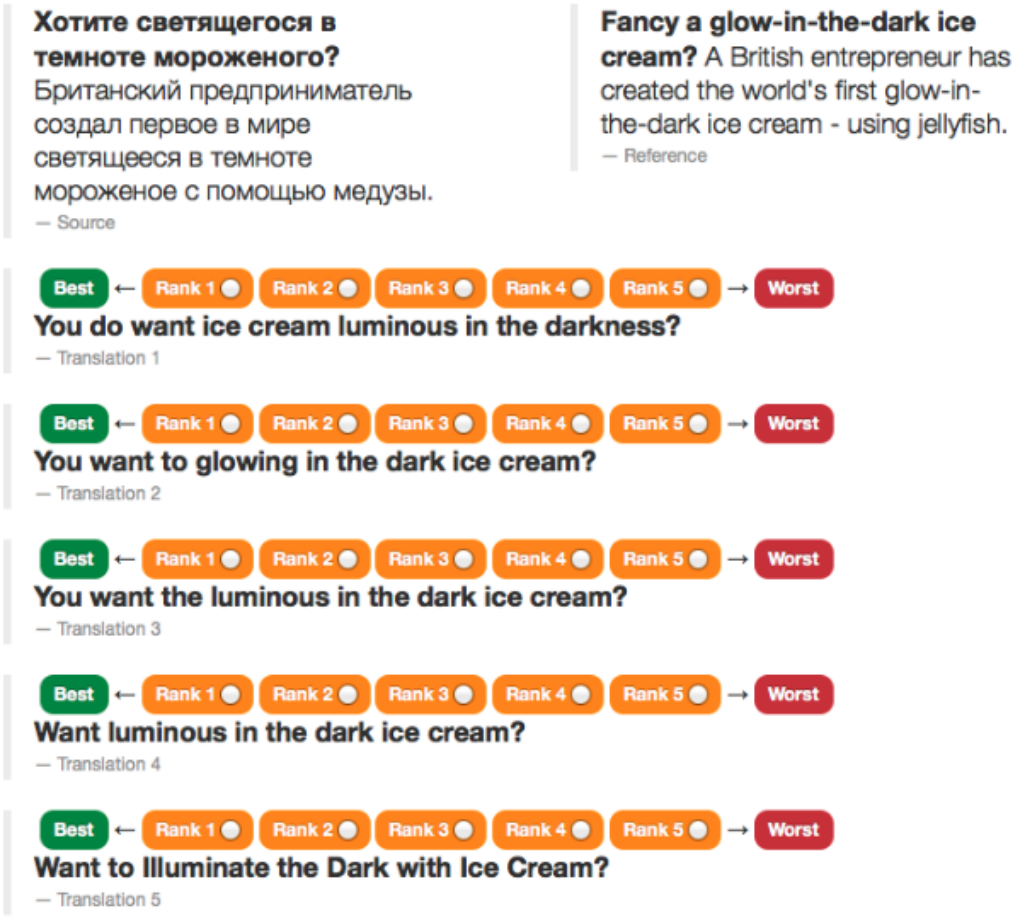


Figure 2.1. Appraise interface used for the evaluation campaign conducted by Bojar et al. (2014). Appraise is an open-source toolkit for manual evaluation of MT outputs (Federmann, 2012). This screenshot is taken from Bojar et al. (2014).

between MT systems without indicating their absolute quality. Figure 2.2 shows a screenshot of the Appraise interface used by human annotators to perform DA-based MT evaluation.

The WMT evaluation campaigns have adopted DA as their primary human evaluation protocol since 2017 (Bojar et al., 2017a, 2018; Barrault et al., 2019, 2020; Akhbardeh et al., 2021; Kocmi et al., 2022), following pilot study by Bojar et al. (2016a) that demonstrated a strong correlation between DA annotations collected via crowd-sourcing and RR annotations collected by researchers. Consequently, in addition to contributions from researchers in participating teams, successive WMT editions have also collected a portion of DA annotations through crowd-sourced non-expert annotators on Amazon Mechanical Turk. Starting from WMT 2020, professional translators have also been employed to provide annotations.

Over the years, several variants of DA have been explored. The original DA setup provided monolingual annotators with only a system translation and a manually-curated reference. Later, Bojar et al. (2018) introduced source-based DA, where bilingual annotators are shown translations alongside the source texts in the original language. This variant enabled the evaluation of reference translations as well, allowing comparisons between human and system translations. As a result, from WMT 2018 onward, source-based DA using bilingual annotators has been systematically employed to evaluate translations out-of-English.[2] Furthermore, after Läubli et al. (2018) challenged early claims of human parity (Hassan et al., 2018) by highlighting the importance of full document context, WMT editions from 2019 onward have collected a portion of DA annotations in which annotators are given access to the full document from which the text was sourced.

Multidimensional Quality Metrics

Multidimensional Quality Metrics (MQM, Freitag et al., 2021a) tasks annotators with identifying error spans in translations and categorizing them by error type and severity (Lommel et al., 2014b; Freitag et al., 2021a). A final score is computed by weighting errors based on their severity, which can be `Neutral`, `Minor`, or `Major`. Freitag et al. (2021a)

[2]Source-based evaluation was used primarily for out-of-English language directions because it requires bilingual annotators who are native speakers of the target language, and it was easier to recruit non-English speakers proficient in English rather than the reverse.

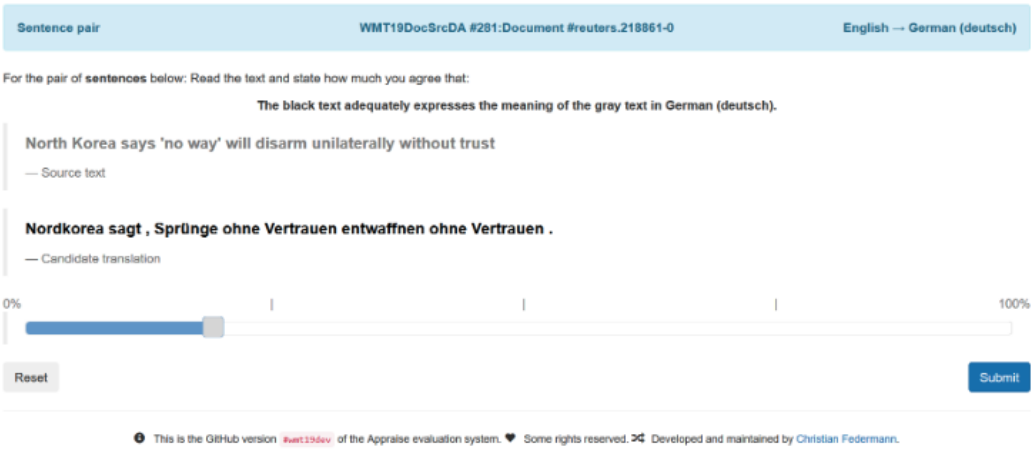


Figure 2.2. Appraise interface used for a part of the evaluation campaign conducted by Barrault et al. (2019). Appraise is an open-source toolkit for manual evaluation of MT outputs (Federmann, 2012). This screenshot is taken from Barrault et al. (2019).

proposed assigning no penalty to `Neutral` errors, and penalties of -1 and -5 to `Minor` and `Major` errors, respectively. Some implementations of the MQM protocol also include `Critical` errors; however, these are typically assigned the same weight as `Major` ones to maintain consistency across annotations collected by different providers (Freitag et al., 2022b, 2023, 2024). The MQM guidelines provided to annotators by Freitag et al. (2021a) are shown in Table 2.1, and the full list of error categories used in their study is presented in Table 2.2.

Compared to RR and DA, the MQM protocol has the advantage of producing interpretable assessments, as span-level annotations explicitly explain the scores assigned to each translation. Furthermore, decomposing the task into specific error types seems to facilitate more consistent evaluations across annotators, yielding higher inter-annotator agreement than other protocols (Freitag et al., 2021a). However, MQM is substantially more time-consuming and costly than other protocols, and requires extensive annotator training. Indeed, the MQM annotations used at the WMT Metrics Shared Tasks from 2021 to 2024 have been produced by professional annotators, rather than crowd-sourced non-expert annotators or researchers from the teams participating in the shared tasks (Freitag et al., 2021b, 2022b, 2023, 2024).

Table 2.1. Annotation guidelines provided to MQM annotators by Freitag et al. (2021a).

MQM Guidelines
You will be assessing translations at the segment level, where a segment may contain one or more sentences. Each segment is aligned with a corresponding source segment, and both segments are displayed within their respective documents. Annotate segments in natural order, as if you were reading the document. You may return to revise previous segments. Please identify all errors within each translated segment, up to a maximum of five. If there are more than five errors, identify only the five most severe. If it is not possible to reliably identify distinct errors because the translation is too badly garbled or is unrelated to the source, then mark a single Non-translation error that spans the entire segment. To identify an error, highlight the relevant span of text, and select a category/sub-category and severity level from the available options. (The span of text may be in the source segment if the error is a source error or an omission.) When identifying errors, please be as fine-grained as possible. For example, if a sentence contains two words that are each mistranslated, two separate mistranslation errors should be recorded. If a single stretch of text contains multiple errors, you only need to indicate the one that is most severe. If all have the same severity, choose the first matching category listed in the error typology (eg, Accuracy, then Fluency, then Terminology, etc). Please pay particular attention to document context when annotating. If a translation might be questionable on its own but is fine in the context of the document, it should not be considered erroneous; conversely, if a translation might be acceptable in some context, but not within the current document, it should be marked as wrong. There are two special error categories: Source error and Non-translation. Source errors should be annotated separately, highlighting the relevant span in the source segment. They do not count against the five-error limit for target errors, which should be handled in the usual way, whether or not they resulted from a source error. There can be at most one Non-translation error per segment, and it should span the entire segment. No other errors should be identified if Non-Translation is selected.

Direct Assessments + Scalar Quality Metrics

Direct Assessment + Scalar Quality Metrics (DA+SQM, Kocmi et al., 2022, 2023) extends the DA protocol by adding seven marked levels to the 0-100 slider used to assign translation

Table 2.2. The MQM Hierarchy used by Freitag et al. (2021a).

Error Category	Subcategory	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 (e.g., 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	Other	Any other issues.
	Source error	An error in the source.
	Non-translation	Impossible to reliably characterize distinct errors.

quality scores, with each level accompanied by a brief description of the corresponding quality. Kocmi et al. (2022) replaced DA with DA+SQM after Freitag et al. (2021a) showed that MT evaluation based on discrete quality levels (from 0 to 6) correlates more strongly with MQM than DA.

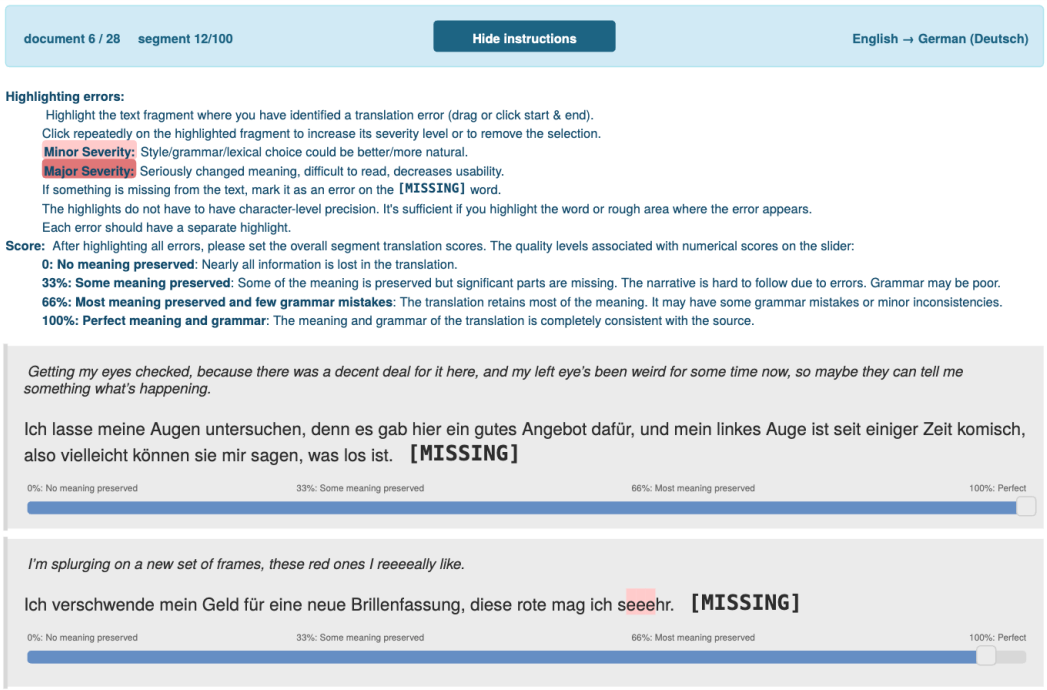


Figure 2.3. ESA interface used by Kocmi et al. (2024b), from which this screenshot is taken.

Error Span Annotation

Error Span Annotation (ESA, Kocmi et al., 2024b) combines elements of DA and MQM. Annotators first identify translation errors, marking their severity as `Minor` or `Major`, and then assign an overall score on a 0-100 scale, informed by the detected errors. ESA strikes a balance between the low annotation cost and efficiency of DA and the higher reliability of MQM.

Unlike MQM, ESA annotators do not provide detailed error categories, which reduces the interpretability of the resulting annotations. They also do not annotate omissions directly on the source text, but instead use a dedicated tag to indicate when an omission is present, without specifying its position. Kocmi et al. (2024b) report that ESA is approximately 32% faster than MQM and does not require extensive annotator training. Figure 2.3 shows the annotation interface used in their study.

The human evaluation conducted at WMT 2024 employed bilingual annotators instructed to conducted the evaluation using the ESA protocol.[3] Annotators were professional

[3]WMT 2024 used ESA for all language directions except Japanese-to-Chinese and English-to-German, for which the MQM protocol was used.

translators and were provided with extended source context alongside the source text and the translation to be evaluated (Kocmi et al., 2024a).

2.1.2 Automatic Evaluation

Automatic MT evaluation refers to the process of assessing the quality of translated text automatically, serving as a proxy for human judgment. Because it is significantly faster and less costly than human evaluation, automatic evaluation has become a crucial component of MT research, enabling rapid experimentation and iterative system development. This section provides an overview of automatic MT evaluation techniques, tracing their evolution from lexical-based approaches to the contemporary state-of-the-art methods that leverage neural models. Furthermore, Table 2.3 summarizes the automatic evaluation techniques discussed, highlighting their main strengths and limitations.

Lexical-Based Metrics

Lexical-based metrics evaluate translation quality by measuring the degree of lexical similarity between a candidate translation and one or more reference translations. The most popular MT metric in this category is BLEU (Papineni et al., 2002), which is a modified precision measure that computes the number of overlapping n-grams between a candidate translation and one or more references. For many years, BLEU has been the dominant MT evaluation metric. Other well-known lexical-based metrics include METEOR (Banerjee and Lavie, 2005) and chrF (Popović, 2015). METEOR aligns words in the candidate and reference translation, incorporating stemming-, synonymy-, and paraphrase-based matching. chrF instead operates at the character level and compares a translation and its reference based on the number of overlapping character n-grams.

Lexical-based metrics are computationally efficient and fully interpretable. Since they rely on simple, non-neural heuristics, the reason behind their assessments can be easily traced. However, they suffer from a fundamental limitation: an over-reliance on surface-form matching, which prevents them from capturing semantic equivalence beyond lexical similarity. Moreover, they require manually curated reference translations, which are expensive to produce and inherently limited in their ability to represent the full spectrum of valid translations.

Neural-Based Metrics

To overcome the main limitation of lexical-based metrics – i.e., the limited extent to which they can capture semantic equivalence via lexical similarity – neural metrics leverage semantically rich representations derived from pre-trained embeddings or from neural models fine-tuned specifically for MT evaluation. Popular examples include BERTscore (Zhang et al., 2020), MoverScore (Zhao et al., 2019), COMET (Rei et al., 2020, 2022a), BLEURT-20 (Sellam et al., 2020; Pu et al., 2021), UniTE (Wan et al., 2022), and MetricX (Juraska et al., 2023, 2024). BERTscore and MoverScore are unsupervised metrics that compute similarity using contextual embeddings extracted from pre-trained language models such as BERT (Devlin et al., 2019). In contrast, COMET, BLEURT-20, UniTE, and MetricX are supervised metrics that are explicitly fine-tuned for the MT evaluation task.

By leveraging multilingual pre-trained language models such as XLM-RoBERTa (Conneau et al., 2020), RemBERT (Chung et al., 2021), and mT5 (Xue et al., 2021), neural metrics provide assessments that extend beyond surface-level similarity and capture deeper semantic relationships between translations and references. Indeed, empirical studies have shown that these metrics correlate more strongly with human judgments (Freitag et al., 2022b, 2023, 2024) than lexical-based ones. Moreover, neural metrics have enabled the development of reference-free (or reference-less) models, i.e., metrics capable of evaluating translation quality without reference translations. This capability has allowed such metrics to function as effective proxies for human evaluators even in inherently reference-free tasks, such as filtering parallel corpora based on translation quality (Peter et al., 2023; Alves et al., 2024).

Despite these advancements, neural metrics also present notable limitations. They are computationally more expensive than lexical-based metrics and, more importantly, their scores are difficult to interpret due to the black-box nature of the underlying neural architectures. In the remainder of this section, we provide a brief description of the most popular neural metrics that we use or evaluate in this dissertation.

BERTscore. BERTscore is one of the earliest neural-based approaches to MT evaluation – and to Natural Language Generation (NLG) evaluation more broadly (Zhang et al., 2020). This metric leverages contextual embeddings from pre-trained BERT and multilingual BERT

models (Devlin et al., 2019) to quantify the semantic similarity between texts. Given a translation and a manually-curated reference (or the source text, when using a multilingual model to extract token embeddings), BERTscore computes their similarity using the cosine similarity between their token embeddings. Specifically, BERTscore constructs a mapping between translation and reference tokens, greedily matching each translation token to the most similar reference token, and vice versa. Precision is then calculated as the average cosine similarity between translation tokens and their matched reference tokens, and recall is defined analogously, averaging cosine similarities between reference tokens and their best matches in the translation. Finally, BERTscore combines precision and recall into a single F-score, which serves as the overall measure of similarity.

The COMET family of metrics. Most models from the COMET family are based on the estimator architecture introduced by Rei et al. (2020), which fine-tunes a pre-trained language model to predict human-annotated quality scores. Specifically, the COMET estimator employs XLM-RoBERTa (Conneau et al., 2020) to extract sentence embeddings for the source text, the translation, and the reference. These embeddings are then combined via element-wise product and difference operations to capture relationships among the three inputs. The resulting features are concatenated and passed through a feed-forward regression module. The entire model is trained end-to-end using Mean Squared Error (MSE) to predict scalar assessments of translation quality. The metric we refer to as COMET is based on XLM-Roberta large and is trained on the DA human judgments released at WMT between 2017 and 2020 (Bojar et al., 2017b; Ma et al., 2018, 2019; Mathur et al., 2020b).

Other metrics from the COMET family include reference-free versions, larger models, and ensembles:

- COMET-QE and COMET-QE-MQM are two regression-based metrics trained to evaluate translation quality without the use of reference translations (Rei et al., 2021). COMET-QE-MQM is further fine-tuned on a split of the MQM-based corpus released by Freitag et al. (2021a), after the initial training on DA.

- CometKiwi and CometKiwi-XL (Rei et al., 2022b, 2023) are reference-free regression-based metrics, built upon InfoXLM (Chi et al., 2021) and XLM-RoBERTa XL (Goyal et al., 2021), respectively. These metrics are trained using DA-based human judgments

released at WMT from 2017 to 2020, as well as DA from the MLQE-PE corpus (Fomicheva et al., 2022). CometKiwi-XL's training data also include the DA for Indian languages released by Blain et al. (2023). Different from other COMET models, CometKiwi and CometKiwi-XL encode the source text and its translation jointly, rather than encoding them in isolation.

- XCOMET-XL and XCOMET-XXL are regression-based metrics built upon XLM-RoBERTa XL and XLM-RoBERTa XXL, respectively, and trained on the concatenation of DA-based human judgments released at WMT from 2017 and 2020 and the MLQE-PE dataset, and further fine-tuned using MQM-based annotations coming from the following datasets: i) WMT data from 2020 to 2022, ii) IndicMT (Sai B et al., 2023), and iii) DEMETR (Karpinska et al., 2022). Given a candidate translation, XCOMET simultaneously identifies its error spans and assigns it a scalar quality score. Similar to CometKiwi models, XCOMET jointly encodes the translation and its source, reference, or both, rather than encoding them in isolation like other models of the COMET family. XCOMET-ENSEMBLE and XCOMET-QE-ENSEMBLE are ensembles between one XL and two XXL XCOMET checkpoints that result from different training stages (Freitag et al., 2023).

BLEURT-20. BLEURT-20 is a regression-based metric built upon the RemBERT pre-trained language model (Chung et al., 2021). RemBERT is fine-tuned to predict translation quality scores using DA-based human assessments from 2015 to 2019, along with synthetic data (Sellam et al., 2020; Pu et al., 2021). Different from models from the COMET family, BLEURT-20 takes as input only the translation and its reference, without using the source text.

The MetricX family of metrics. MetricX metrics (Juraska et al., 2023, 2024) are regression-based metrics built upon the mT5 pre-trained encoder-decoder models (Xue et al., 2021). These metrics are first trained using DA-based human judgments and further fine-tuned on a combination of MQM-based human judgments and synthetic data. Juraska et al. (2024) introduces hybrid variants of MetricX by training mT5 models to conduct the evaluation both in the presence and absence of reference translations, thereby enabling reference-based

and reference-free evaluation using the same model.

LLM-based Metrics

Large Language Models (LLMs) have revolutionized most Natural Language Processing tasks, including Machine Translation and its evaluation. Owing to their strong reasoning and language understanding capabilities, LLMs can be instructed to assess translations in ways that closely resemble human annotation. For example, GEMBA (Kocmi and Federmann, 2023b) uses models from the GPT family (Radford et al., 2019; Brown et al., 2020; OpenAI et al., 2024) to evaluate translations under various annotation schemes, including DA and SQM. Similarly, GEMBA-MQM (Kocmi and Federmann, 2023a) instructs GPT-4 (OpenAI et al., 2024) to perform translation evaluation following MQM-like guidelines.

Beyond simple prompting, other studies have explored fine-tuning LLMs specifically for MT evaluation. For instance, Xu et al. (2023) fine-tune a Llama-7B model (Touvron et al., 2023) using synthetic data – covering both error detection and error explanation – generated by GPT-4. Fernandes et al. (2023) investigate multiple strategies, including zero-shot and few-shot prompting as well as fine-tuning. Their fine-tuning experiments use Palm (Chowdhery et al., 2023) and Palm2 (Anil et al., 2023), trained under two setups: a regression setup, where the embedding of a special vocabulary token is passed to a regression head to produce a scalar score, and a classification setup, where translations are categorized into five quality classes: `very bad`, `bad`, `ok`, `good`, `very good`. More recent work takes this further by treating LLMs as collaborative agents for MT evaluation. For example, Feng et al. (2025); Zhang et al. (2025) propose multi-agent frameworks in which LLMs engage in structured discussions to assess translation quality.

LLM-based evaluation has demonstrated very high correlations with human judgment, achieving performance comparable to or even surpassing that of dedicated neural metrics. Moreover, certain LLM-based approaches go beyond scalar scoring by explicitly identifying error spans and, in some cases, providing natural-language explanations for their assessments. This makes their evaluation more interpretable and informative than that produced by regression-based neural models.

However, like the other approaches discussed so far, LLM-based MT evaluation also presents several limitations. (i) Computational cost. LLMs typically perform evaluations

by generating tokens autoregressively, one at a time, which is substantially slower and more costly than the single-pass inference of neural metrics. Moreover, when using chain-of-thought prompting, which has been shown to improve the ability to perform complex reasoning (Wei et al., 2022), LLMs generate extensive intermediate text before reaching a final decision, further increasing runtime and resource consumption. Finally, approaches employing agentic evaluation settings, where multiple models engage in extended discussion, can incur even higher computational costs. (ii) Systematic biases. LLMs may exhibit biases toward certain translations for spurious or inappropriate reasons. Zheng et al. (2023), for instance, identified several consistent biases in LLM-based judges for open-ended tasks, including position bias (favoring items in particular input positions, verbosity bias (preferring longer responses), and self-enhancement bias (favoring outputs produced by the same model. (iii) Interpretability. Although LLM-based evaluations are often more interpretable than other neural approaches, their decision-making process remains largely opaque: LLMs are still prone to hallucinations (Huang et al., 2025) and their explanations, while often plausible and well-articulated, can be misleading or factually inaccurate, posing challenges for reliable interpretability.

Translation Re-ranking with MT Metrics

Besides their traditional use in measuring incremental improvements during MT systems development, MT metrics are also employed for other applications, among which translation re-ranking is one of the most popular. In translation re-ranking, metrics are used to select the best translation from a pool of candidate translations of the same source text.[4] This is straightforward when using reference-free metrics; in that case, translation re-ranking is referred to as *Quality Estimation (QE) re-ranking*, where the quality of each translation is estimated using a reference-free metric, and the translation with the highest score is selected.

In contrast, when using reference-based metrics, translation re-ranking is typically performed via Minimum Bayes Risk (MBR) decoding. MBR decoding selects the hypothesis with the highest expected utility from a set of hypotheses sampled from the same distribution (Eikema and Aziz, 2022). In MT, the expected utility is commonly approximated by averaging the scores of a reference-based metric across a sample of candidate translations

[4]Chapter 4 of this manuscript evaluates MT metrics in the translation re-ranking task.

Table 2.3. Summary of automatic MT evaluation techniques.

	Lexical	**Neural**		**LLM-based**	
		Unsupervised	**Supervised**	**LLM-as-a-Judge**	**Agentic**
Core idea	Surface overlap between hypothesis and reference(s).	Similarity between contextual embeddings.	Pretrained models fine-tuned to predict human judgments.	LLM-as-a-judge for translation quality (typically with no fine-tuning).	LLMs discuss and collaborate to assess translation quality.
Needs reference?	Yes	Optional	Optional	Optional	Optional
Metrics	BLEU, METEOR, chrF	BERTscore, MoverScore	COMET, UniTE, BLEURT-20, CometKiwi, XCOMET, MetricX	GEMBA, GEMBA-MQM	M-MAD, HiMATE
Main strengths	Fast; cheap; interpretable.	Better semantics understanding than lexical.	High correlation with human judgments.	High correlation with human judgments; higher interpretability than neural metrics.	High correlation with human judgments; higher interpretability than neural metrics.
Main Limitations	Need references; Limited to surface-level similarity.	Lower correlation with human judgments than fine-tuned and LLM-based metrics.	Interpretability.	High cost; slow.	Extremely high cost; extremely slow.

generated by the same MT system. In this setting, the pool of candidate translations is used both as the set of hypotheses and as a set of pseudo-references.

Formally, let x be a source text and $\mathcal{H} = \{h_1, h_2, \ldots, h_N\}$ be a set of translations of x generated by the same MT system. Let also $u(h, y)$ be our utility function, that is, a reference-based MT metric that assigns a quality score to h by comparing it to a reference y. The MBR score of hypothesis h_i is defined as follows:

$$\mathrm{MBR}(h_i) = \frac{1}{N} \sum_{j=1}^{N} u(h_i, h_j) \tag{2.1}$$

The selected translation is the one with the highest MBR score:

$$h^* = \arg\max_{h \in \mathcal{H}} \mathrm{MBR}(h) \tag{2.2}$$

2.2 Machine Translation Meta-Evaluation

As discussed above, research in MT evaluation has led to a variety of approaches, ranging from lexical-based metrics to those based on LLMs. Furthermore, metrics within each category can differ depending on which language models are used, whether or not reference translations are required, and other metric-specific design choices. In this context, it is important to assess which metrics are most effective and to understand their limitations. MT meta-evaluation addresses these questions by evaluating the capabilities of MT metrics and ranking them based on their alignment with human judgments.

The WMT Metrics Shared Task has long been the primary venue for systematic MT meta-evaluation (Freitag et al., 2021b, 2022b, 2023, 2024). Each year, WMT collects source texts, obtains translations from both automatic systems and human translators, and employs human annotators to assess translation quality. The resulting pool of human annotations is then used not only to evaluate MT systems but also to evaluate MT metrics. Participants in the Metrics Shared Task submit scores produced by their automatic metrics, and the organizers measure their agreement with human judgments at two levels:

- At the **segment level**, metrics are evaluated according to their ability to discern between higher- and lower-quality translations.

- At the **system level**, segment-level metric scores are averaged to yield an overall quality score for each MT system; metrics are then evaluated according to their ability to discriminate between superior and inferior MT systems.

Over the years, several standard correlation measures have been used to quantify agreement between metrics and humans, and several novel measures have also been proposed, as different measures may capture distinct aspects of evaluation quality. To date, designing meta-evaluation measures that are better tailored to the specific challenges of MT meta-evaluation remains an open research problem. Below, we briefly summarize the most commonly used measures, providing more detail on Kendall's correlation and Pairwise Accuracy with Tie Calibration, which require detailed explanation to support the discussion in Chapter 5. Furthermore, we provide a summary of these meta-evaluation strategies in Table 2.4.

Pearson's ρ correlation coefficient

The Pearson correlation coefficient, denoted by ρ, measures the degree of linear association between two sets of measurements. Pearson's ρ has been primarily used to assess system-level correlations between automatic metrics and human judgments, with some applications at the segment level (Ma et al., 2018, 2019; Mathur et al., 2020b; Freitag et al., 2021b, 2022b, 2023). Starting from WMT 2020, Pearson's ρ has been computed only after removing outlier systems, as Mathur et al. (2020a) showed that earlier results were heavily confounded by the sensitivity of Pearson's ρ to outliers.

Kendall's τ correlation coefficient

The Kendall τ correlation coefficient measures the ordinal association between two sets of measurements (Kendall, 1945). Kendall's τ (or minor variations of this measure) has been primarily applied at the segment level to assess the ability of metrics to rank translations in the same order as human judges (Ma et al., 2018, 2019; Mathur et al., 2020b; Freitag et al., 2021b, 2022b). In this paragraph, we formally define the Kendall τ coefficient as employed by the organizers of the Metrics Shared Task at WMT21 and WMT22 (Freitag et al., 2021b, 2022b).

Let $\boldsymbol{m}, \boldsymbol{h}$ be the vectors of metric and human assessments, respectively. *Concordant* pairs are pairs of indices whose corresponding translations were ranked in the same order by metric assessments and human judgments. For example, the pair $(1, 2)$ is concordant if $m_1 > m_2$ and $h_1 > h_2$, or $m_1 < m_2$ and $h_1 < h_2$. Conversely, *discordant* pairs are those ranked in opposite order. We define C and D as the number of concordant and discordant pairs, respectively. We also define T_h as the number of pairs only tied in the gold scores, T_m as the number of pairs only tied in the metric scores, and T_{hm} as the number of pairs tied both in gold and metric scores, i.e., the number of correctly predicted ties. The Kendall's τ correlation coefficient is defined as follows (Kendall, 1945):

$$\tau = \frac{C - D}{\sqrt{(C + D + T_h)(C + D + T_m)}}. \tag{2.3}$$

Pairwise Accuracy with Tie calibration

Segment-level Pairwise Accuracy with Tie Calibration (acc_{eq}) modifies the Kendall correlation coefficient to better account for tied assessments. It evaluates metrics based on their ability to rank translations in the same order as human judgments and to correctly predict ties (Deutsch et al., 2023; Freitag et al., 2023, 2024).

As noted by Deutsch et al. (2023), Kendall's τ penalizes the prediction of ties, but never rewards them, as T_m and T_h are in the denominator, and T_{hm} is not used. This issue was not prominent in the earliest editions of the Metrics Shared Task, where ties in human scores were disregarded, and older metrics rarely produced ties. Currently, instead, it is essential to consider the prediction of ties, especially since human MQM annotations contain a lot of them,[5] and some recently-proposed metrics are designed to output evaluation assessments that resemble MQM. For this reason, Deutsch et al. (2023) proposed a measure that mimics the τ coefficient in the way it is computed, but also accounts for correctly predicting ties:

$$\text{acc}_{\text{eq}} = \frac{C + T_{hm}}{C + D + T_h + T_m + T_{hm}}. \tag{2.4}$$

Different from Kendall's τ, acc_{eq} includes T_{hm} in the numerator, and the denominator encompasses the total number of pairs. Notably, discordant pairs are not subtracted from the numerator, rendering this metric a measure of accuracy rather than a correlation, with scores ranging from 0 to 1.

As follows, we provide a numerical example of the computation of both Kendall's τ and acc_{eq}. Let us consider the vectors $\boldsymbol{m}$ and $\boldsymbol{h}$ with values as in the following table:

$\boldsymbol{m}$	0.6	0.5	0.4	0.4
$\boldsymbol{h}$	5	3	5	5

For each vector, there are six pairs of assessments. In particular, the pairs of metric assessments are (m_1, m_2), (m_1, m_3), (m_1, m_4), (m_2, m_3), (m_2, m_4), (m_3, m_4). In Equations 2.3 and 2.4, $C = 1$, since the only concordant pair is (m_1, m_2). Indeed, $m_1 > m_2$ and $h_1 > h_2$. $D = 2$, since the pairs $(m_2, m_3), (m_2, m_4)$ are discordant. $T_m = 0$, since there are no pairs tied only in the metric scores. $T_h = 2$, since the pairs $(h_1, h_3), (h_1, h_4)$ are tied

[5] This is also due to the increasing quality of automatic translation, as perfect translations are assigned the same maximum score.

only in the human scores. $T_{hm} = 1$, since the remaining pair, i.e., (m_3, m_4), is tied in both human and metric scores. In this example, $\tau = -0.258$ and $\text{acc}_{\text{eq}} = 0.333$.

The acc_{eq} measure, as it stands, would unfairly disadvantage continuous metrics. Indeed, it is extremely infrequent for such metrics to assign the same score to two different translations, meaning that they never predict ties. To address this issue, Deutsch et al. (2023) propose the Tie Calibration algorithm that determines, for each metric, a threshold ϵ such that, given two metric assessments m_1 and m_2, they are tied if $|m_1 - m_2| \leq \epsilon$. Accordingly, for each metric, acc_{eq} is computed as the maximum value of Equation 2.4 over varying values of ϵ.

Nonetheless, in Section 5.3, we show that the Tie Calibration algorithm produces the opposite effect: it favors continuous metrics over discrete ones. As a result, a gap remains in MT meta-evaluation, which still lacks a segment-level evaluation strategy capable of fairly assessing discrete and continuous metrics jointly.

Pairwise Accuracy

System-level Pairwise Accuracy (PA) evaluates metrics based on their ability to rank systems in the same order as human judgments (Kocmi et al., 2021; Freitag et al., 2023). Thompson et al. (2024) showed that PA is equivalent to Kendall's τ correlation coefficient, up to a linear rescaling and shift.

Soft Pairwise Accuracy

System-level Soft Pairwise Accuracy (SPA) extends PA by incorporating metric and human confidence in ranking MT systems. It rewards metrics whose confidence in distinguishing between MT systems aligns with that of humans (Thompson et al., 2024; Freitag et al., 2024).

Table 2.4. Summary of MT meta-evaluation strategies.

	Pearson's ρ	**Kendall's τ**	**acc_{eq}**	**PA**	**SPA**
Evaluation level	System-level and segment-level	System-level and segment-level	Segment-level	System-level	System-level
What it evaluates	Linear correlation between metric and human scores.	Ordinal agreement between metric and human rankings.	Ordinal agreement between metric and human rankings, handling tied scores.	Ordinal agreement between metric and human rankings.	Ordinal agreement between metric and human rankings, incorporating metric and human confidence.
Score range	$[-1, 1]$	$[-1, 1]$	$[0, 1]$	$[0, 1]$	$[0, 1]$
Notes	Sensitive to outliers (Mathur et al., 2020a).	Incomplete handling of tied scores (Deutsch et al., 2023).	Penalizes discrete metrics (Shown in Section 5.3)	Equivalent to Kendall's τ up to a linear rescaling and shift (Thompson et al., 2024).	–

Chapter 3

Interpretable Machine Translation Evaluation

By replacing lexical metrics with neural ones, the MT evaluation field has traded interpretability for performance. In this respect, this chapter addresses the problem of *Score Attribution*, introduced in Section 1.1 in the context of metric interpretability. Specifically, a key factor limiting the interpretability of most MT metrics is that they reduce MT evaluation to a single scalar score, making it unclear which translation errors, if any, contributed to that score. To address this issue, we reframe MT evaluation as a sequence tagging problem by introducing the MaTESe metrics,[1] which identify precise error spans in translations and assign severity levels to them.

3.1 The MaTESe metrics

Freitag et al. (2021a) proposed a shift in the standard practices for human MT evaluation, employing the Multidimensional Quality Metrics framework (MQM Lommel et al., 2014b; Freitag et al., 2021a), and moving away from Direct Assessments (DA Graham et al., 2013), which rely on (often non-expert) annotators assigning a scalar value to a candidate translation given a reference. Annotators who follow the MQM guidelines are instructed

[1]The acronym MaTESe stands for "machine translation evaluation as a sequence tagging problem"

to tag the spans of the candidate translations that contain errors.[2] Each identified span is labeled with an error category (e.g., `Fluency/Grammar`, `Fluency/Punctuation`, or `Style/Awkward`) and a severity level (e.g., `Major` or `Minor`), which together determine the score penalty associated with that error span. Finally, a scalar quality score for the entire sentence is derived from the annotated spans.

Inspired by the MQM evaluation framework, our work aims to employ a similar paradigm for automatic evaluation. We propose the MaTESe metrics, which, given a candidate translation and its reference, identify spans of errors in the translation, assigning to them a severity level between `Major` and `Minor`. We also develop a reference-less version of MaTESe, i.e., MaTESe-QE, which predicts error spans using the source text in the place of the reference translation.

To derive error spans, we use the BIO scheme (Ramshaw and Marcus, 1995), assigning to each token a label in {`O`, `B-Minor`, `I-Minor`, `B-Major`, `I-Major`}. Then, we derive a score for the entire tagged sentence by weighting the identified error spans similarly to Freitag et al. (2021a): each error span is assigned a score of -5 or -1 for `Major` and `Minor`, respectively. Following Freitag et al. (2021a), the score assigned to a translation is the sum of the scores assigned to its error spans, with a minimum total score of -25. System-level scores are computed by averaging the scores of the translations produced by the same MT system. Figure 3.1 shows an example of the annotations returned by our metrics.

3.1.1 Formal definition

In our experiments, we investigate the effectiveness of different configurations for the MaTESe metrics. Here, we formally define the best-performing one, namely, the configuration that uses XLM-RoBERTa (XLM-R, Conneau et al., 2020) large encoder as the backbone pre-trained language model, and leverages reference translations for the evaluation.

Let us consider a source text $\boldsymbol{x}$, its translation $\boldsymbol{h}$, and a manually curated reference $\boldsymbol{y}$. MaTESe concatenates $\boldsymbol{h}$ and $\boldsymbol{y}$, separated by a `</s>` token, and passes the resulting sequence to XLM-R to derive embedding representations of each token. These representations

[2] In some cases, the source sentence may also be annotated. For instance, omission errors require annotators to report the spans of the source sentence that are missing from the candidate translation.

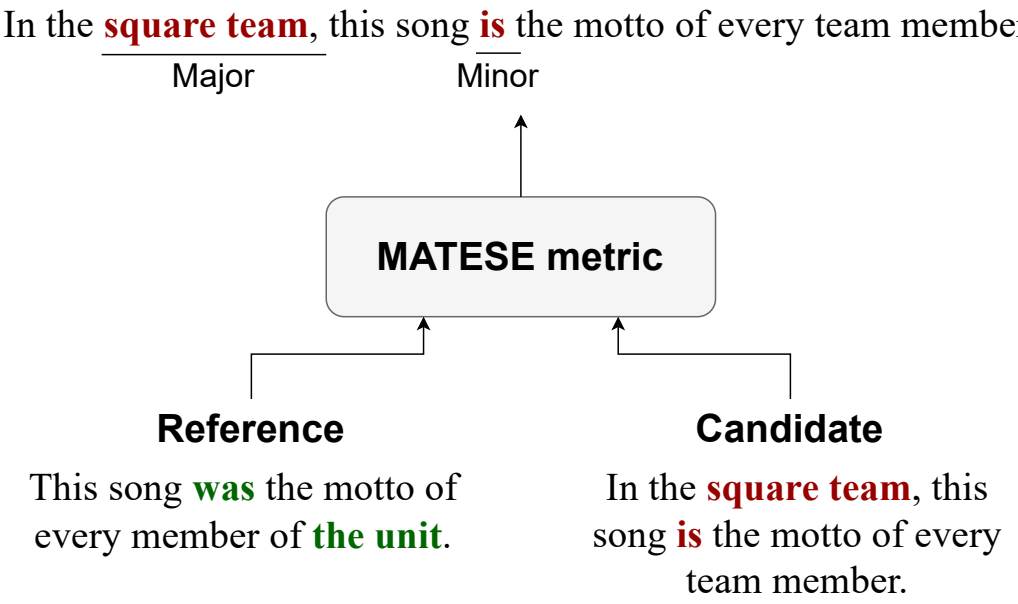


Figure 3.1. Example of the annotation returned by the MaTESe metrics, given a candidate translation and its reference. The final score of the translation is -6, that is the sum of -5 and -1, assigned to the `Major` and `Minor` errors, respectively.

are passed to two additional encoder layers before being sent to a classification head that maps them into the five classes defined above. Formally, let the source, hypothesis, and reference be token sequences

$$\boldsymbol{x} = (x_1, \ldots, x_{T_x}), \quad \boldsymbol{h} = (h_1, \ldots, h_{T_h}), \quad \boldsymbol{y} = (y_1, \ldots, y_{T_y}).$$

The input to the model is the concatenation of the hypothesis and reference separated by a `</s>` token:

$$\boldsymbol{s} = (h_1, \ldots, h_{T_h}, \texttt{</s>}, y_1, \ldots, y_{T_y}),$$

with total length $T = T_h + T_y + 1$.

Let XLMR_ϕ denote the pre-trained XLM-R large encoder with L layers and parameters ϕ. For each layer $\ell \in \{1, \ldots, L\}$, let

$$\mathbf{E}^{(\ell)} = \mathrm{XLMR}_\phi^{(\ell)}(\boldsymbol{s}) \in \mathbb{R}^{T \times d},$$

where d is the hidden size and the t-th row $\mathbf{e}_t^{(\ell)} \in \mathbb{R}^d$ corresponds to the representation of the t-th token at layer ℓ.

We obtain the XLM-R output used by MaTESe by summing the representations of the last four layers:

$$\mathbf{E} = \sum_{\ell=L-3}^{L} \mathbf{E}^{(\ell)} \in \mathbb{R}^{T\times d}.$$

The output $\mathbf{E}$ is then passed through two additional transformer encoder layers $\mathrm{Enc}_{\theta_1}^{(1)}$ and $\mathrm{Enc}_{\theta_2}^{(2)}$:

$$\mathbf{R}^{(0)} = \mathbf{E}, \qquad \mathbf{R}^{(\ell)} = \mathrm{Enc}_{\theta_\ell}^{(\ell)}(\mathbf{R}^{(\ell-1)}), \quad \ell = 1, 2,$$

yielding

$$\mathbf{R} = \mathbf{R}^{(2)} \in \mathbb{R}^{T\times d},$$

where the t-th row $\mathbf{r}_t \in \mathbb{R}^d$ is the contextual representation of the t-th token after the two additional encoder layers. Each encoder layer is a standard transformer encoder block with 8 self-attention heads and GELU activations (Hendrycks and Gimpel, 2017).

Let the output classes be

$$\mathcal{C} = \{\texttt{O}, \texttt{B-Minor}, \texttt{I-Minor}, \texttt{B-Major}, \texttt{I-Major}\}.$$

with $K = |\mathcal{C}|$. A token-level classification head is applied independently at each position $t \in \{1, \ldots, T_h\}$ corresponding to the tokens of the hypothesis $\boldsymbol{h}$. Before the classifier, we apply a dropout module D_ρ with dropout rate $\rho = 0.1$:

$$\tilde{\mathbf{r}}_t = D_\rho(\mathbf{r}_t).$$

The classification head is a single feed-forward layer with parameters $W \in \mathbb{R}^{K\times d}$ and no bias:

$$\mathbf{o}_t = W\tilde{\mathbf{r}}_t \in \mathbb{R}^K, \quad t = 1, \ldots, T_h.$$

The corresponding class probabilities are

$$p(c \mid \boldsymbol{h}, \boldsymbol{y}, t) = \mathrm{softmax}(\mathbf{o}_t)_c = \frac{\exp(o_{t,c})}{\sum_{c'\in\mathcal{C}} \exp(o_{t,c'})}, \quad \forall c \in \mathcal{C},\ t = 1, \ldots, T_h.$$

Each token of the hypothesis is assigned the class with the highest probability:

$$\hat{c}_t = \arg\max_{c\in\mathcal{C}} p(c \mid \boldsymbol{h}, \boldsymbol{y}, t), \quad t = 1, \ldots, T_h.$$

Thus, given a pair $(\boldsymbol{h}, \boldsymbol{y})$, MaTESe outputs the sequence of predicted classes for the hypothesis tokens,

$$\mathrm{MaTESe}(\boldsymbol{h}, \boldsymbol{y}) = (\hat{c}_1, \ldots, \hat{c}_{T_h}),$$

assigning one class label for each token in the candidate translation.

After obtaining the sequence of class labels, we construct translation error spans according to the BIO scheme: tokens labeled with `O` are treated as correct, labels `B-Minor` and `B-Major` mark the beginning of `Minor` and `Major` error spans, respectively, and labels `I-Minor` and `I-Major` mark their continuations. In this way, both the boundaries and the severity of each error span are determined using a single set of labels.

3.1.2 Data pre-processing

The MaTESe metrics are trained and evaluated using the MQM data released by Freitag et al. (2021a) and the test sets of the WMT 2021 Metrics Shared Task (Freitag et al., 2021b). However, to facilitate training, we apply several transformations to the original data:

1. We discard annotations belonging to the `Non-translation` category (Table 2.2). Errors from this category are weighted -25 by Freitag et al. (2021a), which would require special handling. However, they might be too scarce ($< 0.1\%$ of the whole data) for our model to learn how to assign them;

2. We discard annotations referring to `Accuracy/Omission` and `Source error` categories. In both cases, the source sentence should be annotated, but our models are trained to tag only the candidate translation;

3. We discard annotations of errors with `Neutral` severity. These annotations are subjective and do not contribute to the final quality score (Freitag et al., 2021a);

4. We replace `Critical` severity labels with `Major` to make the English→Russian dataset consistent with the rest of the data;

5. We discard all error categories. While we acknowledge the importance of error categorization, we chose to remove it because it makes the classification problem sparser, which could be problematic given the limited amount of training data.

Furthermore, since in the MQM data released by Freitag et al. (2021a) every sentence was annotated three times by different raters, we create a single training sample per sentence by merging different raters' annotations. Specifically, we retain all non-overlapping errors

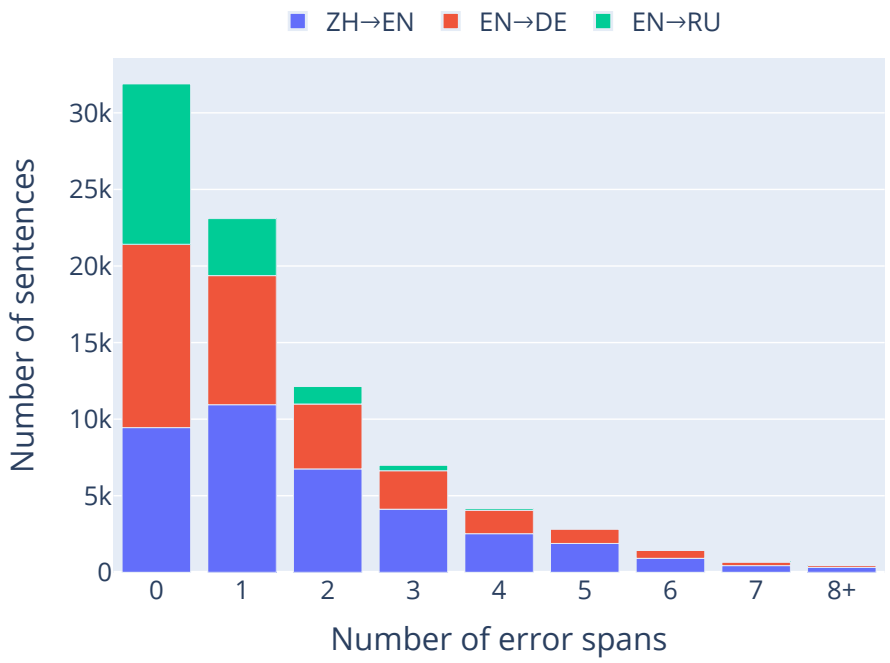


Figure 3.2. Distribution of the number of error spans in our data, after the pre-processing illustrated in Section 3.1.2.

and, in case where two error spans partial overlap, we discard the one associated with the `Minor` error in favor of the `Major` one, or select one at random if both have the same severity. We prioritize `Major` over `Minor` errors because Freitag et al. (2021a) observed that MT system rankings remain nearly identical when considering only `Major` errors compared to using the full MQM score. Figure 3.2 reports the distribution of error spans across our dataset.

3.2 Experimental Setup

In this Section, we describe the different architectures we experiment with, the training and evaluation data, and the meta-metrics we use to measure performance.

3.2.1 Architectures

In order to have a single model capable of evaluating text in multiple languages, we leverage multilingual pre-trained models like XLM-RoBERTa (Conneau et al., 2020) and mBART

(Liu et al., 2020b). Nonetheless, we compare the performances of multilingual models with their English-only counterparts by also experimenting with RoBERTa (Liu et al., 2019). Specifically, RoBERTa can be employed only with language pairs that have English as the target language – in our case, this is only Chinese→English – and only for reference-based evaluation.

Encoder-only models XLM-RoBERTa and RoBERTa consist of only the encoder part of the standard Transformer architecture (Vaswani et al., 2017). As input for these models, we concatenate the candidate translation and its reference (or source, for MaTESe-QE), separated by a `</s>` token. Furthermore, we add two randomly initialized encoder layers on top of the last layer and a classification head. Due to computational constraints, we keep the embedding layer frozen during training.

Encoder-decoder model When using mBART, we feed the reference translation (or the source, for MaTESe-QE) to the encoder, and the candidate to the decoder. This way, we maintain similarity with mBART pre-training. We highlight that we do not use the decoder autoregressively; instead, following standard practices for sequence classification with encoder-decoder models, we force the candidate to be processed all at once, and collect the contextualized embeddings at the last layer. On top of the decoder, we add two randomly initialized encoder layers and a classification head. As with the encoder-only models, we keep the embedding layer frozen.

3.2.2 Training and validation data

As mentioned, our data consists of the MQM data released by Freitag et al. (2021a), concatenated to the test sets of the WMT 2021 Metrics Shared Task (Freitag et al., 2021b). This data contains translations in three language pairs: Chinese→English, English→German, and English→Russian. However, we experiment only with the first two, since the English→Russian language direction is included only in the test sets of WMT21, but not in the data released by Freitag et al. (2021a).

We train our models on 90% of all the data, and evaluate them on the remaining 10%. We wish to highlight that this split is performed on *unique source sentences*. Indeed, since

Figure 3.3. An example of evaluation with the Hypothesis and Target Span Hit metrics. The **turquoise line —** (below) and **amber line —** (above) represent the hypothesis and target annotation, respectively. HSH $= 2/3$ (2 out of 3 spans are *hit*), TSH $= 2/2$ (2 out of 2 spans are *hit*).

each source sentence was translated by multiple MT systems, this split prevents translations of the same source sentence from being present both in the training and the validation splits.

Furthermore, to allow a fair comparison between the MaTESe metrics and those submitted to the WMT21 Metrics Shared Task, we also retrained them using solely the data released by Freitag et al. (2021a). We dub these metrics MaTESe^{21} and MaTESe-QE^{21}.

3.2.3 Evaluation metrics

Correlations with MQM scores paint only a partial picture, since the final score assigned to a translation depends only on the number of error spans (with their severity), but not on their position in the sentence. For example, if a system flagged a span as a `Major` error, but the target annotation had a different span tagged as `Major`, the MQM scores would be identical despite the tagging error. However, differently from other neural metrics, the MaTESe metrics can be evaluated both according to the final scores and in terms of the produced token-level annotations. Thus, we use correlation with human judgments to measure metric performance at the score level, and Precision, Recall, and F-score, to measure our metrics' ability at classifying each token into one of the following classes: {`O`, `B-Minor`, `I-Minor`, `B-Major`, `I-Major`}.[3] Given that our data is highly imbalanced (see Table 3.1), we employ macro versions of these metrics and, in particular, use macro F-score to select the best checkpoint of our models on the validation set. Finally, we also assess the span-level error detection capabilities of our systems. To do this, we introduce two new span-level evaluation metrics, which we call the Hypothesis and Target Span Hit metrics.

[3]We compute Precision, Recall, and F1 using the TorchMetrics[4] modules.

Table 3.1. Distribution of the token-level gold annotations in the concatenation of the training and test sets of WMT 2021 Metrics Shared Task, after the pre-processing we described in Section 3.1.2.

	O	B-Minor	I-Minor	B-Major	I-Major
EN→DE	818 945	32 667	37 897	8516	25 192
ZH→EN	1 053 663	33 633	48 333	33 996	76 984
EN→RU	343 449	614	1015	7271	3189
ALL	2 216 057	66 914	87 245	49 783	105 365

The Span Hit Metrics

Besides the token-level Precision, Recall, and F-score measures, we are also interested in determining how well our metrics are able to flag, even partially, a true error span, regardless of its severity or length. Unfortunately, the existing span-level metrics, such as Span Precision, Span Recall and Span F1, focus on exact matches between the predicted and target spans, which we believe to be too strict. To address this issue, we introduce the Hypothesis Span Hit (HSH) and Target Span Hit (TSH) metrics: HSH represents the percentage of predicted error spans that are also, at least partially, true; instead, TSH represents the percentage of true error spans that the metric has predicted, even partially. An example of using HSH and TSH is given in Figure 3.3.

Formal definition Let us consider a candidate translation $\boldsymbol{c}$ as a sequence of tokens $(c_1, c_2, \ldots, c_n)$; moreover, let us define an error span s as a set of contiguous tokens in $\boldsymbol{c}$, e.g., $\{c_1, c_2, c_3\}$, and an error annotation A as a set of disjoint error spans, i.e., that satisfies $\bigcap_{s' \in A} s' = \emptyset$. Furthermore, we define the Span Hit Indicator as

$$\mathrm{SHI}(s, A) = \mathbb{I}(s \cap \sigma(A) \neq \emptyset)$$

where $\sigma(A) = \bigcup_{s' \in A} s'$, i.e., the set of all tokens in annotation A. In simpler terms, $\mathrm{SHI}(s, A)$ is 1 if at least one of the tokens in s belongs to the set of all tokens of the error spans in A.

Finally, let us take two error annotations: A_h represents the hypothesis spans produced by a model, while A_t represents the target spans that $\boldsymbol{c}$ was originally annotated with. We

Table 3.2. Comparison of different architectures in terms of Precision, Recall and F-score in their macro versions; HSH and TSH are Hypothesis Span Hit and Target Span Hit metrics.

	P	R	F1	HSH	TSH
XLM-R$^{\text{LARGE}}$	47.38	**38.40**	**41.72**	57.73	**46.08**
XLM-R$^{\text{BASE}}$	46.64	34.12	37.93	**58.01**	38.70
mBART	**47.97**	31.94	36.01	55.85	32.66

define the Hypothesis Span Hit and Target Span Hit metrics as follows:

$$\text{HSH}(A_h, A_t) = \frac{\sum_{s_h \in A_h} \text{SHI}(s_h, A_t)}{|A_h|}$$

$$\text{TSH}(A_t, A_h) = \frac{\sum_{s_t \in A_t} \text{SHI}(s_t, A_h)}{|A_t|}$$

Both metrics are defined as the average number of span hits of one error annotation with respect to the other. To compute the metrics for an entire dataset we employ micro-averaging, i.e., we concatenate all hypotheses into a single one, do the same for the targets, and then measure Span Hit metrics on the newly-created pair of hypothesis and target. We avoid averaging the single results because the number of spans varies widely across samples (Figure 3.2).

3.3 Results

In this section, we show the results obtained by our metrics. Unless explicitly stated otherwise, all experiments were conducted using the reference-based version of the MaTESe metrics.

3.3.1 Architectures comparison

The results of the architecture comparison are shown in Table 3.2. The best performing architecture is XLM-R$^{\text{LARGE}}$, which attains the highest F-score. Given the complexity of the task and the data imbalance, we conjecture that the other architectures obtain higher Precision because they tend to detect only the most evident errors, assigning `O`s more frequently. This is supported by the TSH score, which excludes `O` labels and thus amplifies

Table 3.3. Model performance on monolingual and multilingual settings. XLM-R$^{\text{ALL}}$ is trained and evaluated on the concatenation of English→German (EN→DE) and Chinese→English (ZH→EN) datasets, whereas XLM-R$^{\text{MONO}}$ refers to two separate models, each trained and evaluated on a single dataset. RoBERTa is an English language model and can therefore be applied to the ZH→EN data only.

	EN→DE					ZH→EN				
	P	R	F1	HSH	TSH	P	R	F1	HSH	TSH
XLM-R$^{\text{ALL}}$	43.03	**33.09**	**35.89**	54.35	**44.54**	47.86	**39.39**	42.63	59.90	**47.04**
XLM-R$^{\text{MONO}}$	**43.98**	30.64	33.52	**56.56**	39.67	50.85	38.51	**42.77**	63.51	44.38
RoBERTa	–	–	–	–	–	**51.49**	37.91	42.41	**64.33**	42.81

the differences between architectures, with XLM-R$^{\text{BASE}}$ and mBART achieving substantially lower scores than XLM-R$^{\text{LARGE}}$. Interestingly, XLM-R outperforms mBART despite having less than half of its parameters. We hypothesize that this advantage might stem from XLM-R's self-attention mechanism, which encodes both sequences (the translation and the reference sentence) simultaneously, potentially yielding token representations that better integrate information from the two inputs.

3.3.2 Monolingual-multilingual comparison

Table 3.3 reports the results obtained when training the XLM-R model on a single language pair at a time or on both. This setup also allows us to investigate whether an English model such as RoBERTa can outperform XLM-R when dealing exclusively with English data.

Our results show that training on the full dataset is beneficial: XLM-R$^{\text{ALL}}$ achieves an F-score that is higher than or comparable to its monolingual variants. Furthermore, XLM-R$^{\text{ALL}}$ attains higher Recall and Target Span Hit for both language pairs. Similar to what we observed with different architectures, we hypothesize that training on a larger and more diverse dataset enables the models to detect a broader range of errors, even if the additional data comes from a different language. Finally, we do not observe sizable differences between the results obtained by RoBERTa and those of XLM-R$^{\text{MONO}}$ on the Chinese→English data.

Table 3.4. Comparison of our reference-based and reference-free models, i.e., MaTESe and MaTESe-QE, respectively.

	P	R	F1	HSH	TSH
MaTESe	47.38	**38.40**	**41.72**	57.73	**46.08**
MaTESe-QE	**49.34**	34.53	38.89	**59.89**	36.84

3.3.3 MaTESe-QE

A desirable feature of evaluation metrics is to function both in the presence and absence of human-curated references. To achieve this, we investigate whether it is feasible to tag the errors in the candidate translation by looking at the source sentence rather than the reference. Table 3.4 reports the results obtained by the best architecture, i.e., XLM-R$^{\text{LARGE}}$, trained on both English→German and Chinese→English, when using the reference sentence and when operating without it.

MaTESe outperforms MaTESe-QE in terms of Recall, F-score, and Target Span Hit metrics. Unsurprisingly, the information provided by the reference is easier to exploit, allowing the reference-based model to detect a much wider range of errors. At the same time, MaTESe-QE proves to be a viable alternative in the absence of manually-curated references: it achieves higher levels of Precision and Hypothesis Span Hit, indicating that its predictions are more accurate than those of MaTESe, albeit over a narrower range of detectable errors.

3.3.4 Correlations with Human Judgments

Tables 3.5 and 3.6 report the correlations with human judgments that our metrics attained on `newstest2021` (in-domain) and `TED` (out-of-domain) test sets of the WMT 2021 Metrics Shared Task: `w/ HT` indicates that manually-curated references have been scored together with system outputs, while `w/o HT` indicates that they have been kept out of the evaluation. Aside from our models, i.e., MaTESe [21] and MaTESe-QE [21], we also report two additional baselines: #1 WMT and #2 WMT. These are the top-1 and top-2 results reported by Freitag et al. (2021b) in the corresponding tables (Tables 23, 24, 27, and 28). Since those positions are held by different systems, we assign each submission a unique symbol and report the mapping as follows:

Table 3.5. System-level correlations with human judgments as measured in the WMT 2021 Metrics Shared Task (Freitag et al., 2021b). MaTESe [21] and MaTESe-QE [21] are MaTESe metrics that have been re-trained using only the training set of such a Shared Task. A legend of the other symbols is found in Section 3.3.4.

	EN→DE			ZH→EN		
	w/o HT	w/ HT	TED	w/o HT	w/ HT	TED
#1 WMT	‡0.938	⊥0.823	‖**0.818**	∧**0.834**	∧**0.727**	∨**0.421**
#2 WMT	†0.937	⊥0.822	⊥0.802	‖0.628	‖0.619	⊤0.403
MaTESe[21]	**0.946**	**0.863**	0.621	0.636	0.701	0.017
MaTESe-QE[21]	0.910	0.806	0.584	0.502	0.600	0.056

- ‡: cushLEPOR(LM) (Han et al., 2021);
- ⊥: C-SPEC and C-SPECpn (Takahashi et al., 2021);
- ∧: tgt-regEMT and tgt-regEMT-baseline (Stefanik et al., 2021);
- ‖: COMET-MQM_2021 and COMET-QE-MQM_2021-src (Rei et al., 2021);
- ∨: TER (Snover et al., 2006);
- †: BLEU (Papineni et al., 2002);
- ⊤: MTEQA (Krubiński et al., 2021a,b).

In in-domain settings,[5] MaTESe[21] and MaTESe-QE[21] achieve correlations on par or better than the top-2 WMT 2021 submissions in English→German, while they are sub-optimal in Chinese→English. Interestingly, in out-of-domain settings (i.e., on the TED split of the test set), we observe a sizable drop in correlation on both translation directions. We attribute this drop to the very limited amount of training data, which probably hinders proper generalization capabilities. Finally, we observe that MaTESe-QE[21] lags behind MaTESe[21] in all settings.

[5]The TED split of the WMT 2021 test set is out-of-domain for the MaTESe metrics, which, for this experiment, were trained only on the data released by Freitag et al. (2021a).

Table 3.6. Segment-level correlations with human judgments as measured in the WMT 2021 Metrics Shared Task (Freitag et al., 2021b). MaTESe[21] and MaTESe-QE[21] are MaTESe metrics that have been re-trained using only the training set of such a Shared Task. A legend of the other symbols is found in Section 3.3.4.

	EN→DE			ZH→EN		
	w/o HT	w/ HT	TED	w/o HT	w/ HT	TED
#1 WMT	$^{\perp}0.267$	$^{\perp}0.256$	$^{\wedge}\mathbf{0.290}$	$^{\perp}\mathbf{0.402}$	$^{\perp}\mathbf{0.390}$	$^{\wedge}0.248$
#2 WMT	$^{\perp}0.266$	$^{\perp}0.254$	$^{\perp}0.285$	$^{\perp}0.401$	$^{\perp}0.388$	$^{\perp}0.241$
MaTESe[21]	**0.323**	**0.310**	0.271	0.358	0.346	**0.257**
MaTESe-QE[21]	0.288	0.277	0.210	0.343	0.332	0.196

3.3.5 Submissions to WMT

We submitted the MaTESe metrics to the WMT 2022 and WMT 2023 Metrics Shared Tasks (Freitag et al., 2022b, 2023). The WMT22 submission includes MaTESe and MaTESe-QE, as described in this chapter, trained on the MQM annotations released by Freitag et al. (2021a) and at WMT21 (Freitag et al., 2021b). For this submission, we further include the English→Russian data into our training recipe. Moreover, we split the whole data 5 times, each time using 90% for training and 10% for validation, and train 5 different systems (10 if we also consider MaTESe-QE). In our submission, each score is the median prediction of the systems trained on the 5 different data splits.

Instead, the WMT23 submission includes a new version of MaTESe implemented as an ensemble of two models: one based on Info-XLM (Chi et al., 2021), used to evaluate translations into German and Russian, and the other based on DeBERTa-v3-large (He et al., 2023a), used to evaluate translations into English. The training sets for these models also incorporated the MQM annotations released at WMT22 (Freitag et al., 2022b). MaTESe-QE instead uses Info-XLM exclusively.

Extracts from the WMT22 and WMT23 rankings are reported in Table 3.7. In both cases, the MaTESe metrics achieved strong results. In WMT22, MaTESe ranked just below Comet-20 (Rei et al., 2020) and BLEURT-20 (Sellam et al., 2020). In WMT23, MaTESe ranked within the same statistical significance cluster as state-of-the-art metrics such as

COMET (Rei et al., 2022a) and BLEURT-20, and just below much larger models such as GEMBA-MQM (Kocmi and Federmann, 2023a) and MetricX-23 (Juraska et al., 2023).

We also compared the WMT23 version of MaTESe with xCOMET-XL at the span level. Notably, among the primary submission of WMT23, the xCOMET family is the only one fine-tuned to produce error spans (a more detailed description of xCOMET metrics is provided in Section 2.1.2).[6] For this comparison, we rely on the Span Hit metrics (introduced in Section 3.2.3) and evaluate performance on three test sets:

- WMT23 EN→DE (Freitag et al., 2023).
- WMT23 ZH→EN (Freitag et al., 2023).
- WMT24 EN→DE (Freitag et al., 2024).

We selected test sets because they cover translation directions supported by both MaTESe and xCOMET. The results are presented in Table 3.8.

As shown, there is no clear winner. MaTESe appears to be substantially more precise, achieving higher HSH values, whereas xCOMET excels in terms of TSH, indicating that it identifies more ground-truth errors. These findings confirm MaTESe as a state-of-the-art MT metric, particularly when considering the difference in model size between the two approaches: MaTESe is based on Deberta-v3-large for evaluating translations into English and on Info-XLM for translations into other languages, with 304M and 550M parameters, respectively. In contrast, xCOMET-XL is based on XLM-R XL, which features 3.5B parameters (Conneau et al., 2020).

3.4 Conclusions

In this chapter, we addressed the problem of *Score Attribution* (Section 1.1) by introducing the MaTESe metrics, neural metrics that identify error spans in translations while also assigning severity levels to them. Besides the enhanced evaluation interpretability, we showed that the MaTESe metrics display high levels of correlation with human judgments, achieving performance in the same ballpark as the best-performing metrics proposed in the WMT21, WMT22, and WMT23 Metrics Shared Tasks.

[6] Although GEMBA-MQM also produces error spans, it is based on the much larger GPT-4, and its predictions for the WMT23 test set were not released publicly.

Table 3.7. Official ranking of primary submissions of the WMT22 (on the left) and WMT23 (on the right) Metrics Shared Task. 'Avg. rank' (WMT22) is an average over more than 201 different evaluation scenarios. Instead, 'Avg. corr' (WMT23) is averaged over 10 different tasks, and 'Rank' indicates clusters of statistical significance. Starred metrics are reference-free.

(a) WMT22

Metric	**avg rank**
MetricX XXL	1.20
Comet-22	1.32
UniTE	1.86
BLEURT-20	1.91
Comet-20	2.36
MaTESe	2.57
CometKiwi*	2.70
MS-Comet-22	2.84
UniTE-src*	3.03
YiSi-1	3.27
Comet-QE*	3.33
MaTESe-QE *	3.85
MEE4	3.87
BERTScore	4.06
MS-Comet-QE-22*	4.06
chrF	4.70
f101spBLEU	4.97
HWTSC-Teacher-Sim*	5.17
BLEU	5.31
REUSE*	6.69

(b) WMT23

Metric	**Rank**	**Avg. corr**
XCOMET-Ensemble	1	0.825
XCOMET-QE-Ensemble*	2	0.805
MetricX-23	2	0.802
GEMBA-MQM*	2	0.800
MetricX-23-QE*	2	0.800
mbr-metricx-qe*	2	0.788
MaTESe	3	0.782
CometKiwi*	3	0.782
COMET	3	0.779
BLEURT-20	3	0.776
KG-BERTScore*	3	0.774
sescoreX	3	0.772
cometoid22-wmt22*	4	0.772
docWMT22CometDA	4	0.768
docWMT22CometKiwiDA*	4	0.767
Calibri-COMET22	4	0.767
Calibiri-COMET22-QE*	4	0.755
YiSi-1	4	0.754
MS-COMET-QE-22*	5	0.744
...	5	0.744

Table 3.8. Span-level comparison between MaTESe and xCOMET-XL using the Span Hit metrics defined in Section 3.2.3

	WMT23 EN→DE		WMT23 ZH→EN		WMT24 EN→DE	
Metric	**HSH**	**TSH**	**HSH**	**TSH**	**HSH**	**TSH**
MaTESe	**52.86**	34.11	**63.81**	30.36	**37.27**	37.21
MaTESe-QE	51.92	35.28	63.22	24.77	38.69	30.95
xCOMET-XL	34.00	**59.02**	48.48	**61.05**	21.82	**62.64**
xCOMET-XL-QE	30.91	57.11	46.06	60.37	19.75	58.44

Chapter 4

Interpretable Machine Translation Meta-Evaluation

In Section 1.1, we identified *Performance* as another important factor limiting the interpretability of MT evaluation. Specifically, when metrics are assessed solely in terms of their correlation with human judgments, we gain little insight into their capabilities in absolute terms, rather than relative to each other. To address this limitation, this chapter introduces a novel, interpretable MT meta-evaluation framework that measures metric performance in terms of Precision, Recall, F-score, and Re-Ranking Precision.

4.1 Introduction

The WMT2020 and WMT2021 Metrics Shared Tasks (Mathur et al., 2020b; Freitag et al., 2021b) have seen the rise of reference-free metrics, which assess translation quality without the need for human-curated references. More recently, reference-free metrics have demonstrated performance on par with, and sometimes superior to, their reference-based counterparts (Freitag et al., 2023; Kocmi et al., 2024c). Thanks to these advancements and the ability to use metrics without references, several new use cases for MT metrics have emerged. Freitag et al. (2022a), Fernandes et al. (2022), Farinhas et al. (2023), Ramos et al. (2024), and Finkelstein and Freitag (2024) used MT metrics as utility functions for Minimum Bayes Risk (MBR) decoding (Kumar and Byrne, 2004; Eikema and Aziz, 2020,

2022) and Quality Estimation (QE) re-ranking.[1] Ramos et al. (2024), Gulcehre et al. (2023), He et al. (2024), and Xu et al. (2024) used MT metrics as a proxy for human preferences to fine-tune MT systems using Reinforcement Learning (RL)- and Direct Preference Optimization (DPO)-like training objectives. Peter et al. (2023), Alves et al. (2024), and Gulcehre et al. (2023) used reference-free metrics to filter parallel corpora – discarding all translations assigned with a metric score that is below a certain threshold – with the goal of training MT systems using higher quality data. These works leverage MT metrics for applications beyond their traditional use of measuring incremental improvements in the development of MT systems. However, the lack of a dedicated evaluation, paired with the inherent opacity of MT metrics, makes it challenging to determine whether one metric suits a given task better and what the impact of various design choices is. For example, Alves et al. (2024), Peter et al. (2023), and Gulcehre et al. (2023) filter MT datasets using different MT metrics and thresholds, leaving it unclear whether an optimal choice exists. Furthermore, considering the ever-increasing number of metrics available, researchers are often forced to grid-search the best configuration for each new application, as do Fernandes et al. (2022) and Ramos et al. (2024), who explore by grid-search whether certain metrics are better suited than others for MBR decoding, QE re-ranking, and as reward models for RL-based training. Unfortunately, the lack of dedicated evaluation setups often requires revisiting these studies to assess whether their findings hold with newer metrics, resulting in a non-negligible increase in experimentation time.

In this chapter, we address these issues by introducing a novel and more interpretable evaluation framework for MT metrics. Specifically, with regard to the interpretability issue presented in Section 1.1, this chapter focuses on mitigating the *Performance* factor: while useful for ranking metrics, correlations with human judgments alone do not provide users with a clear understanding of metric performance in absolute terms, nor do they offer guidance on which metrics are best suited for specific applications such as data filtering and translation re-ranking. To overcome this limitation, we measure metric capabilities using Precision, Recall, F-score, and Re-Ranking Precision across two evaluation scenarios designed as proxies for these applications. Thus, in addition to ranking metrics – typically done via correlation with human judgments – we quantitatively assess their effectiveness

[1] We refer the reader to Section 2.1.2) for details on QE re-ranking and MBR decoding.

Table 4.1. Number of MT systems, source segments, and the total number of annotations in $\text{WMT23}_{\text{MQM}}$ and $\text{WMT23}_{\text{DA+SQM}}$, excluding official WMT23 references employed by reference-based metrics. Concerning $\text{WMT23}_{\text{DA+SQM}}$, we restrict the annotations to those available in $\text{WMT23}_{\text{MQM}}$ and discard the rest.

	ZH→EN			EN→DE			HE→EN		
	Sys.	**Seg.**	**Total**	**Sys.**	**Seg.**	**Total**	**Sys.**	**Seg.**	**Total**
$\text{WMT23}_{\text{MQM}}$	15	1177	17655	12	460	5520	13	820	10660
$\text{WMT23}_{\text{DA+SQM}}$	15	884	13260	12	460	5520	–	–	–

in data filtering and translation re-ranking, offering insights into their performance in absolute terms. We release our evaluation framework as open-source software at `https://github.com/SapienzaNLP/interpretable-mt-metrics-eval`.

4.2 An Interpretable Evaluation Framework for MT Metrics

First, let us illustrate data filtering and translation re-ranking.

In data filtering, MT metrics separate good-quality from poor-quality translations. After choosing a threshold value, all translations scoring below the threshold are labeled as poor quality and discarded. To evaluate metrics in data filtering, we assess their performance as binary classifiers tasked to separate good-quality (GOOD) from poor-quality (BAD) translations.

Instead, translation re-ranking involves selecting the best translation from a pool of candidates for the same source text (Section 2.1.2). In this scenario, we measure each metric's ability to identify the best translation.

4.2.1 Metrics as Binary Classifiers for Data Filtering

Let us consider the MT metric $\mathcal{M}$, which outputs scores in the range $[m_1, m_2]$. Let us define $\mathcal{M}(t) \in [m_1, m_2]$ as the score assigned by metric $\mathcal{M}$ to translation t. By selecting an arbitrary threshold value $\tau \in [m_1, m_2]$, we repurpose $\mathcal{M}$ as a binary classifier: a translation t is deemed as GOOD by metric $\mathcal{M}$, with threshold τ, if $\mathcal{M}(t) \geq \tau$, otherwise it is deemed as BAD .

Precision, Recall, and F-score Assuming that we have an oracle $\mathcal{H}$ telling us whether a translation is GOOD or BAD, we can measure the performance of metric $\mathcal{M}$, with threshold τ, in terms of standard measures such as Precision, Recall, and F-score, which we refer to as $P^{\mathcal{M}_\tau}$, $R^{\mathcal{M}_\tau}$, and $F^{\mathcal{M}_\tau}$. Given metric $\mathcal{M}$, oracle $\mathcal{H}$, translation t, and threshold τ, $P^{\mathcal{M}_\tau}$ estimates the probability that translation t is GOOD, given that metric $\mathcal{M}$ deems it as such:

$$P^{\mathcal{M}_\tau} = \hat{\Pr}(\mathcal{H}(t) = \text{GOOD} \mid \mathcal{M}(t) \geq \tau). \tag{4.1}$$

Similarly, $R^{\mathcal{M}_\tau}$ estimates the probability that metric $\mathcal{M}$ deems translation t as GOOD, given that the oracle deems it as such:

$$R^{\mathcal{M}_\tau} = \hat{\Pr}(\mathcal{M}(t) \geq \tau \mid \mathcal{H}(t) = \text{GOOD}). \tag{4.2}$$

Finally, we aggregate Precision and Recall using F_β-score, with $\beta = \frac{1}{\sqrt{2}}$, which weights Precision higher than Recall compared to the more common F_1-score. Arguably, false positives – i.e., translations of low quality that are mistakenly considered GOOD – could be detrimental to the applications that see metrics employed as binary classifiers. For example, in data filtering, false positives correspond to low-quality translations that survive the filtering, compromising the quality of filtered data. In contrast, false negatives – i.e., translations of high quality that are mistakenly assigned with the BAD label – would more frequently lead to minor inconveniences, as they correspond to good-quality translations that are mistakenly discarded. Moreover, we note that MT metrics struggle to achieve high Precision, meaning that metrics' differences can be best highlighted if Precision is weighted higher than Recall.

Therefore, the F-score of metric $\mathcal{M}$, with threshold τ, is defined as follows:

$$F^{\mathcal{M}_\tau} = \frac{3}{2} \frac{P^{\mathcal{M}_\tau} R^{\mathcal{M}_\tau}}{\frac{1}{2} P^{\mathcal{M}_\tau} + R^{\mathcal{M}_\tau}}. \tag{4.3}$$

4.2.2 Metrics as Utility Functions for Translation Re-Ranking

Let us consider the set $T = \{t_1, t_2, ..., t_n\}$ containing translations of the same source text. We are interested in assessing metric performance in ranking the best translation (as determined by human annotators) in the first position. However, metrics and humans might return tied assessments, placing two or more translations together in the first position. Therefore, we define $T^{\mathcal{M}}$ as the subset of T containing all translations assigned with the

Table 4.2. Average and Median MQM scores of the translations in $\text{WMT23}_{\text{MQM}}$ and $\text{WMT22}_{\text{MQM}}$.

	ZH→EN		EN→DE		HE→EN	
	Avg.	**Median**	**Avg.**	**Median**	**Avg.**	**Median**
$\text{WMT23}_{\text{MQM}}$	-4.21	-2.00	-7.47	-3.00	-2.35	0.00
$\text{WMT22}_{\text{MQM}}$	-3.21	-1.00	-1.31	0.00	–	–

highest score by $\mathcal{M}$. Similarly, $T^{\mathcal{H}}$ contains the translations of T ranked highest by human annotators. The Re-Ranking Precision of metric $\mathcal{M}$ is defined as follows:

$$RRP^{\mathcal{M}} = \frac{|T^{\mathcal{M}} \cap T^{\mathcal{H}}|}{|T^{\mathcal{M}}|}. \tag{4.4}$$

Unlike in the data filtering scenario, we focus solely on Re-Ranking Precision, not Recall. This is because, to serve as a proxy for translation re-ranking applications, what matters is whether the returned translation is the best – or among the best – rather than identifying all the translations ranked highest by human annotators.

4.3 Experimental Setup

This section outlines the data employed, the metrics evaluated, and our methodology.

4.3.1 The Data

We employ **$\text{WMT23}_{\text{MQM}}$** (Freitag et al., 2023), which contains human annotations collected within the Multidimensional Quality Metrics framework (MQM, Lommel et al., 2014a), and **$\text{WMT23}_{\text{DA+SQM}}$** (Kocmi et al., 2023), which includes human annotations as Direct Assessments + Scalar Quality Metrics (DA+SQM, Kocmi et al., 2022). Both datasets consist of source texts translated by multiple MT systems, with translation quality assessed by professional human annotators. Additional information regarding these datasets can be found in Table 4.1.

We conduct the evaluation using $\text{WMT23}_{\text{MQM}}$ in the ZH→EN, EN→DE and HE→EN language directions, and use $\text{WMT23}_{\text{DA+SQM}}$ as a reference for human performance, given that it contains a subset of the translations in $\text{WMT23}_{\text{MQM}}$ annotated using a different human evaluation scheme.

4.3.2 The Metrics

We consider the following metrics (most of these metrics have been introduced in Chapter 2):

- COMET, COMET-QE, and COMET-QE-MQM (Rei et al., 2020, 2021).[2]
- BLEURT-20 (Sellam et al., 2020; Pu et al., 2021).[3]
- BERTscore (Zhang et al., 2020).[4]
- MetricX-23, MetricX-23-QE, MetricX-23-XL and MetricX-23-QE-XL from the MetricX family (Juraska et al., 2023). The first two use mT5-XXL and the latter two use mT5-XL (Xue et al., 2021).[5]
- CometKiwi and CometKiwi-XL (Rei et al., 2022b, 2023).[6]
- XCOMET-XL, XCOMET-ENSEMBLE, and XCOMET-QE-ENSEMBLE (Guerreiro et al., 2024).[7]
- MaTESe and MaTESe-QE. Specifically, we use the versions of the MaTESe metrics submitted to the WMT23 Metrics Shared Task, i.e., those built upon InfoXLM and DeBERTaV3 (He et al., 2023b).
- GEMBA-MQM (Kocmi and Federmann, 2023a).[8]
- MBR-MetricX-QE, which is a metric based on the MBR decoding strategy (Naskar et al., 2023). Specifically, given a translation, it uses an MT system to generate pseudo-references and MetricX-23 as the MBR utility function. The final score assigned to the given translation is its MBR score.
- BLEU (Papineni et al., 2002).[9]
- chrF (Popović, 2015).[10]

[2] https://github.com/Unbabel/COMET. We used version 2.2.1 of the COMET framework.
[3] https://github.com/google-research/bleurt.
[4] https://github.com/Tiiiger/bert_score.
[5] https://github.com/google-research/metricx.
[6] See footnote 2.
[7] See footnote 2.
[8] https://github.com/MicrosoftTranslator/GEMBA.
[9] https://github.com/mjpost/sacrebleu.
[10] See footnote 9.

Additionally, following Freitag et al. (2023), we include the results from a random baseline, i.e., Random-sysname, which outputs discrete scores drawn from several Gaussian distributions, one for each MT system that translated the texts in the test set. Each Gaussian has a randomly assigned mean between 0 and 9, with a standard deviation of 2.

4.3.3 Selecting the Thresholds τ

In the data filtering scenario, we measure two different aspects of metric performance, depending on how we select the τ value:

1. By selecting τ to maximize the F-score **on the test set**, we measure MT metrics' ability to separate GOOD from BAD translations under ideal conditions. This scenario allows us to measure the maximum achievable F-score for each metric on the test data, effectively evaluating the metric's discriminative power. Metrics whose assessments are not accurate enough, noisy, or, more generally, poorly aligned with human judgments, will achieve a lower optimal F-score than the others.

2. By selecting τ to maximize the F-score **on a development set** and measuring performance on the test set, we estimate the true values of MT metrics' Precision, Recall, and F-score in data filtering applications.

We use **WMT22$_{\text{MQM}}$** as a development set. However, since some of the tested metrics were trained using WMT22$_{\text{MQM}}$ data, we restrict the experiment involving a development set to the metrics that did not use this data at training time.

4.3.4 Extracting Binary Labels from Manually-Annotated Datasets

Within the MQM annotation framework, professional annotators identify span-level translation errors and assign each error a category and severity. The final MQM score is calculated based on these errors using the following weighting scheme (Freitag et al., 2021a):

Error severity	Category	Penalty
Major	Non-translation	-25
	Others	-5
Minor	Punctuation	-0.1
	Others	-1

We map the annotations in $\text{WMT23}_{\text{MQM}}$ and $\text{WMT22}_{\text{MQM}}$ to binary labels by considering translations with a score above a certain threshold as GOOD. Specifically, if a translation is assigned an MQM score h, we label it as GOOD if $h \geq -4$, meaning it contains no Major errors and at most 4 Minor ones (or more if Minor punctuation errors are present). Additionally, we classify translations as PERFECT if they contain at most 1 Minor error, i.e., those with $h \geq -1$.[11] This allows us to investigate metrics' ability to distinguish between PERFECT and OTHER translations.

4.4 Results

In this section, we report the performance obtained by MT metrics when used as binary classifiers to distinguish between GOOD and BAD, as well as PERFECT and OTHER translations, and in terms of their effectiveness in translation re-ranking, i.e., in selecting the best translations among candidates for the same source text.

4.4.1 Binary Classification

First, we discuss the results obtained in the binary classification task.

τ selected on the test set

Tables 4.3, 4.4, and 4.5 show MT metrics' threshold values, Precision, Recall, and F-score in distinguishing GOOD from BAD and PERFECT from OTHER translations, with the optimal threshold τ selected on the test set ($\text{WMT23}_{\text{MQM}}$).

[11] We use $h \geq -1$ instead of $h = 0$ because inter-annotator agreement in MT evaluation is not particularly high (Freitag et al., 2021a), even under high-cost annotation frameworks such as MQM. Therefore, we argue that selecting only translations with a score of 0 might excessively restrict the set of PERFECT translations while overly depending on individual annotators' preferences.

Table 4.3. Metrics' Precision, Recall, and F-score in binary classification, distinguishing GOOD from BAD, and PERFECT from OTHER translations. τ is selected to maximize the F-score **on the test set**. In the last two columns, we report metrics' Precision in translation re-ranking and the average MQM score of the selected translations. The test set is $\text{WMT23}_{\text{MQM}}$ and the translation direction is ZH→EN. The metrics highlighted in grey are not openly available.

		GOOD vs BAD				PERFECT vs OTHER				Re-ranking	
	Metric	τ	**P**	**R**	F	τ	**P**	**R**	F	**RRP**	**Avg.**
REFERENCE BASED	XCOMET-ENSEMBLE	0.83	79.91	84.42	81.36	0.91	68.25	68.93	68.47	43.17	−2.38
	XCOMET-XL	0.80	78.33	83.63	80.02	0.92	67.55	67.46	67.52	37.49	−2.75
	MetricX-23	−4.79	77.43	86.23	80.15	−2.25	63.99	73.20	66.79	39.63	−2.72
	MetricX-23-XL	−3.52	77.80	84.46	79.90	−1.74	65.60	72.54	67.76	39.52	−2.71
	MaTESe	−4.00	76.53	78.10	77.05	−1.00	55.75	79.88	61.99	33.07	−3.18
	COMET	0.76	74.56	78.76	75.91	0.82	61.25	64.38	62.26	34.25	−3.06
	BLEURT-20	0.60	72.76	82.76	75.81	0.67	55.88	69.21	59.71	33.35	−3.07
	BERTscore	0.84	64.33	99.47	72.91	0.92	48.20	69.15	53.62	32.29	−3.20
REFERENCE FREE	XCOMET-QE-ENSEMBLE	0.83	80.40	83.47	81.40	0.92	70.00	63.60	67.73	41.40	−2.47
	MBR-MetricX-QE	0.73	79.00	82.81	80.23	0.80	67.02	65.91	66.64	38.47	−2.40
	MetricX-23-QE	−3.90	76.73	87.70	80.07	−1.31	67.76	67.85	67.79	37.55	−2.59
	MetricX-23-QE-XL	−3.57	77.91	83.36	79.64	−1.64	67.15	70.08	68.10	36.09	−2.83
	GEMBA-MQM	−5.00	82.41	79.99	81.59	−1.00	64.12	74.12	67.14	42.58	−2.30
	MaTESe-QE	−4.00	73.72	85.64	77.30	0.00	55.43	75.05	60.72	30.34	−3.59
	COMET-QE	−0.01	75.35	82.53	77.60	0.05	59.64	68.59	62.35	37.35	−2.66
	COMET-QE-MQM	0.08	75.40	86.33	78.72	0.10	61.63	73.84	65.22	33.52	−3.59
	CometKiwi	0.76	78.62	80.90	79.37	0.80	64.79	66.52	65.35	39.28	−2.61
	CometKiwi-XL	0.64	78.04	79.81	78.62	0.71	64.73	65.51	64.99	38.78	−2.60
LEXICAL BASED	BLEU	0.00	64.06	100.00	72.78	0.00	42.13	100.00	52.20	30.09	−3.50
	chrF	0.00	64.06	100.00	72.78	0.00	42.13	100.00	52.20	31.51	−3.39
	Random-sysname	−5.00	64.06	100.00	72.78	−4.00	42.14	99.99	52.21	29.04	−3.74
	DA+SQM	63.50	67.83	95.95	75.18	74.67	48.30	82.61	56.06	32.99	−3.22

Table Note: Due to its dependence on GPT-4, we include GEMBA-MQM in the group of not openly available metrics. Instead, openly available versions of MetricX-23 and MetricX-23-QE can be found at `https://github.com/google-research/metricx`. However, we could not compute their predictions due to their high parameter count and thus resorted to using their outputs as submitted to WMT23, which were computed using different checkpoints than those currently available. Therefore, we include MetricX-23 and MetricX-23-QE in the group of not openly available metrics as well

As can be seen, no metric achieves an F-score higher than 84.10, obtained by XCOMET-ENSEMBLE in separating GOOD from BAD translations in HE→EN. Further-

Table 4.4. Metrics' Precision, Recall, and F-score in binary classification, distinguishing GOOD from BAD, and PERFECT from OTHER translations. τ is selected to maximize the F-score **on the test set**. In the last two columns, we report metrics' Precision in translation re-ranking and the average MQM score of the selected translations. The test set is $\text{WMT23}_{\text{MQM}}$ and the translation direction is EN→DE. The metrics highlighted in grey are not openly available.

		GOOD vs BAD				PERFECT vs OTHER				Re-ranking	
	Metric	τ	**P**	**R**	F	τ	**P**	**R**	F	**RRP**	**Avg.**
REFERENCE BASED	XCOMET-ENSEMBLE	0.90	80.55	70.52	76.90	0.92	75.90	67.86	73.02	48.79	−3.58
	XCOMET-XL	0.90	78.49	71.17	75.89	0.94	78.17	65.44	73.41	47.31	−3.91
	MetricX-23	−1.71	79.56	67.61	75.14	−1.18	75.21	68.18	72.71	52.61	−3.47
	MetricX-23-XL	−1.55	77.62	73.98	76.37	−1.07	72.59	70.88	72.01	47.81	−3.74
	MaTESe	−1.00	71.87	72.92	72.21	0.00	67.42	60.67	65.01	43.18	−4.56
	COMET	0.83	72.56	70.81	71.97	0.88	81.10	55.66	70.38	48.26	−3.50
	BLEURT-20	0.70	76.49	69.54	74.03	0.74	72.66	63.26	69.23	48.27	−3.64
	BERTscore	0.85	57.66	81.22	63.83	0.92	68.12	40.42	55.45	43.11	−4.55
REFERENCE FREE	XCOMET-QE-ENSEMBLE	0.87	79.99	70.39	76.51	0.91	75.58	66.10	72.13	46.70	−3.90
	MBR-MetricX-QE	0.76	78.82	70.50	75.84	0.80	74.25	66.75	71.57	48.81	−3.78
	MetricX-23-QE	−2.07	75.65	75.84	75.71	−1.09	76.81	64.88	72.37	48.04	−3.58
	MetricX-23-QE-XL	−1.99	75.86	73.18	74.94	−1.27	74.08	68.44	72.10	45.57	−3.96
	GEMBA-MQM	−1.00	79.69	66.77	74.86	0.00	75.07	62.04	70.16	42.52	−4.04
	MaTESe-QE	−2.00	67.48	82.89	71.93	0.00	68.16	63.48	66.52	41.03	−5.14
	COMET-QE	0.04	68.75	73.61	70.30	0.07	65.26	62.99	64.49	45.71	−3.84
	COMET-QE-MQM	0.08	74.18	73.98	74.12	0.09	73.50	65.64	70.68	41.25	−4.82
	CometKiwi	0.82	75.86	68.51	73.24	0.82	64.08	71.37	66.34	41.75	−4.32
	CometKiwi-XL	0.69	73.52	70.71	72.56	0.73	67.44	64.13	66.30	43.67	−4.45
LEXICAL BASED	BLEU	3.29	52.19	99.25	61.99	0.00	39.95	100.00	49.94	42.95	−4.28
	chrF	28.67	52.59	99.47	62.39	73.11	62.32	39.61	52.32	41.43	−4.48
	Random-sysname	−4.00	52.07	99.94	61.96	−4.00	39.96	99.92	49.95	40.56	−5.36
	DA+SQM	77.33	59.60	85.39	66.27	82.67	48.90	77.21	55.71	37.11	−5.01

more, lower performance is observed when differentiating between PERFECT and OTHER translations, with the highest F-score being 82.49, again from XCOMET-ENSEMBLE. We also note that Precision is almost always lower than Recall, despite the optimal threshold τ being selected to maximize F_β-score with $\beta = \frac{1}{\sqrt{2}}$, which gives more weight to Precision over Recall. These results suggest that the metrics may lack the sensitivity required to distinguish between high-quality translations that differ in minor nuances rather than major errors. As a result, the optimization process may be selecting lower thresholds to compensate

Table 4.5. Metrics' Precision, Recall, and F-score in binary classification, distinguishing GOOD from BAD, and PERFECT from OTHER translations. τ is selected to maximize the F-score **on the test set**. In the last two columns, we report metrics' Precision in translation re-ranking and the average MQM score of the selected translations. The test set is $\text{WMT23}_{\text{MQM}}$ and the translation direction is HE→EN. The metrics highlighted in grey are not openly available.

		GOOD vs BAD				PERFECT vs OTHER				Re-ranking	
	Metric	τ	**P**	**R**	F	τ	**P**	**R**	F	**RRP**	**Avg.**
REFERENCE BASED	XCOMET-ENSEMBLE	0.84	83.28	85.81	84.10	0.87	83.42	80.69	82.49	69.21	−0.99
	XCOMET-XL	0.81	82.00	87.14	83.64	0.85	81.24	82.82	81.76	68.31	−0.98
	MetricX-23	−4.26	82.09	86.51	83.51	−3.34	81.78	81.96	81.84	67.20	−1.11
	MetricX-23-XL	−3.44	81.94	87.23	83.63	−3.39	79.11	88.20	81.92	67.17	−1.07
	MaTESe	−4.00	84.49	76.57	81.67	−4.00	81.86	77.99	80.53	61.99	−1.51
	COMET	0.77	78.96	87.93	81.74	0.81	78.95	81.31	79.72	70.01	−1.03
	BLEURT-20	0.67	80.85	83.39	81.68	0.67	77.48	84.02	79.55	68.33	−1.04
	BERTscore	0.92	76.33	88.14	79.90	0.93	73.89	85.97	77.52	69.88	−0.88
REFERENCE FREE	XCOMET-QE-ENSEMBLE	0.82	80.92	86.90	82.82	0.85	80.96	80.60	80.84	66.22	−1.40
	MBR-MetricX-QE	0.74	81.57	86.17	83.05	0.74	78.01	86.65	80.69	68.09	−1.25
	MetricX-23-QE	−1.79	80.27	89.66	83.17	−1.28	79.42	85.60	81.38	63.17	−1.46
	MetricX-23-QE-XL	−3.46	80.32	86.03	82.13	−3.19	78.27	85.08	80.41	63.25	−1.64
	GEMBA-MQM	−7.00	79.70	89.24	82.64	−5.00	79.01	83.83	80.55	65.22	−1.26
	MaTESe-QE	−6.00	74.35	95.01	80.16	−3.00	75.99	81.43	77.72	53.95	−2.28
	COMET-QE	−0.03	75.62	91.71	80.32	−0.00	74.21	86.43	77.88	61.06	−1.70
	COMET-QE-MQM	0.08	76.28	89.22	80.16	0.09	74.16	87.36	78.09	52.57	−2.32
	CometKiwi	0.77	80.14	86.48	82.14	0.80	80.02	79.85	79.96	60.42	−1.55
	CometKiwi-XL	0.60	77.83	89.83	81.46	0.63	76.90	84.57	79.30	65.61	−1.28
LEXICAL BASED	BLEU	0.00	71.31	100.00	78.85	0.00	67.89	100.00	76.03	64.46	−1.38
	chrF	0.00	71.31	100.00	78.85	0.00	67.89	100.00	76.03	65.61	−1.27
	Random-sysname	−5.00	71.32	99.99	78.85	−5.00	67.89	99.99	76.03	53.51	−2.08

for their lack of Precision with a higher Recall.

It is worth noting that the best-performing openly available, reference-free metric is MetricX-23-QE-XL across all three language directions. Therefore, we recommend using MetricX-23-QE-XL for data filtering applications.

Table 4.6. Metrics' Precision, Recall, and F-score in binary classification, distinguish GOOD from BAD, and PERFECT from OTHER translations. τ is selected to maximize the F-score **on the development set**, i.e., $\text{WMT22}_{\text{MQM}}$. The test set is $\text{WMT23}_{\text{MQM}}$ and the translation direction is ZH→EN.

		GOOD vs BAD				PERFECT vs OTHER			
	Metric	τ	**P**	**R**	F	τ	**P**	**R**	F
REFERENCE BASED	MetricX-23-XL	−3.93	76.58	87.01	79.77	−2.97	57.70	88.80	65.32
	COMET	0.77	75.95	75.28	75.72	0.79	55.51	74.57	60.68
	BLEURT-20	0.61	73.81	79.92	75.74	0.64	52.45	76.89	58.67
	BERTscore	0.92	71.44	63.39	68.54	0.93	49.99	59.81	52.88
REFERENCE FREE	MetricX-23-QE-XL	−5.45	73.36	91.88	78.64	−3.54	55.63	90.32	63.80
	COMET-QE-MQM	0.07	72.57	92.41	78.17	0.08	52.59	93.19	61.53
	COMET-QE	−0.02	73.77	85.81	77.39	−0.02	50.54	88.44	58.96
	CometKiwi	0.74	75.57	85.35	78.58	0.76	53.38	86.73	61.23
	CometKiwi-XL	0.62	75.47	83.37	77.93	0.64	53.70	85.03	61.22
LEXICAL BASED	BLEU	5.44	64.78	85.99	70.58	5.43	42.73	86.42	51.39
	chrF	2.08	64.04	99.84	72.73	2.08	42.09	99.77	52.14

τ selected on the development set

Tables 4.6 and 4.7 report threshold values, Precision, Recall, and F-score when the threshold is optimized on the development set ($\text{WMT22}_{\text{MQM}}$).[12] These values are estimates of the true performance of metrics in data filtering applications.

As expected, the F-score values are lower than the optimal ones reported in Tables 4.3 and 4.4. Nonetheless, the metric rankings remain stable across the two settings, with MetricX-23-XL and MetricX-23-QE-XL outperforming the other metrics among the reference-based and reference-free ones, respectively. Finally, the best-performing openly available, reference-free metric is again MetricX-23-QE-XL, confirming itself as our recommended metric for data filtering applications.

[12]The HE→EN translation direction is not included in these results because it was introduced at WMT23, and we did not have any previous dataset to use as a development set. Furthermore, we restrict the set of tested metrics to those that are openly available and do not employ the $\text{WMT22}_{\text{MQM}}$ data for training.

Table 4.7. Metrics' Precision, Recall, and F-score in binary classification, distinguish GOOD from BAD, and PERFECT from OTHER translations. τ is selected to maximize the F-score **on the development set**, i.e., $\text{WMT22}_{\text{MQM}}$. The test set is $\text{WMT23}_{\text{MQM}}$ and the translation direction is EN→DE.

		GOOD vs BAD				PERFECT vs OTHER			
	Metric	τ	**P**	**R**	F	τ	**P**	**R**	F
REFERENCE BASED	MetricX-23-XL	−2.10	73.10	81.34	75.65	−1.10	71.84	71.41	71.70
	COMET	0.72	57.94	94.05	66.44	0.80	53.88	84.32	61.25
	BLEURT-20	0.60	63.64	87.97	70.11	0.66	56.99	82.09	63.45
	BERTscore	0.68	52.15	99.84	62.03	0.68	40.00	99.79	49.99
REFERENCE FREE	MetricX-23-QE-XL	−2.71	70.68	81.67	74.00	−1.66	67.12	76.24	69.91
	COMET-QE-MQM	0.07	72.32	76.27	73.59	0.09	69.07	71.16	69.75
	COMET-QE	−0.05	56.09	94.26	64.84	−0.01	47.75	87.66	56.29
	CometKiwi	0.73	60.98	90.91	68.50	0.79	55.90	83.61	62.84
	CometKiwi-XL	0.52	59.13	91.29	67.00	0.62	51.57	86.45	59.59
LEXICAL BASED	BLEU	0.00	52.05	100.00	61.95	0.00	39.95	100.00	49.94
	chrF	0.00	52.05	100.00	61.95	0.00	39.95	100.00	49.94

Thresholds reliability Our results show that optimal thresholds tend to vary when moving from $\text{WMT23}_{\text{MQM}}$ to $\text{WMT22}_{\text{MQM}}$ for their calculation. For example, MetricX-23-QE-XL's τ shifts from -3.57 to -5.45, and from -1.64 to -3.54, when separating between GOOD and BAD and PERFECT and OTHER, respectively, and other metrics display a similar pattern. Optimal threshold values do not appear stable across language pairs either, as illustrated in Figures 4.1 and 4.2. Specifically, optimal thresholds frequently differ between EN→DE and the other translation directions in the GOOD vs BAD scenario, and are substantially lower for HE→EN in the PERFECT vs OTHER scenario. Such differences in the optimal thresholds might suggest that metric scores have different meanings depending on the translation direction. Furthermore, and as already discussed, an overly small optimal threshold might suggest that metrics are not precise enough. However, threshold values might also be influenced by the quality of the translations in the dataset. Indeed, on average, the HE→EN dataset contains higher-quality translations compared to the other language pairs (Table 4.2). This might incentivize the metrics to "settle" for lower threshold values in

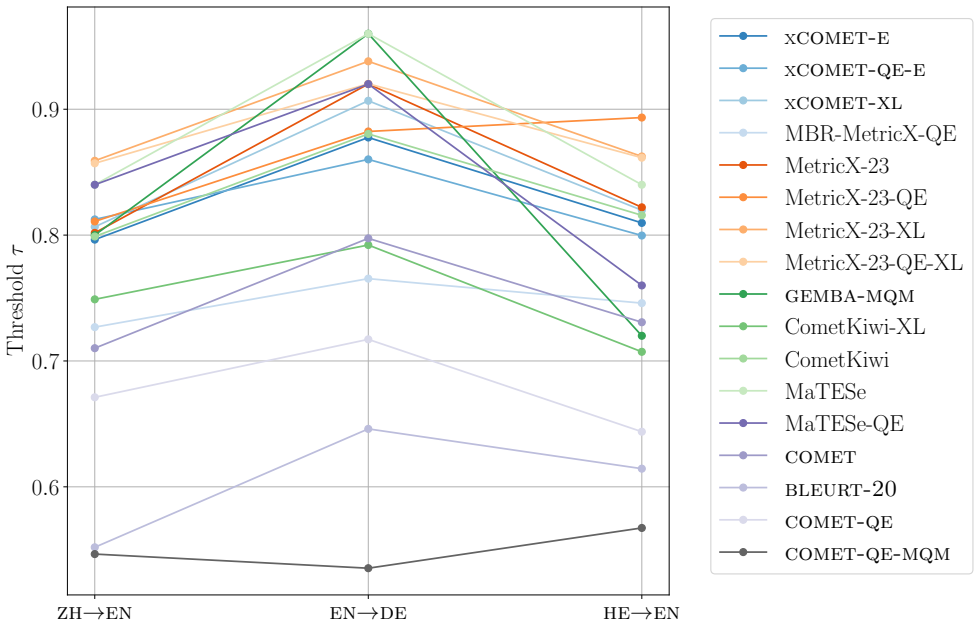


Figure 4.1. Tested metrics' optimal threshold values across different language directions. The thresholds were selected to maximize the F-score on the test set in the GOOD vs BAD binary classification scenario. Thresholds are normalized between 0 and 1 for improved clarity.

order to maximize Recall.

In general, we believe that the characteristics of the development set play an important role in determining appropriate thresholds for data filtering applications. Aside from the translation direction, evaluation datasets can differ in the MT systems employed to translate the source texts, the data domains included, and the human annotators who assessed translation quality. Therefore, while leaving the investigation of these phenomena to future work, we recommend estimating optimal thresholds using as much annotated data as possible, to prevent the peculiarities of any single dataset from overly influencing the estimated values. To support this, we are releasing our evaluation framework with options for estimating optimal metric thresholds across several datasets, depending on the user's Precision and Recall requirements.

What is the human performance?

As a reference for human performance, we examine the agreement between two annotation schemas: DA+SQM and MQM. We use DA+SQM as a metric and show the results in

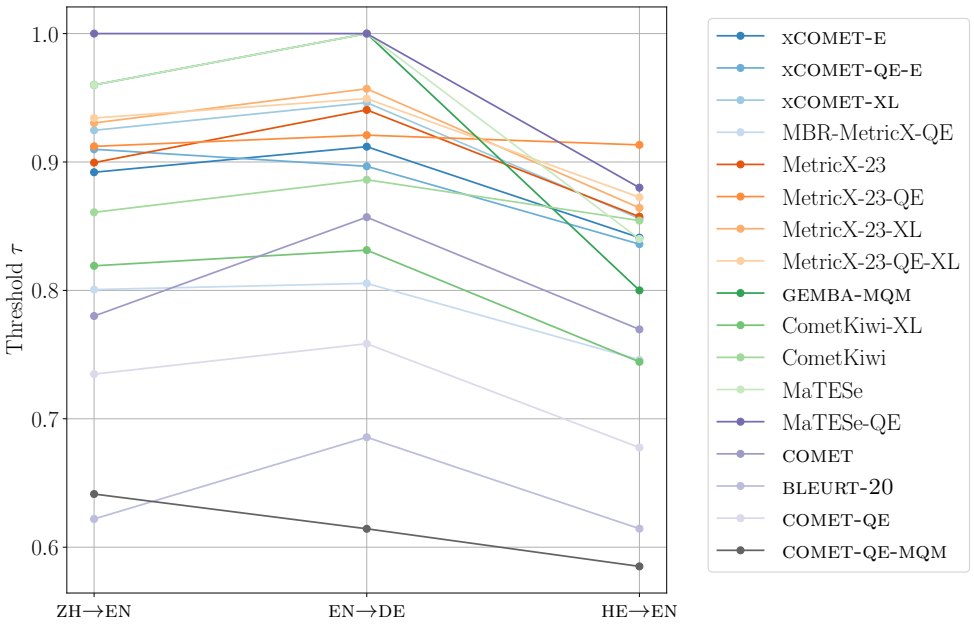


Figure 4.2. Tested metrics' optimal threshold values across different language directions. The thresholds were selected to maximize the F-score on the test set in the PERFECT vs OTHER binary classification scenario. Thresholds are normalized between 0 and 1 for improved clarity.

the last row of Table 4.3. It is important to note that DA+SQM annotations do not fully encompass the set of MQM annotations. Consequently, we evaluate its performance on a subset of the data, including $\approx 70\%$ of the data used for metrics, which means that their scores are not directly comparable. However, we observe that DA+SQM performs poorly in absolute terms, with particularly low Precision and a much higher Recall. We hypothesize that this might be due to DA+SQM's annotation guidelines, which instruct annotators to assign only a general quality score to translations rather than identifying specific translation errors, as is done within the MQM framework.[13] This might lead to noisy annotations, rendering DA+SQM less suitable for fine-grained translation quality assessments.

We investigate this matter further by restricting the data available to MT metrics to that of DA+SQM. This way, we measure the segment-level correlations with MQM scores of metrics and DA+SQM annotations jointly. We report results in Tables 4.8a and 4.8b. We find DA+SQM annotations correlate less strongly with MQM than all tested automatic metrics.

[13]Within DA+SQM, human annotators rate a translation from 0 to 100 using a slider with 7 marked levels, where each level is paired with a description of the corresponding translation quality.

Table 4.8. Kendall τ and Pearson ρ correlation coefficients, and acc_{eq} accuracy (Deutsch et al., 2023), measured between the DA+SQM- and MQM-based annotations, and between MT metrics and MQM. The data is the intersection between $\text{WMT23}_{\text{MQM}}$ and $\text{WMT23}_{\text{DA+SQM}}$.

(a) ZH→EN.

Metric	τ	ρ	acc_{eq}
GEMBA-MQM	0.36	0.43	0.52
xCOMET-ENSEMBLE	0.30	0.42	0.54
xCOMET-QE-ENSEMBLE	0.26	0.37	0.53
xCOMET-XL	0.26	0.38	0.52
MBR-MetricX-QE	0.29	0.43	0.53
MetricX-23	0.26	0.37	0.53
MetricX-23-QE	0.24	0.35	0.52
MetricX-23-XL	0.25	0.36	0.52
CometKiwi	0.25	0.37	0.52
CometKiwi-XL	0.25	0.38	0.52
COMET	0.25	0.36	0.51
BLEURT-20	0.26	0.37	0.52
MaTESe	0.27	0.33	0.48
COMET-QE-MQM	0.16	0.21	0.48
MaTESe-QE	0.21	0.24	0.44
DA+SQM	0.11	0.20	0.42
Random-sysname	0.02	0.02	0.38

(b) EN→DE.

Metric	τ	ρ	acc_{eq}
GEMBA-MQM	0.40	0.48	0.57
MBR-MetricX-QE	0.40	0.54	0.58
xCOMET-ENSEMBLE	0.38	0.54	0.60
xCOMET-XL	0.37	0.51	0.60
MetricX-23	0.37	0.51	0.60
COMET	0.37	0.51	0.58
BLEURT-20	0.37	0.49	0.57
MetricX-23-XL	0.36	0.49	0.59
xCOMET-QE-ENSEMBLE	0.36	0.51	0.59
MetricX-23-QE	0.36	0.51	0.60
CometKiwi-XL	0.35	0.50	0.57
CometKiwi	0.33	0.46	0.57
COMET-QE-MQM	0.29	0.39	0.54
MaTESe	0.29	0.33	0.53
MaTESe-QE	0.28	0.34	0.52
DA+SQM	0.17	0.29	0.46
Random-sysname	0.08	0.12	0.41

We wish to emphasize that Kocmi et al. (2023) have already noted that DA+SQM-based annotations exhibit reduced precision in distinguishing between MT systems with similar performance, compared to MQM-based annotations. Furthermore, in a recent study, Kocmi et al. (2024d) compared the correlation of several human annotation schemes with MQM and found that DA+SQM performed poorly. In line with these findings, our results raise additional concerns regarding DA+SQM reliability, showing lower correlation with MQM than all tested automatic metrics.

4.4.2 Translation Re-Ranking

We present the results in the last two columns of Tables 4.3, 4.4, and 4.5. To facilitate interpretation of these results, we also calculate the average MQM score of the translations ranked highest by MT metrics, and report it in the last column. Additionally, it is important to note that there are 15, 12, and 13 translations per source text in ZH→EN, EN→DE, and HE→EN, respectively.

As shown, metric precision ranges from 30% to 70%, with the highest precision rate of 70.01% achieved by COMET in HE→EN. Furthermore, the top-performing metrics yield average MQM scores of ≈ -1 in HE→EN, indicating that their highest-ranked translations contain, on average, one Minor error. In translation re-ranking as well, we record the highest scores on HE→EN, likely because this split of the test set is inherently easier, given the distribution of MQM scores (Table 4.2).

Notably, reference-based metrics consistently outperform reference-free ones. By looking at pairs of metrics of the same family, XCOMET-ENSEMBLE, MetricX-23, and MetricX-23-XL outperform XCOMET-QE-ENSEMBLE, MetricX-23-QE, and MetricX-23-QE-XL, respectively, across all translation directions. However, in real-world translation re-ranking scenarios, references are not available. To account for this, we assess the performance of reference-based metrics when used as the utility function in an MBR decoding-like scenario, therefore not relying on the presence of reference translations.

Metric performance in MBR decoding Table 4.9 shows the translation re-ranking performance of reference-based metrics when used as the utility function for MBR decoding, and comparing it to the reference-based re-ranking scenario presented in the last columns of Tables 4.3, 4.4, and 4.5. On average, the absence of reference translations reduces performance. However, this is not true for all language directions, as MBR decoding outperforms reference-based re-ranking in ZH→EN. We believe this exception is due to the particularly poor quality of the references in the ZH→EN split of $\text{WMT23}_{\text{MQM}}$, as discussed by Freitag et al. (2023).

Furthermore, the last two rows of Table 4.9 present the performance of the best and second-best openly available, reference-free metrics in translation re-ranking. Our results indicate that reference-based metrics used as the utility function for MBR decoding tend

Table 4.9. Re-Ranking Precision of reference-based metrics when used as the utility function for MBR decoding, compared with the reference-based re-ranking scenario in the last two columns of Tables 4.3, 4.4, and 4.5. The last two rows of this table show the performance of the best and second-best reference-free metrics in translation re-ranking.

	ZH→EN		EN→DE		HE→EN	
Metric	**MBR**	**Tab 4.3**	**MBR**	**Tab 4.4**	**MBR**	**Tab 4.5**
XCOMET-XL	40.27	37.49	44.03	47.31	65.05	68.31
MetricX-23-XL	41.10	39.52	48.24	47.81	63.71	67.17
MaTESe	35.27	33.07	44.53	43.18	60.20	61.99
COMET	37.20	34.25	45.12	48.26	66.38	70.01
BLEURT-20	36.07	33.35	47.76	48.27	63.70	68.33
#1 REF FREE	–	39.28	–	45.71	–	65.61
#2 REF FREE	–	38.78	–	45.57	–	63.25

to outperform reference-free metrics. Specifically, the best performance is achieved by two reference-based metrics: MetricX-23-XL, in ZH→EN and EN→DE, and COMET in HE→EN. Again, by looking at metrics of the same family, MetricX-23-XL outperforms MetricX-23-QE-XL across the board. These findings suggest that translation re-ranking using MBR decoding may be more reliable than QE re-ranking.[14] Previous studies have already compared MT metrics in MBR decoding and QE re-ranking. For example, Freitag et al. (2022a) and Fernandes et al. (2022) employ human annotators to evaluate the quality of translations produced by MT systems when using one or the other re-ranking technique. However, due to the high cost of human annotations, these studies were limited to including only a few metrics. Furthermore, their findings may need to be revisited to determine whether they remain valid with the introduction of new metrics. In contrast, by relying solely on the human annotations released annually at WMT, our setup facilitates updating results as soon as new metrics or datasets become available.

[14]This is based on the assumption that using translations generated by distinct MT systems can serve as an effective approximation of sampling from a single system. We delve into the differences between MBR decoding and our evaluation setup in the Limitations.

4.5 Conclusion

In this chapter, we continued to mitigate the interpretability issue introduced in Section 1.1. Specifically, we tackled the *Performance* factor: correlation with human judgments fails to provide users with a clear understanding of metric performance. To address this issue, we introduced a novel evaluation framework for MT metrics, which we used to measure metric performance in i) binary classification, i.e., distinguishing between GOOD and BAD, and PERFECT and OTHER translations, and ii) in a proxy scenario for translation re-ranking, selecting the best among the translations of the same source text. By measuring performance in terms of Precision, Recall, F-score, and Re-Ranking Precision, we fulfilled a dual purpose. First, we offered a more intuitive interpretation of metrics' capabilities – as compared to correlation with human judgment – providing insights into absolute metric performance, and second, we provided concrete user recommendations concerning novel MT metric applications.[15] We found that MT metrics perform relatively well in distinguishing between GOOD and BAD translations, but struggle with Precision, especially when dealing with higher-quality translations like in the PERFECT vs OTHER scenario. Our results show that MetricX-23-QE-XL is the best openly available metric for data filtering applications, while MetricX-23-XL and COMET achieve the highest performance in translation re-ranking. Additionally, we demonstrated that reference-based MT metrics, when used as the utility function in an MBR decoding-like scenario, outperform reference-free ones, suggesting that MBR decoding may be superior to QE re-ranking. Finally, we reported notably poor performance for DA+SQM annotations used as a metric within our evaluation framework, raising concerns about its reliability.

[15] We emphasize that this study is not intended to replace correlation as the standard meta-evaluation measure. On the contrary, we argue that our meta-evaluation framework should be used alongside correlation-based analyses to provide complementary insights into metric behavior.

Chapter 5

Improving the MT Meta-Evaluation with Sentinel Metrics

As mentioned in Section 1.1, a significant limitation of most neural-based metrics is that they are black-box systems trained to predict human judgments in the form of scalar scores. Besides rendering them challenging to interpret, such opacity also makes them challenging to evaluate, complicating the MT meta-evaluation. However, a flawed meta-evaluation risks steering the development of MT metrics in the wrong direction, which, in turn, may compromise the evaluation of MT systems.

In this chapter, we identify issues with certain meta-evaluation strategies recently employed in the context of WMT, which we found to favor specific metric categories and potentially encourage undesirable metric behavior. To demonstrate these problems, we introduce the concept of sentinel metrics, i.e., a suite of metrics serving as a probe to identify pitfalls in the meta-evaluation process.

5.1 The MT Meta-Evaluation

In this section, we briefly remind the reader of the standard practices in the MT meta-evaluation, i.e., the evaluation of MT metrics (illustrated in Section 2.2). Specifically, we focus on the meta-evaluation as conducted at the WMT Metrics Shared Task.

Yearly, the WMT Metrics Shared Task organizes a competition among metrics, including participants' submissions and baselines, to identify the metric that most closely

aligns with human judgments. First, multiple MT systems are employed to translate source segments found in one or more test datasets.[1] Consequently, test datasets contain several translations of the same source segment. Second, a manual evaluation campaign is carried out to assess the quality of all translations. Finally, metrics' capabilities are assessed based on their alignment with human judgments. Such alignment is typically estimated using correlation with human judgments. As already mentioned in Chapter 2, metrics are evaluated at two granularity levels:

- At the **segment level**, metrics are evaluated according to their ability to discern between higher- and lower-quality translations.
- At the **system level**, segment-level metric scores are averaged to produce a quality score for each MT system; then, metrics are evaluated based on their ability to discriminate between superior and inferior MT systems.

At both granularity levels, metrics can be evaluated using several statistical methods, such as Kendall's τ and Pearson's ρ correlation coefficients, which have traditionally been applied at the segment and system levels, respectively. A final metrics ranking is derived by aggregating results from all the chosen statistics. For example, at WMT23, the final ranking was computed from the following three statistics:

1. System-level pairwise ranking accuracy (Kocmi et al., 2021), which evaluates metrics based on their ability to rank systems in the same order as human judgments.
2. System- and segment-level Pearson correlation, which measures the degree to which metric scores and human scores are correlated linearly.
3. Segment-level pairwise ranking accuracy with tie calibration (Deutsch et al., 2023), which evaluates metrics based on their ability to rank segments in the same order as human judgments, or their ability to predict ties correctly.

In this chapter, we identify two critical issues related to the second and third statistics, and provide the following recommendations to address them:

[1] A segment typically refers to a single sentence, but can also include multiple sentences. For instance, at WMT23, a segment represents an entire paragraph rather than a single sentence for the English-to-German translation direction.

- **Translations should be grouped by their source segment before calculating segment-level correlations** (Section 5.2).
- **Tie calibration should not be conducted on the test set** (Section 5.3).

In the following sections, we illustrate the flaws of the aforementioned statistics, demonstrating their impact by leveraging sentinel metrics.

5.2 To Group or Not to Group?

At early editions of the WMT Metrics Shared Task (Macháček and Bojar, 2013, 2014; Stanojević et al., 2015; Bojar et al., 2016b), human assessments were collected in the form of Relative Rankings (RR). Specifically, the annotators were tasked to rank up to 5 translations of the same source sentence, produced by different MT systems. From each ranking, up to 10 pairwise comparisons were extracted. Despite metrics assessments being scalar scores – which theoretically enabled the comparison of all pairs of translated segments – correlation was measured only on those pairs of translations for which RR annotations were available. Therefore, only translations of the same source sentence were compared. Later on, at subsequent editions of WMT, new techniques for human evaluation were adopted: first, Direct Assessments (DA, Graham et al., 2013) – where annotators rate individual translations on a scale from 0 to 100 – then, Multidimensional Quality Metrics (MQM, Lommel et al., 2014b; Freitag et al., 2021a) – where annotators tag the spans of a translation that contain errors, specifying their category and severity.[2] With both the new annotation schemas, each translated segment was assigned a scalar quality score independently of the other translations, which made it possible to compare all translations, not only those of the same source sentence. This new possibility raised doubts regarding the best way to compute the correlation between metrics and human assessments. Indeed, it could be computed using all translations at once – *No Grouping* – or by first grouping translations based on either their source segment – *Segment Grouping* – or the system that produced them – *System Grouping* – and then returning the average correlation of these groups.

At the WMT21 Metrics Shared Task, Freitag et al. (2021b) chose the *No Grouping* strategy, arguing that the other options would provide only a partial view of the overall

[2]Please refer to Section 2.1.1 for further details about techniques for human MT evaluation.

picture. At WMT22, all three grouping strategies were used (Freitag et al., 2022b), and later at WMT23, Freitag et al. (2023) chose *No Grouping* again. Although *No Grouping* is the only strategy that assesses the MT metrics' ability to discern between higher- and lower-quality translations in absolute terms, irrespective of the source segment or MT system, we show that both *No Grouping* and *System Grouping* may introduce unfairness and favor trained metrics over the rest.

5.2.1 The Relation Between Spurious Correlations and Grouping

Most neural-based metrics are trained with a regression objective to approximate human judgments. They are expected to infer by pattern-matching the relation between human judgments and various phenomena, such as omissions, additions, or other translation errors. However, this mechanism might inadvertently lead to the detection of patterns that are not in a causal relation with the concept of translation quality, but are instead spurious correlations, e.g., the length of a translation, or the number of named entities in it, among others. Arguably, the meta-evaluation should not reward metrics for basing their assessments on spurious correlations between the features of the source, translation, or reference, and the human judgments. However, our intuition is that *No Grouping* and *System Grouping* strategies might be doing so by allowing the comparison of translations from different sources. To simplify, consider a metric that unfairly penalizes a translation solely because it contains many named entities. Using *No Grouping* or *System Grouping*, such a metric might have a non-negative correlation with human judgments if, on average, translating sentences containing many named entities is more challenging than translating other sentences, because MT systems would be making more mistakes in translating them. Therefore, exploiting such a pattern might be beneficial even though it is not causally related to the quality of a translation. In contrast, when using *Segment Grouping*, such a pattern would be ineffective, as different translations of the same source sentence should contain the same amount of named entities. More generally, we expect *Segment Grouping* to lessen the impact of most spurious correlations derived from features shared by a source sentence and its translations.

To assess the extent of this issue, we incorporate three sentinel metrics into the current meta-evaluation framework and recompute the metrics' rankings using all grouping strategies. Crucially, we find that the impact of spurious correlations when *No Grouping* and

System Grouping strategies are employed is substantial – favoring trained metrics over the rest[3] – and is significantly reduced with *Segment Grouping*.

5.2.2 The Sentinel Metrics

This section describes the three sentinel metrics employed to measure the impact of grouping strategies on the meta-evaluation process:

1. **sentinel-cand**, which assesses the quality of a translation without taking its source or reference as input.

2. **sentinel-src**, which predicts the quality of a translation solely based on its source.

3. **sentinel-ref**, which predicts the quality of a translation solely based on its reference.

Lacking information about the translation to be evaluated, the training process of sentinel-src and sentinel-ref (described in the next section) can only lead them to learn spurious correlations between the features of the source and reference sentences, respectively, and the human judgments. sentinel-cand, on the other hand, is a metric with partial information. It is indeed possible to assess a translation's fluency and grammatical correctness without comparing it to its source or reference sentences, though not its adequacy. Nonetheless, we also expect sentinel-cand to rely, at least in part, on spurious correlations.

5.2.3 Training the Sentinel Metrics

Sentinel metrics employ XLM-RoBERTa large (Conneau et al., 2020) as their backbone model, with a multi-layer fully-connected neural network on top of the [CLS] token, which is used to output predictions in the form of scalar scores. We train sentinel metrics to minimize the Mean Squared Error (MSE) between their predicted scores and human judgments. Our dataset comprises a selection of data from WMT spanning 2017 to 2022, incorporating DA and MQM scores.

Following Rei et al. (2022a), we train sentinel metrics for a single epoch using DA from 2017 to 2020 and fine-tune them for a further epoch using MQM data. The input for

[3]Indeed, overlap-based metrics such as BLEU (Papineni et al., 2002) and chrF (Popović, 2015), or LLM-based metrics such as GEMBA-MQM (Kocmi and Federmann, 2023a), were not trained to mimic human assessments and should not be able to leverage spurious correlations.

Hyperparameter	Value
Optimizer	RAdam (Liu et al., 2021)
Learning Rate	1e-6
Number of Epochs	1
Batch Size	8
Accumulation Steps	2
Dropout	0.1
Dimension of $\boldsymbol{h}_t^{(1)}$	512
Dimension of $\boldsymbol{h}_t^{(2)}$	128

Table 5.1. Sentinel metrics training hyperparameters.

the sentinel metrics consists of either the source text (sentinel-src), candidate translation (sentinel-cand), or reference translation (sentinel-ref). Each sentence is tokenized and passed to the XLM-RoBERTa large model, which serves as a feature extractor. Then, we pass the embedding of the [CLS] token to a multi-layer, fully-connected neural network, which outputs the final scalar score. More formally, considering t as the input text for a sentinel metric:

$$
\begin{aligned}
\boldsymbol{e}_t &= \text{XLM-R}\,(t) \\
\boldsymbol{h}_t^{(1)} &= \text{Dropout}\left(\text{Tanh}\left(W_h^{(1)}\boldsymbol{e}_t + \boldsymbol{b}_h^{(1)}\right)\right) \\
\boldsymbol{h}_t^{(2)} &= \text{Dropout}\left(\text{Tanh}\left(W_h^{(2)}\boldsymbol{h}_t^{(1)} + \boldsymbol{b}_h^{(2)}\right)\right) \\
s_t &= W_o\boldsymbol{h}_t^{(2)} + \boldsymbol{b}_o
\end{aligned}
$$

Where:

- t is the tokenized input sentence.
- $\boldsymbol{e}_t$ is the [CLS] token embedding at the output of XLM-RoBERTa large.
- $\boldsymbol{h}_t^{(i)}$ represents the output of the i^{th} layer of the fully-connected neural network. Each layer consists of a linear transformation, using weight matrix $W_h^{(i)}$ and bias vector $\boldsymbol{b}_h^{(i)}$, followed by a Tanh activation function and a dropout layer.
- W_o and $\boldsymbol{b}_o$ are the output layer's weight matrix and bias vector, respectively.

- s_t is the output scalar score assigned to sentence t.

Both training phases (i.e., the first, using DA-based human judgments, and the second, using MQM-based ones) employ the same set of hyperparameters, detailed in Table 5.1.

5.2.4 Results

In Tables 5.2, 5.3, and 5.4, we report the ranking derived from the segment-level Pearson correlation of the primary submissions to the Metrics Shared Task of WMT23, where we included sentinel metrics, in the language directions ZH→EN, EN→DE, and HE→EN, respectively, under all three grouping strategies.

Although our results are consistent across the three tested language directions, we focus here on ZH→EN, where the patterns are particularly striking. Specifically, we find that sentinel-src ranks fourth and third under *No Grouping* and *System Grouping*, respectively, outperforming strong baselines like COMET or BLEURT-20, and even state-of-the-art metrics like GEMBA-MQM. The only metrics that surpass it are large regression-based models such as XCOMET-ENSEMBLE (Guerreiro et al., 2024) and MetricX-23 (Juraska et al., 2023), which might have learned the same spurious correlations leveraged by the sentinel metrics, in addition to non-spurious patterns (cf. Section 5.2.4).

In contrast, under *Segment Grouping*, both sentinel-src and sentinel-ref are correctly positioned at the bottom of the ranking,[4] while sentinel-cand drops to the 11th place, compared to the 3rd and 2nd with *No Grouping* and *System Grouping*, respectively. A notable difference between the three grouping strategies is the positioning of GEMBA-MQM: It ranks 7th and 9th under *No Grouping* and *System Grouping*, respectively, but rises to first place with *Segment Grouping*. We hypothesize that this is due to GEMBA-MQM being based on GPT-4, which has not been explicitly fine-tuned on human assessments and is thus less likely to leverage spurious correlations such as those described in Section 5.2.1. Interestingly, under both *No Grouping* and *System Grouping*, all sentinel metrics outperform GEMBA-MQM.

Finally, we note that sentinel-cand is the only sentinel metric that does not rank at the very bottom under *Segment Grouping*. It outperforms prismSrc (Thompson and Post, 2020) and embed_llama (Dreano et al., 2023), in ZH→EN, while it ranks as high as 6th in EN→DE.

[4]This is expected, as both metrics assign identical scores to all translations of the same source segment.

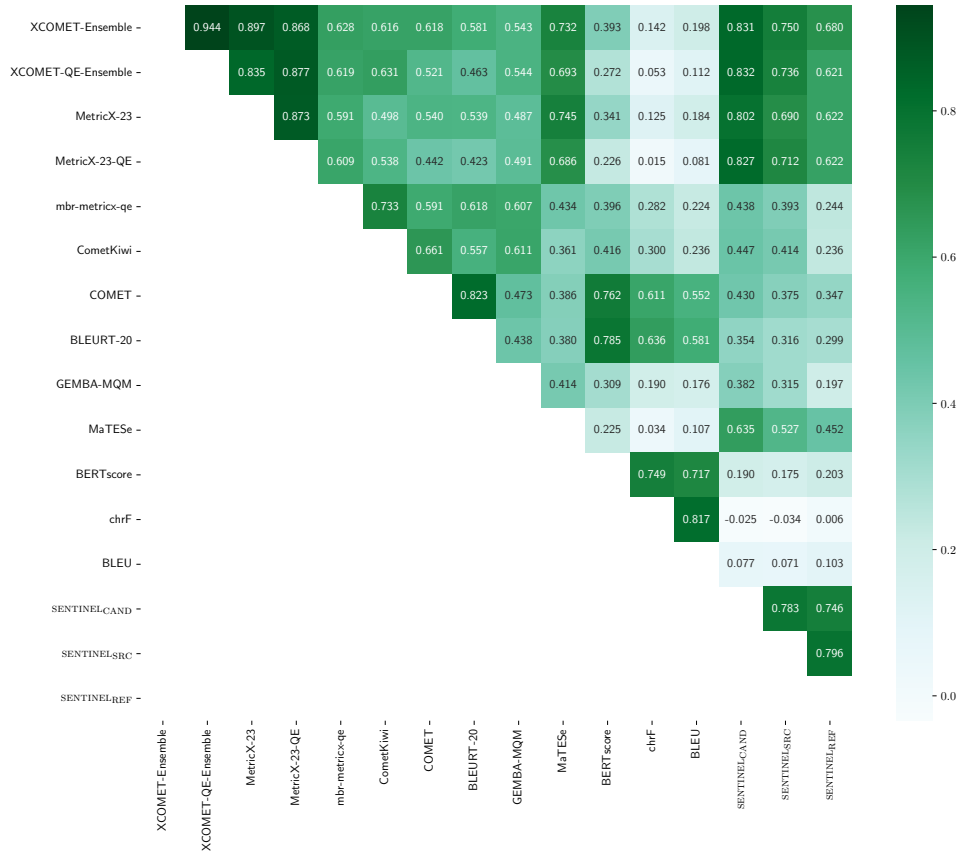


Figure 5.1. Pairwise correlation between a part of the primary submissions and baselines of WMT23, and sentinel metrics. Correlation is Pearson with *No Grouping*, and the language direction is ZH→EN.

This suggests that focusing solely on the candidate translation – specifically, its fluency and grammatical correctness – can be sufficient to outperform some weaker metrics, at least in terms of Pearson correlation with human judgments. Moreover, our results may shed light on the open issue noted at WMT23 regarding the inconsistency between segment-level and system-level correlations for prismSrc. Freitag et al. (2023) noticed that prismSrc, despite achieving moderate segment-level correlation, displayed negative correlation at the system level. As shown in Table 5.2, prismSrc ranks 15th out of 24 under *No Grouping* but 13th out of 14 under *Segment Grouping* – that is, in the second-to-last significance cluster, close to the sentinel metrics – and this performance drop is consistent across language directions. This aligns with the negative correlation achieved by prismSrc at the system level.

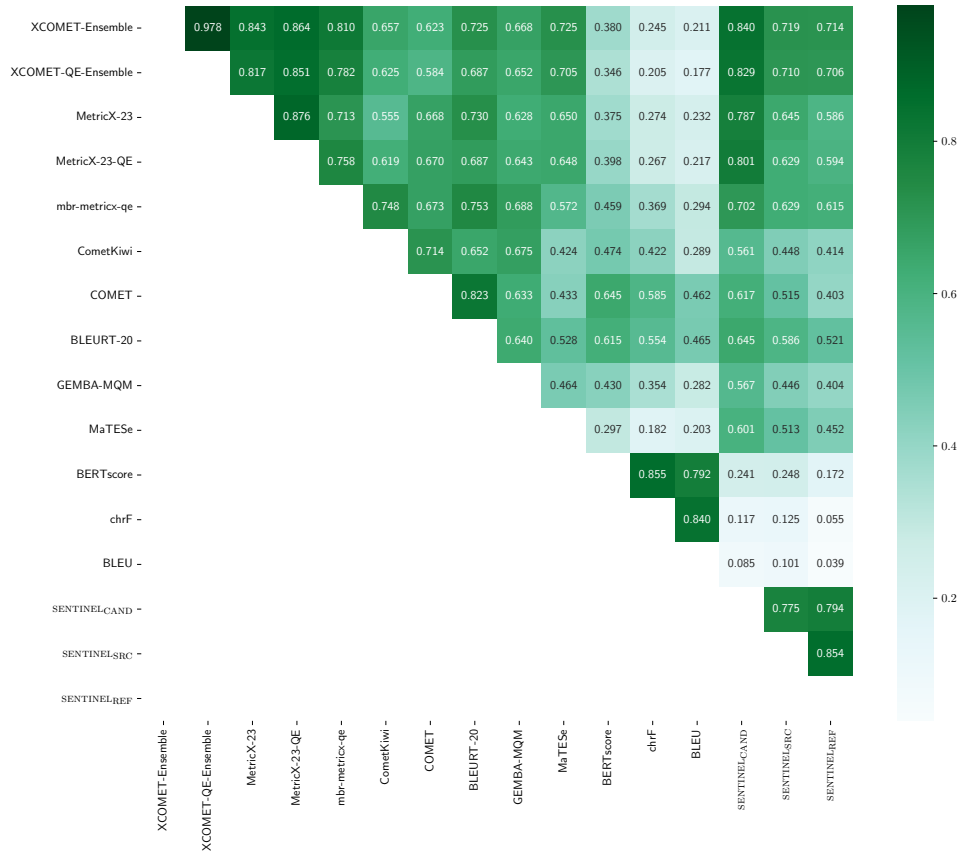


Figure 5.2. Pairwise correlation between a part of the primary submissions and baselines of WMT23, and sentinel metrics. Correlation is Pearson with *No Grouping*, and the language direction is EN→DE.

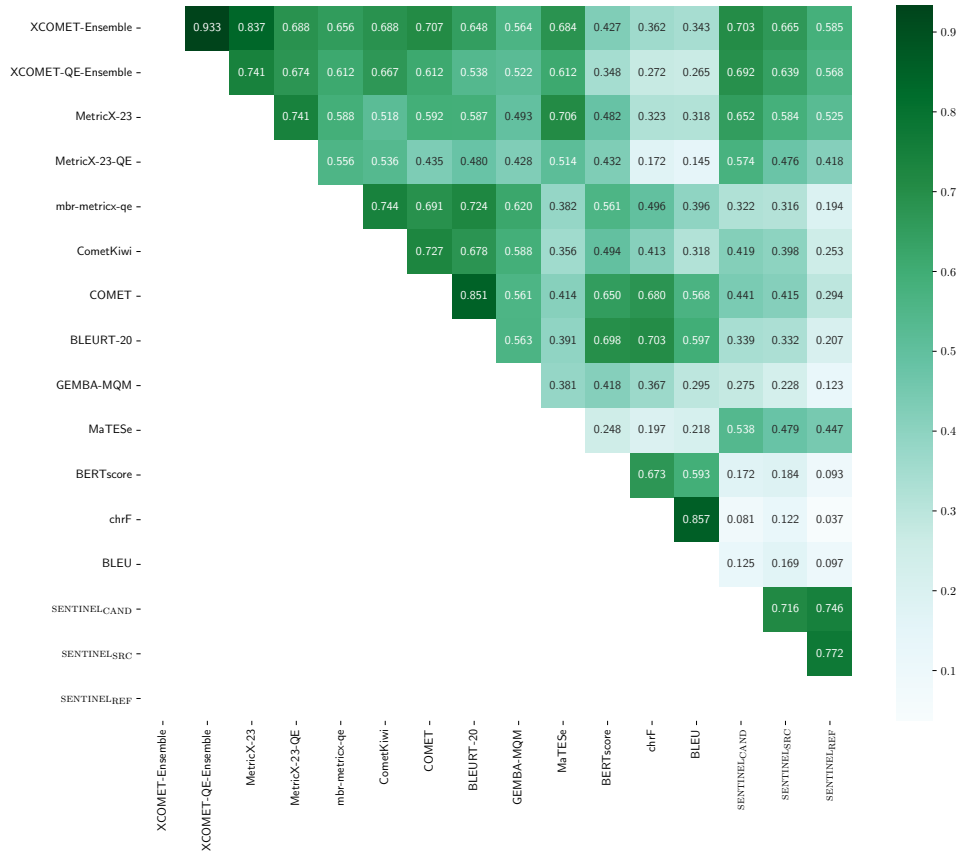


Figure 5.3. Pairwise correlation between a part of the primary submissions and baselines of WMT23, and sentinel metrics. Correlation is Pearson with *No Grouping*, and the language direction is HE→EN.

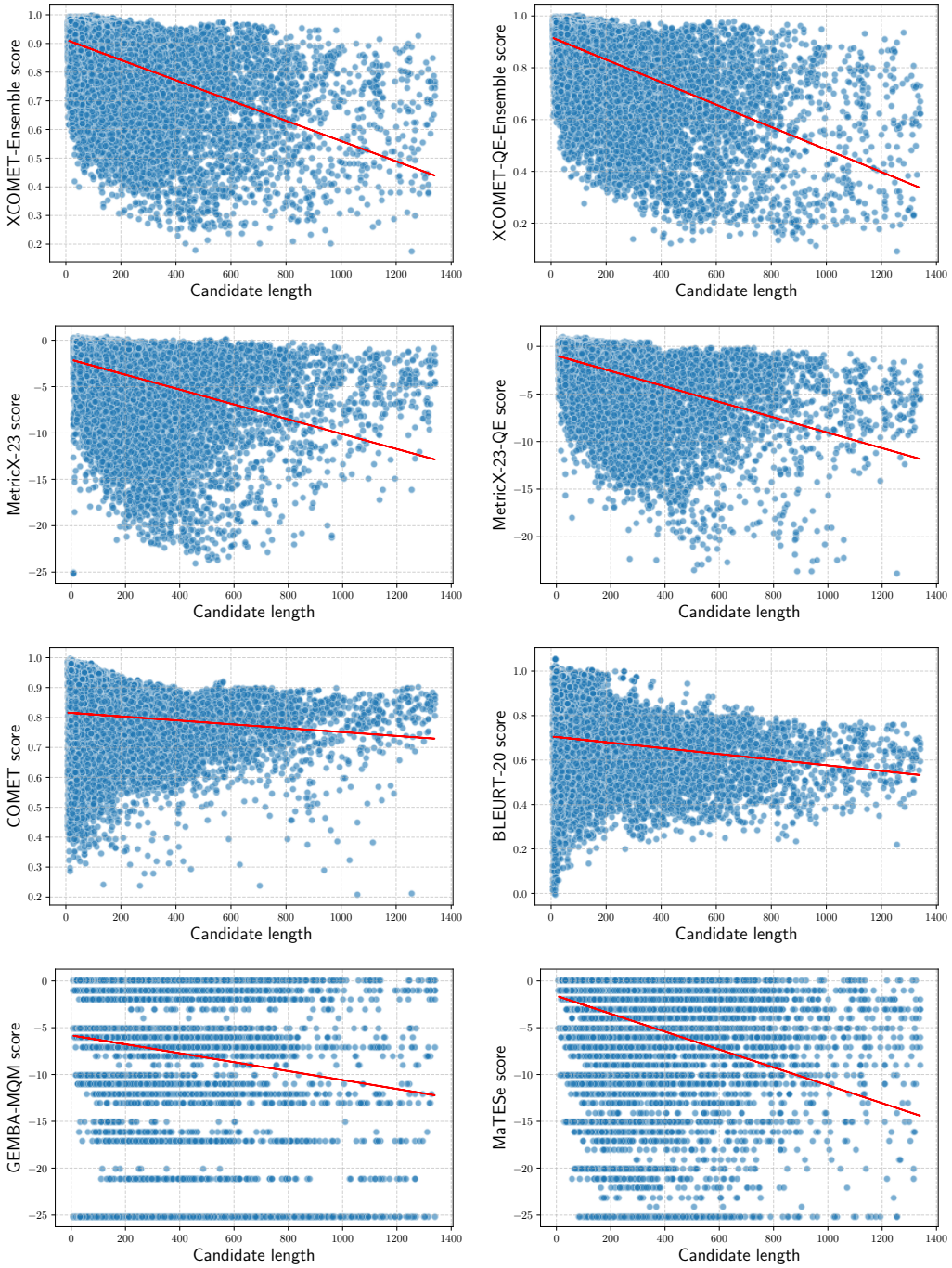


Figure 5.4. Metric assessments over translation length for a subset of the metrics that participated in WMT23. The red line represents the least-squares regression.

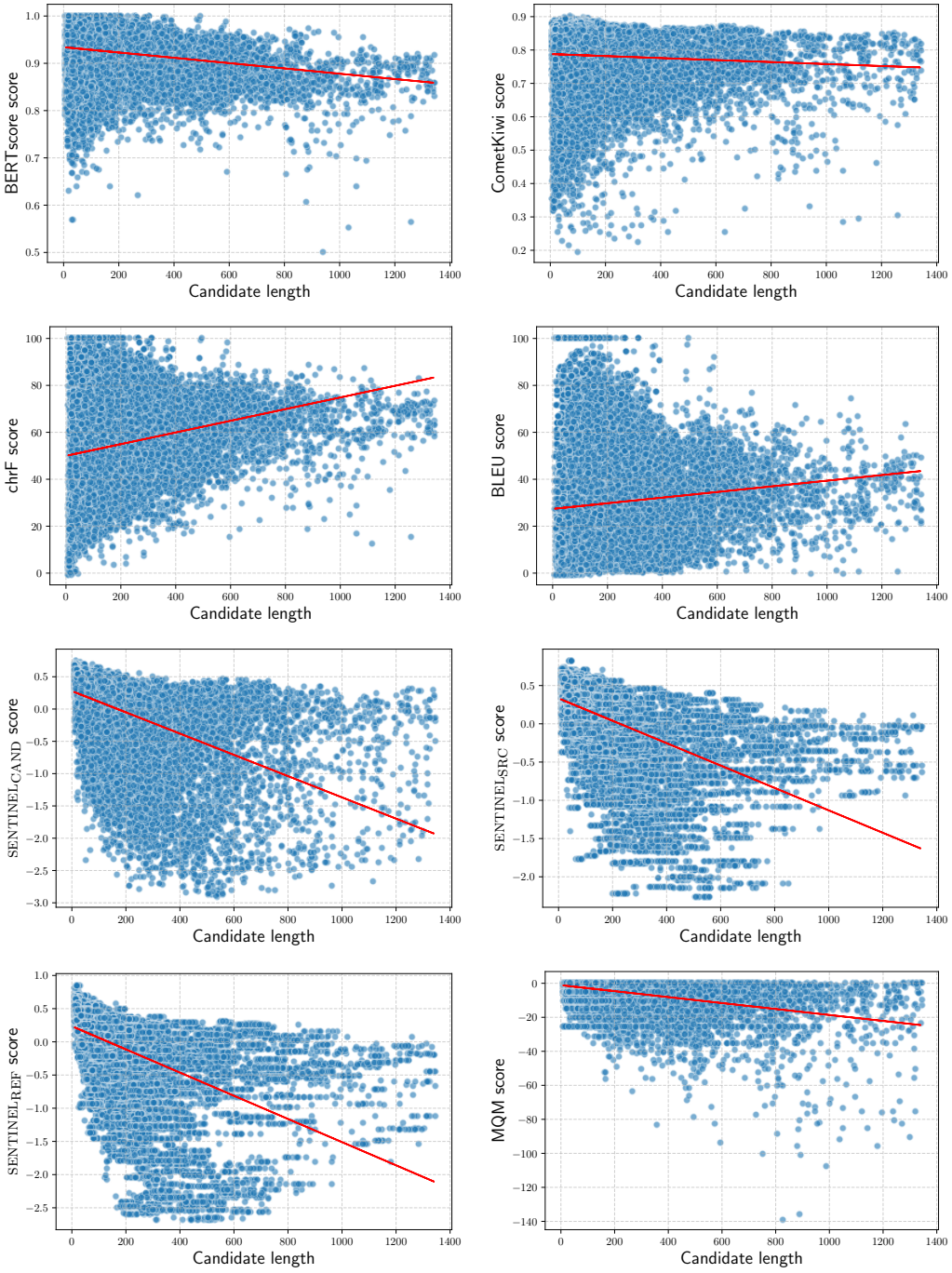


Figure 5.5. Metric assessments over translation length for a subset of the metrics that participated in WMT23, together with sentinel metrics. The red line represents the least-squares regression.

Are MT metrics learning from spurious correlations?

We hypothesize that certain trained metrics may be basing their assessments on the same spurious correlations as those leveraged by the sentinel metrics. To delve deeper into this, we measure their segment-level Pearson correlation with the sentinel metrics using *No Grouping*, and report results in Figures 5.1, 5.2, and 5.3. Surprisingly, XCOMET-ENSEMBLE, XCOMET-QE-ENSEMBLE, MetricX-23, and MetricX-23-QE, which are the only metrics that consistently surpass the sentinels in Tables 5.2, 5.3, and 5.4, display a high correlation with all three sentinel metrics. In ZH→EN, their correlation with sentinel-src is 0.750, 0.736, 0.690, and 0.712 (Figure 5.1), respectively, higher than their correlation with human judgment, which is 0.650, 0.647, 0.625, and 0.647, respectively (Table 5.2), and similar patterns are observed in the other translation directions.

We recognize that these metrics share many similarities with our sentinels, as both are neural transformer-based systems and both were trained with the same regression-based objective, using largely the same data. This similarity likely contributes to the high correlation values observed. However, with access limited to only the source segment, sentinel-src relies exclusively on spurious correlations to conduct the evaluation. For this reason, we argue that these results raise concerns about the reliability of state-of-the-art MT metrics, which may be learning to exploit spurious correlations to minimize the Mean Squared Error with human judgments during training. To further support our hypothesis, we measure the relation between a subset of the tested metrics' assessments and translation length, which serves as a simple spurious correlate of translation quality.[5] Additionally, we plot the distribution of the corresponding MQM human judgments over translation length. All results are presented in Figures 5.4 and 5.5, where we concatenate the data from all three language directions used in the MQM-based meta-evaluation of WMT23, i.e., ZH→EN, EN→DE, and HE→EN.

Let us take XCOMET-ENSEMBLE as the running metric to discuss our results. XCOMET-ENSEMBLE scores decrease at increasing candidate lengths, with the metric almost never assigning scores higher than 0.9 to translations longer than 400 characters. However,

[5]We expect that longer sentences are, on average, more challenging to translate. Therefore, we anticipate that MT metrics might have learned to assign lower scores to longer translations, despite the length and quality of translations not being causally related.

the distribution of human judgments shows that human annotators rated many of those translations as perfect or near-perfect, indicating that XCOMET-ENSEMBLE might be biased to assign lower scores to longer translations, irrespective of their quality.

More generally, most regression-based metrics, sentinels included, tend to never assign particularly high scores to long translations, even when they are correct. This is in marked contrast to metrics trained with different objectives, such as MaTESe, or not fine-tuned to mimic the human judgment, such as GEMBA-MQM. Indeed, both these metrics assign their highest score to several translations longer than 1200 characters. Notably, there are several metrics whose assessments converge to a very narrow range of values as length increases. For example, BLEURT-20's assessments seem to be confined between approximately 0.4 and 0.8 for translations longer than 1000 characters, and a similar pattern is observed for COMET.

Finally, by looking at the relation between MQM scores and length (At the bottom of Figure 5.5), we note that the least-squares regression lines indicate that longer translations contain more errors than shorter ones and are therefore assigned lower scores by human annotators. This is expected, as longer sources are more difficult to translate, and therefore their translations will contain more errors, on average. However, this suggests that detecting biases of this type may be particularly complex without datasets specifically crafted for the purpose.

5.3 The Evaluation of Ties

In this Section, we focus on the third statistic among those described in Section 2.2, i.e., the segment-level pairwise ranking accuracy with tie calibration, dubbed acc_{eq} by Deutsch et al. (2023). Prior to WMT23, the organizers of the Metrics Shared Task used to employ Kendall's τ coefficient – a statistic used to estimate the rank-based agreement between two sets of measurements (Kendall, 1945) – to measure the correlation between metrics and human judgments at the segment level. Deutsch et al. (2023) pointed out that Kendall's τ coefficient does not account for metrics correctly predicting ties,[6] and introduced acc_{eq} to address this issue (illustrated in Section 2.2). Unfortunately, our analysis indicates that the

[6]Given a pair of translations whose quality has been assessed by human annotators, such a pair is tied if both translations are assigned the same score.

tie calibration algorithm used to compute acc_{eq} inadvertently compromises meta-evaluation fairness in order to accommodate ties, ultimately biasing the results in favor of continuous metrics[7] over discrete ones.

5.3.1 The Problem with Tie Calibration

As mentioned in Section 2.2, the tie calibration algorithm determines, for each metric, a threshold ϵ such that, given two metric assessments m_1 and m_2, they are tied if $|m_1 - m_2| \leq \epsilon$. Deutsch et al. (2023) propose selecting ϵ as the value that maximizes acc_{eq} on the same test set used for the meta-evaluation. This enables metrics to output the number of tied scores that best fit the distribution of human ties in the considered test set. This distribution is not stable across test sets (Table 5.5), and Deutsch et al. (2023) show that ϵ values are not stable either. Nonetheless, they argue that this would not impact the fairness of the meta-evaluation. Unfortunately, our analysis shows that this is not the case. Specifically, despite all metrics' ϵ values being selected on the same test data, we demonstrate that continuous metrics are more flexible in fitting the underlying distribution of human ties, compared to discrete ones, leading to unfairly higher acc_{eq} values.

5.3.2 Two New Sentinel Metrics

To demonstrate the impact of this phenomenon, we introduce two additional sentinel metrics, i.e., $\text{sentinel}_{\text{GEMBA}}$ and $\text{sentinel}_{\text{MaTESe}}$. GEMBA-MQM (Kocmi and Federmann, 2023a) and MaTESe (Chapter 3) are MT metrics that output discrete scores in the form of MQM quality assessments and were submitted to the WMT23 Metrics Shared Task. $\text{sentinel}_{\text{GEMBA}}$ and $\text{sentinel}_{\text{MaTESe}}$ are perturbed versions of GEMBA-MQM and MaTESe, respectively, obtained by adding Gaussian noise – $\mathcal{N}(0, 0.0001)$ – to their predictions. By making their output continuous in the neighborhood of discrete values, we partially fill their gap with continuous metrics, while preventing any two different discrete assessments from inverting their ordering. That is, if two GEMBA-MQM's assessments m_1, m_2 are such that $m_1 > m_2$, this relation

[7] By continuous, we refer to those metrics whose assessments can take on any value within a given range, as opposed to discrete metrics, which can take on a limited set of values. Metrics from the COMET family such as COMET, XCOMET-ENSEMBLE, and CometKiwi (Rei et al., 2022b) are continuous, whereas GEMBA-MQM (Kocmi and Federmann, 2023a) and MaTESe (Chapter 3) are examples of discrete metrics.

is preserved by $\text{sentinel}_{\text{GEMBA}}$. In general, we expect a fair meta-evaluation to rank these sentinels on par or below their discrete counterparts. Furthermore, we wish to remark that this solution is sub-optimal compared to metrics that are continuous by design. Indeed, due to the addition of Gaussian noise, the ordering of all $\text{sentinel}_{\text{GEMBA}}$ and $\text{sentinel}_{\text{MaTESe}}$'s assessments in the neighborhood of discrete values is randomized.

To demonstrate that $\text{sentinel}_{\text{GEMBA}}$ and $\text{sentinel}_{\text{MaTESe}}$ can better fit the distribution of human ties compared to their discrete counterparts, we modify such a distribution in the test data. Specifically, we repeatedly sub-sample the test data, such that for each pair of tied human assessments we remove that pair from the test data with a certain probability p_t, and do the same for non-tied pairs, which are removed with probability p_n. We extract 13 samples by assigning various values to p_t and p_n and report the chosen values in Table 5.6. As a consequence, each pair (p_t, p_n) represents a different sub-sample of test data, with a different percentage of tied human pairs. Then, for each metric, we select the best ϵ and compute acc_{eq} on each of these samples.

5.3.3 Results

In Figure 5.6 (left), we present the acc_{eq} results for a subset of continuous metrics, together with GEMBA-MQM, MaTESe, $\text{sentinel}_{\text{GEMBA}}$, and $\text{sentinel}_{\text{MaTESe}}$. At first glance, it is evident that discrete metrics exhibit a distinct acc_{eq} pattern compared to continuous and sentinel metrics. Notably, at lower percentages of tied human pairs, $\text{sentinel}_{\text{GEMBA}}$ and $\text{sentinel}_{\text{MaTESe}}$ significantly outperform GEMBA-MQM and MaTESe.[8] This discrepancy arises because the tie calibration algorithm selects very small ϵ values, close to 0 for every metric, allowing the number of ties predicted by continuous metrics to potentially drop to 0. Conversely, metrics that yield discrete scores inherently produce a certain number of ties, placing them at a disadvantage, and thus ranking conceptually identical metrics like $\text{sentinel}_{\text{GEMBA}}$ and GEMBA-MQM at significantly different positions. Interestingly, in the hypothetical scenario in which there are no tied human pairs in the dataset, $\text{sentinel}_{\text{GEMBA}}$ would rank second (despite several of its assessments having a random ordering), whereas GEMBA-MQM would

[8] It is important to highlight that the range of human tie percentages explored in our analysis is similar to that found in the WMT test sets. Indeed, as shown in Table 5.5, such percentages range from a minimum of 15.14% to a maximum of 53.35%, observed in the WMT22 EN→DE test set.

be second to last. At increasing percentages of gold ties, instead, the acc_{eq} values obtained by $\text{sentinel}_{\text{GEMBA}}$ and $\text{sentinel}_{\text{MaTESe}}$ converge to those of their discrete counterparts. However, this is a limitation of these sentinels' design and does not imply that the meta-evaluation is fair at higher percentages of human ties.

To better investigate the source of unfairness, in Figure 5.6 (right), we show how the optimal ϵ changes at varying percentages of human ties. As can be seen, continuous metrics' ϵ is dynamically adjusted with heightened sensitivity, contrary to what happens for discrete metrics. For example, in ZH→EN, discrete metrics' ϵ is exactly 0 until the percentage of human ties over all pairs is 39%. Additionally, for MaTESe, it remains constant between 44% and 56%, and between 61% and 68%, and the same happens for GEMBA-MQM between 47% and 51% and between 56% and 68%. In contrast, the values change for all the other metrics in the same intervals, enabling them to better fit the distribution of gold ties found in the test set.

Can we use a held-out set for tie calibration?

We have demonstrated that conducting the tie calibration on the same test set used for the meta-evaluation favors continuous metrics over discrete ones. Nonetheless, this does not necessarily mean using a held-out dataset would ensure a fair meta-evaluation. Indeed, our experiments show that unfairness stems from the different levels of adaptability between continuous and discrete metrics to the distribution of human ties found in the dataset used for tie calibration. Therefore, we expect that using a held-out dataset would still advantage continuous metrics if the distribution of human ties in the held-out set resembled that of the test set, and disadvantage them if such a distribution differed from that of the test set. In both cases, continuous metrics' increased adaptability compared to discrete metrics would impair the fairness of the meta-evaluation. To investigate this further, we compute an 80-20 split of the test set to obtain an evaluation set for tie calibration. Then, we repeatedly sub-sample such an evaluation set to modify its distribution of human ties and compute acc_{eq} on the new test set. The results are shown in Figure 5.7. We observe that the ranking is unstable at varying percentages of human ties, putting continuous metrics at a disadvantage if the proportion of ties in the evaluation set deviates significantly from that in the test set.

5.4 Conclusion

In this chapter, we identified two issues with the current MT meta-evaluation, as conducted at the WMT Metrics Shared Task. We proposed a suite of sentinel metrics designed to highlight these issues and demonstrate their impact on the metrics rankings, revealing that certain metric categories are unfairly advantaged. Indeed, the *None Grouping* and *System Grouping* strategies favor trained metrics over overlap- and LLM-based ones, and the tie calibration algorithm favors continuous metrics over discrete ones when tie calibration is conducted on the same test set used for the meta-evaluation. Finally, we observed a notably high correlation between sentinel metrics and state-of-the-art metrics, raising concerns about their reliability and suggesting that their assessments might be based on spurious correlations present in the training data.

Table 5.2. Segment-level Pearson correlation for the primary submissions to the WMT23 Metrics Shared Task, with sentinel metrics. The language direction is ZH→EN. Starred metrics are reference-free, underlined metrics are baselines, and highlighted metrics are sentinels. Ranks represent clusters of statistical significance and are computed following Freitag et al. (2023), which leverage the PERM-BOTH hypothesis test introduced by Deutsch et al. (2021).

Metric	**No**		**Segment**		**System**	
XCOMET-Ensemble	1	0.650	2	0.421	1	0.610
MetricX-23-QE*	1	0.647	4	0.359	1	0.610
XCOMET-QE-Ensemble*	1	0.647	3	0.380	1	0.612
MetricX-23	2	0.625	3	0.373	2	0.580
sentinel-cand*	3	0.580	11	0.201	2	0.578
sentinel-src*	4	0.540	14	0.000	3	0.561
sescoreX	4	0.536	7	0.295	5	0.505
MaTESe	5	0.511	6	0.325	6	0.441
sentinel-ref	5	0.506	14	0.000	4	0.525
mbr-metricx-qe*	6	0.489	1	0.436	7	0.431
cometoid22-wmt22*	6	0.479	4	0.357	6	0.446
GEMBA-MQM*	7	0.449	1	0.434	9	0.378
Calibri-COMET22-QE*	7	0.443	5	0.355	8	0.411
CometKiwi*	7	0.442	3	0.388	9	0.388
KG-BERTScore*	8	0.430	4	0.369	10	0.374
COMET	9	0.396	4	0.364	12	0.345
Calibri-COMET22	9	0.396	7	0.311	11	0.360
docWMT22CometKiwiDA*	10	0.387	6	0.340	13	0.320
BLEURT-20	10	0.378	4	0.371	13	0.330
MS-COMET-QE-22*	11	0.367	7	0.306	14	0.313
docWMT22CometDA	12	0.353	6	0.327	15	0.291
YiSi-1	13	0.290	6	0.329	16	0.237
BERTscore	14	0.236	7	0.309	17	0.186
prismSrc*	15	0.223	13	0.078	16	0.243
prismRef	16	0.183	6	0.332	18	0.135
embed_llama	17	0.161	12	0.138	18	0.139
mre-score-labse-regular	18	0.145	8	0.251	19	0.123
BLEU	19	0.119	11	0.208	20	0.093
XLsim	19	0.111	10	0.218	21	0.069
f200spBLEU	20	0.108	10	0.220	21	0.077
MEE4	20	0.105	9	0.236	21	0.070
chrF	21	0.063	8	0.263	22	0.020
tokengram_F	22	0.060	8	0.262	23	0.015
Random-sysname*	23	0.018	14	0.019	23	0.002
eBLEU	24	−0.084	10	0.219	24	−0.115

Table 5.3. Segment-level Pearson correlation for the primary submissions to the WMT23 Metrics Shared Task, with sentinel metrics. The language direction is EN→DE. Starred metrics are reference-free, underlined metrics are baselines, and highlighted metrics are sentinels. Ranks represent clusters of statistical significance and are computed following Freitag et al. (2023), which leverage the PERM-BOTH hypothesis test introduced by Deutsch et al. (2021).

Metric	No		Segment		System	
XCOMET-Ensemble	1	0.695	1	0.538	1	0.676
XCOMET-QE-Ensemble*	2	0.679	2	0.507	2	0.658
MetricX-23-QE*	3	0.626	2	0.511	3	0.564
MetricX-23	4	0.585	2	0.507	4	0.547
mbr-metricx-qe*	4	0.571	1	0.543	3	0.551
sentinel-cand*	5	0.561	6	0.396	5	0.522
MaTESe	5	0.554	8	0.330	4	0.526
sescoreX	6	0.519	3	0.459	6	0.502
prismRef	6	0.516	7	0.349	4	0.528
GEMBA-MQM*	6	0.502	3	0.482	7	0.446
BLEURT-20	7	0.484	2	0.492	7	0.455
CometKiwi*	7	0.475	3	0.463	7	0.451
sentinel-src*	8	0.469	12	0.000	6	0.502
sentinel-ref	8	0.464	12	0.000	6	0.492
KG-BERTScore*	8	0.451	4	0.456	8	0.421
docWMT22CometKiwiDA*	9	0.444	5	0.426	9	0.404
cometoid22-wmt22*	9	0.441	2	0.499	9	0.385
Calibri-COMET22-QE*	9	0.441	5	0.432	8	0.414
COMET	9	0.432	2	0.508	10	0.363
prismSrc*	9	0.425	11	0.102	6	0.487
Calibri-COMET22	10	0.413	3	0.477	10	0.370
docWMT22CometDA	11	0.394	3	0.484	11	0.310
YiSi-1	12	0.366	5	0.404	12	0.284
BERTscore	13	0.325	7	0.355	13	0.250
MS-COMET-QE-22*	13	0.310	6	0.400	13	0.241
embed_llama	14	0.250	10	0.242	14	0.180
XLsim	14	0.239	6	0.372	16	0.151
f200spBLEU	14	0.237	7	0.343	14	0.178
chrF	15	0.232	8	0.336	15	0.157
tokengram_F	16	0.227	8	0.340	16	0.153
MEE4	17	0.202	7	0.360	16	0.145
BLEU	17	0.192	9	0.310	17	0.140
mre-score-labse-regular	18	0.111	6	0.376	18	0.087
Random-sysname*	19	0.064	11	0.124	19	−0.015
eBLEU	20	−0.011	8	0.317	19	−0.030

Table 5.4. Segment-level Pearson correlation for the primary submissions to the WMT23 Metrics Shared Task, with sentinel metrics. The language direction is HE→EN. Starred metrics are reference-free, underlined metrics are baselines, and highlighted metrics are sentinels. Ranks represent clusters of statistical significance and are computed following Freitag et al. (2023), which leverage the PERM-BOTH hypothesis test introduced by Deutsch et al. (2021).

Metric	No		Segment		System	
XCOMET-Ensemble	1	0.556	1	0.479	1	0.515
MetricX-23	1	0.548	2	0.441	1	0.509
MetricX-23-QE*	2	0.520	5	0.387	2	0.480
XCOMET-QE-Ensemble*	3	0.498	4	0.397	3	0.458
MaTESe	4	0.459	5	0.373	4	0.408
mbr-metricx-qe*	5	0.411	2	0.448	5	0.362
GEMBA-MQM*	5	0.401	2	0.431	6	0.354
COMET	5	0.401	3	0.421	5	0.367
Calibri-COMET22	5	0.401	4	0.397	5	0.371
YiSi-1	6	0.395	2	0.439	6	0.348
Calibri-COMET22-QE*	6	0.395	6	0.354	5	0.369
CometKiwi*	7	0.387	5	0.375	6	0.353
sescoreX	7	0.385	5	0.370	6	0.352
KG-BERTScore*	8	0.382	5	0.375	7	0.347
BLEURT-20	8	0.382	3	0.418	7	0.344
mre-score-labse-regular	8	0.378	4	0.407	8	0.335
cometoid22-wmt22*	9	0.365	7	0.309	7	0.346
docWMT22CometDA	10	0.339	5	0.379	9	0.294
sentinel-cand*	10	0.339	11	0.104	7	0.343
BERTscore	10	0.335	4	0.412	9	0.293
sentinel-src*	10	0.334	13	0.000	7	0.336
prismRef	11	0.319	3	0.428	10	0.276
sentinel-ref	11	0.301	13	0.000	9	0.299
MS-COMET-QE-22*	12	0.295	9	0.252	10	0.274
docWMT22CometKiwiDA*	12	0.286	7	0.324	11	0.234
MEE4	13	0.256	8	0.291	11	0.222
XLsim	14	0.233	7	0.314	12	0.198
f200spBLEU	14	0.230	8	0.287	12	0.195
tokengram_F	14	0.226	7	0.311	13	0.184
chrF	15	0.221	7	0.308	14	0.179
BLEU	15	0.220	9	0.260	13	0.189
embed_llama	15	0.215	10	0.188	13	0.187
prismSrc*	16	0.140	11	0.100	15	0.150
eBLEU	16	0.131	8	0.280	16	0.104
Random-sysname*	17	0.041	12	0.057	17	0.001

Table 5.5. Percentage of tied pairs in the MQM data released over different years at the Metrics Shared Task (or by Freitag et al. (2021a), for 2020), and regarding different translation directions.

	2020	2021	2022	2023
EN→DE	15.14	44.62	53.35	23.11
ZH→EN	17.01	30.31	41.55	24.03
EN→RU	–	53.24	44.42	–
HE→EN	–	–	–	42.84

Table 5.6. p_t is the probability of removing a tied human pair, and p_n is that of removing a non-tied human pair. The considered test set is WMT23. For each sub-table, which reports the statistics of an individual language pair, each column indicates a pair (p_t, p_n), i.e., it represents a sub-sample of the test set in which tied and non-tied pairs have been removed with probabilities p_t and p_n. The third row contains the percentage of tied human pairs over all pairs, as a result of the sub-sampling, and the last row contains the total number of pairs remaining in the test set after the sub-sampling.

ZH→EN													
p_t	1.00	0.65	0.30	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
p_n	0.00	0.00	0.00	0.00	0.20	0.40	0.50	0.60	0.65	0.70	0.75	0.80	0.85
%	0	10	18	24	28	35	39	44	47	51	56	61	68
#	93890	104304	114664	123585	104888	85969	76522	67237	62624	57948	53110	48491	43730
EN→DE													
p_t	1.00	0.65	0.30	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
p_n	0.00	0.00	0.00	0.00	0.20	0.40	0.50	0.60	0.65	0.7	0.75	0.80	0.85
%	0	10	17	23	27	33	38	43	46	50	54	60	67
#	23343	25803	28236	30360	25694	21021	18689	16353	15184	14014	12899	11698	10493
HE→EN													
p_t	1.0	0.90	0.80	0.65	0.50	0.35	0.20	0.00	0.00	0.00	0.00	0.00	0.00
p_n	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.20	0.40	0.55	0.65	0.75
%	0	7	13	21	27	33	38	43	48	56	62	68	75
#	36561	39254	42038	46202	50272	54435	58516	63960	56679	49315	43918	40145	36530

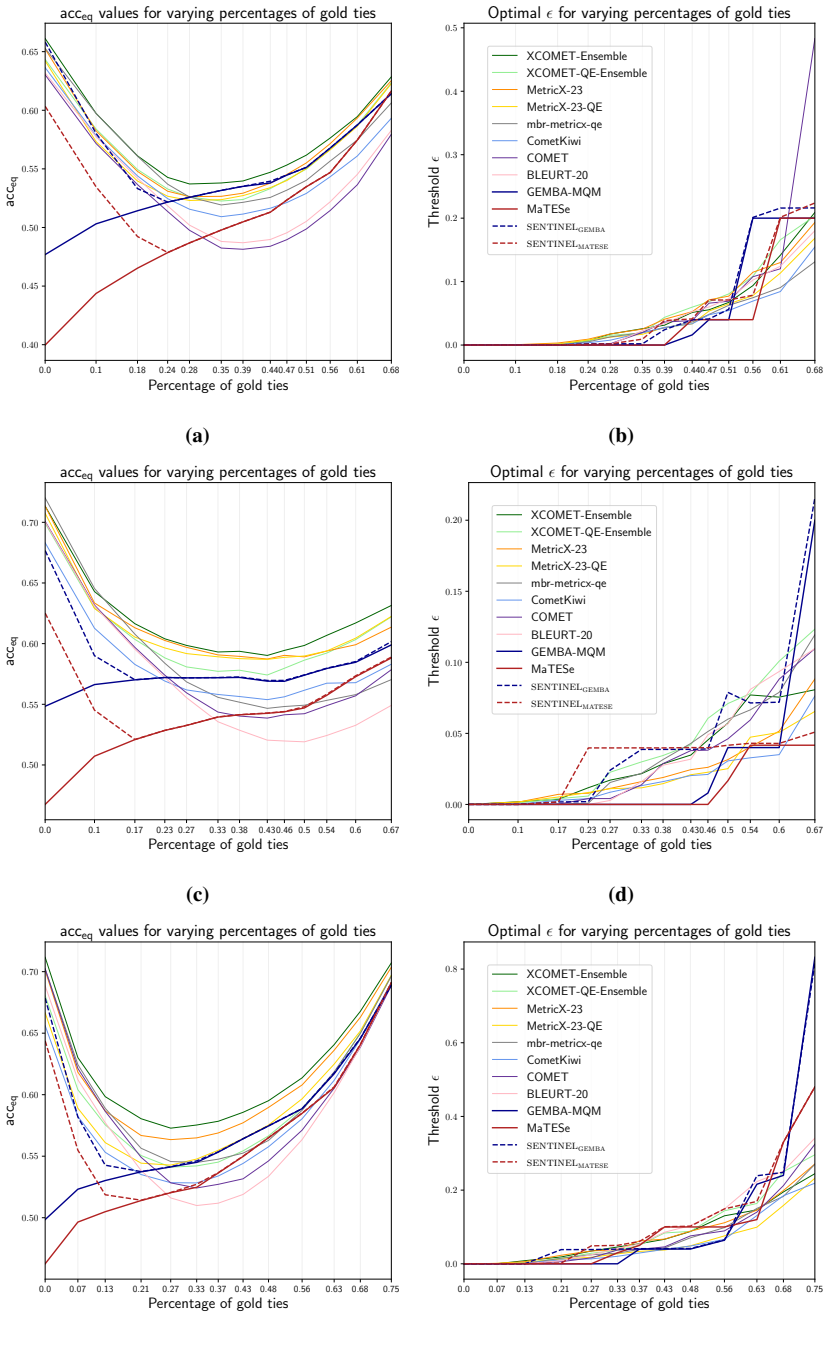
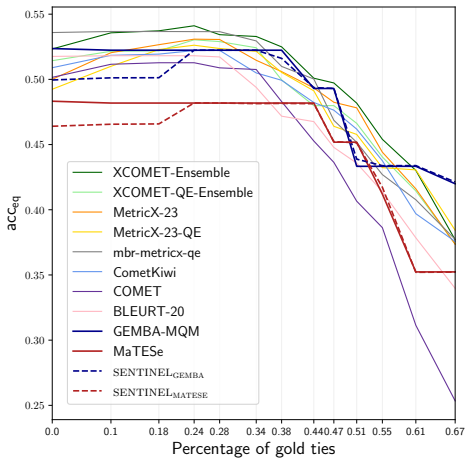
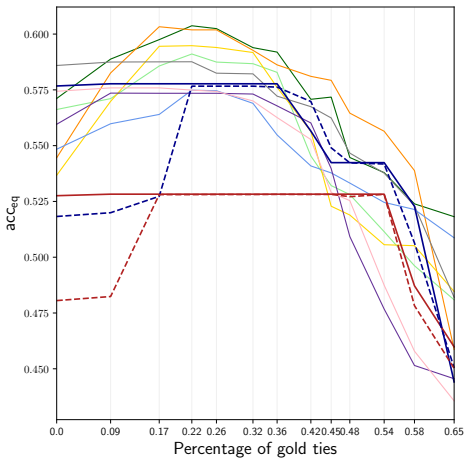


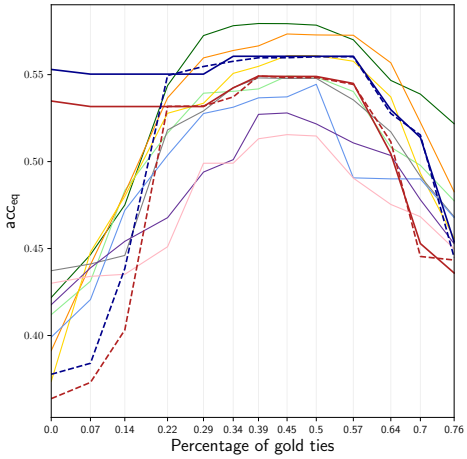
Figure 5.6. acc_{eq} (5.6a, 5.6c, 5.6e) and optimal ϵ (5.6b, 5.6d, 5.6f) of the considered metrics for varying percentages of human ties in the test dataset. 0.24, 0.23, and 0.43 are the percentage of human ties in the ZH→EN (5.6a, 5.6b), EN→DE (5.6c, 5.6d), and HE→EN (5.6e, 5.6f) datasets, respectively, obtained when p_t and p_n are both 0. For each metric, the minimum ϵ is the optimal ϵ at 0% of human ties, and the maximum is the optimal ϵ at 100%. For each percentage of human ties, we use 5 different seeds to sub-sample the test data and average results.



(a) The language pair is ZH→EN. The percentage of human ties in the 80% split of the test set is 24%.



(b) The language pair is EN→DE. The percentage of human ties in the 80% split of the test set is 23%.



(c) The language pair is HE→EN. The percentage of human ties in the 80% split of the test set is 42%.

Figure 5.7. acc$_{eq}$ of the considered metrics when tie calibration is conducted on a held-out set, derived as a 20% split of the test set, and repeatedly sub-sampled to modify its percentage of tied scores. The x-axis represents the percentage of ties in the held-out set, while the y-axis represents the acc$_{eq}$, as computed on the remaining 80% of the test set. For each percentage of human ties, we use 5 different seeds to sub-sample the held-out set and report averaged results.

Chapter 6

Estimating Machine Translation Difficulty

Previous chapters discussed how we contributed to improving MT evaluation by enhancing the evaluation methods: we developed interpretable evaluation techniques (Chapter 3), designed interpretable meta-evaluation strategies (Chapter 4), and introduced the concept of sentinel metrics to identify and address issues in the meta-evaluation (Chapter 5). In this chapter, we shift our focus to improving MT evaluation by refining the selection of test data used in MT benchmarks. Specifically, we introduce the task of Translation Difficulty Estimation, which aims to predict the translation difficulty of a given source text in advance and can be used to construct more challenging MT benchmarks.

6.1 Why Translation Difficulty?

Not all data samples are equal: Machine learning models may struggle with some samples more than others. The ability to automatically assess sample difficulty is indispensable at various stages of model development. For example, during training, organizing samples from the easiest to the hardest, known as Curriculum Learning, improves both performance and training efficiency (Bengio et al., 2009; Wang et al., 2022; Soviany et al., 2022). Even during inference, computational costs can be reduced by early-exiting on easy examples (Teerapittayanon et al., 2016; Schwartz et al., 2020).

Evaluating models also benefits from estimates of sample difficulty, as too easy or

too difficult benchmarks fail to effectively differentiate between models (Lalor et al., 2018; Rodriguez et al., 2021). This issue is particularly relevant in Machine Translation, with recent state-of-the-art models obtaining near-perfect scores and performing close to the human level (Kocmi et al., 2024a; Proietti et al., 2025a). With easy test sets, practitioners might struggle to differentiate between top-performing models and assess whether there is headroom for further model improvement. Additionally, while the MT Test Suites subtask of WMT (Kocmi et al., 2024a) targets specific complex translation phenomena, no systematic investigation of the broader concept of general translation difficulty has been carried out.

To address this gap, we explore the notion of sample difficulty in Machine Translation. First, we propose a definition of translation difficulty and formally introduce Translation Difficulty Estimation as a novel task, where the source text's difficulty is automatically predicted. We then present Difficulty Estimation Correlation (DEC), a measure designed to evaluate the performance of difficulty estimation methods. Finally, we test baselines and newly proposed approaches to difficulty estimation and validate their practical utility in the downstream task of creating a challenging benchmark, which involves automatically selecting subsets of challenging samples from a large corpus.

We find that approaches such as word rarity, syntactic complexity, or even LLM-as-a-Judge underperform dedicated solutions in capturing translation difficulty. Specifically, we show that sentinel-src – a model trained to predict the expected translation quality of a given text based solely on the source text itself (Chapter 5) – outperforms other methods at estimating translation difficulty. Therefore, we train two improved versions, called sentinel-src-24 and sentinel-src-25, and release them publicly.[1]

6.2 Related Work

Previous works can be divided into two categories depending on whether their focus is on human or Machine Translation difficulty.

Human translation difficulty. The earliest works (Fang, 1959; Hale and Campbell, 2002) attempted to connect general text complexity to translation difficulty for humans. A more

[1]Models: hf.co/collections/Prosho/translation-difficulty-estimators-6816665c008e1d22426eb6c4.
Code: github.com/zouharvi/translation-difficulty-estimation.

modern investigation by Mishra et al. (2013) framed human translation difficulty as the time needed to translate a sentence, and estimated it using data on translators' eye movements. They used text length, word polysemy degree, and syntactic complexity as predictors of translation difficulty. Vanroy et al. (2019) examined the correlation between error count, word translation entropy, and syntactic equivalence with translation duration, gaze, and other proxies for human translation difficulty. More recently, Lim et al. (2023, 2024) used word alignment distributions and decoder perplexity to predict human translation difficulty.

Machine translation difficulty. To implement a Curriculum Learning training schedule, Kocmi and Bojar (2017) estimated sample difficulty based on sentence length, word rarity, and the number of coordinating conjunctions in the text. Similarly, Platanios et al. (2019) used sentence length and rarity as proxies for difficulty. Beyond these linguistically-motivated criteria, Zhang et al. (2018) and Liu et al. (2020a) predicted translation difficulty using the confidence and other intrinsics of the translation model in generating the text. Almeida (2017) treated difficulty estimation as a binary classification task, although they also used features from the target text, making it closer to quality estimation. Zhan et al. (2021b) used an artificial crowd-based approach that leverages automatic metrics and discovered that long segments, low-frequency words, and proper nouns are the most challenging to machine translate. Finally, Zhan et al. (2021a) estimated a text's difficulty using the embedding similarity between its tokens and those of its translations.

Closer to our work, Don-Yehiya et al. (2022) defined the PreQuEL task as predicting the quality of a given text's translation before the translation is generated. However, they adopted the evaluation of the WMT 2020 Quality Estimation Shared Task (Specia et al., 2020), which was designed for quality estimation rather than for assessing difficulty estimators. Furthermore, their test set included only two language directions, with all translations produced by the same MT model. Additionally, they did not explore the broader space of difficulty estimators or investigate their use in constructing challenging benchmarks.

In contrast, we define translation difficulty estimation as a distinct task with a dedicated evaluation metric. Moreover, we benchmark a wide range of difficulty estimation approaches using test sets that span 11 language directions, with 11 to 19 translations per segment across language pairs, produced by both MT models and human translators. As a

result, our work constitutes the first extensive evaluation of translation difficulty estimators, establishing a new state of the art for the task.

6.3 The Difficulty Estimation Task

The difficulty of translating a given text can depend on multiple factors. A text may be challenging, for example, due to its length, syntactic complexity, idiomatic language, or the presence of rare or specialized vocabulary. Some aspects that affect translation difficulty may even depend on the translation direction, meaning that the same source text might be more difficult to translate into one language than into another. Moreover, translation difficulty might not be uniform across translators, as it can vary with the translator's cultural background and linguistic familiarity – in the case of human translators – or based on factors such as the number of parameters, training data, and model architecture – in the case of Machine Translation models.

Given these considerations, we avoid defining translation difficulty in absolute terms, as such a definition may not generalize well. Instead, we define difficulty relative to a given target language and to the ability of a particular translator, whether human or automatic. More specifically, given a text x, a model[2] m, and a target language l, we assign to x a difficulty score $d_{m,l}(x)$ equal to the quality score assigned to a translation of x into language l produced by m. Lower scores indicate a lower translation quality and, therefore, greater difficulty associated with the source text.

As an example, suppose we have two texts, x_1 and x_2, and their respective translations t_1 and t_2 into language l, both produced by model m. A human rater evaluates these translations on a scale from 1 to 100, assigning a score of 60 to the first and 90 to the second. Then, $d_{m,l}(x_1) = 60$ and $d_{m,l}(x_2) = 90$. Since $d_{m,l}(x_1) < d_{m,l}(x_2)$, then x_1 is more difficult to translate into l than x_2, for model m. Importantly, the lower the score d, the higher the difficulty and vice versa.

Task Definition. Given a source text x, a model m, and a target language l, Difficulty Estimation is the task of automatically predicting $d_{m,l}(x)$. Different from Quality Estimation, difficulty estimation models do not have access to the translations whose quality is being

[2]For brevity, we use "model" to refer both to human translators and automatic models.

estimated. Indeed, difficulty estimation can be seen as the task of estimating the *expected* quality of a given text's translation.

Evaluation. We evaluate difficulty estimation methods according to their ability to rank texts based on difficulty. Consider a collection of texts $\mathcal{X} = x_1, x_2, \ldots, x_N$, a collection of target languages $\mathcal{L} = l_1, l_2, \ldots, l_L$, and a collection of models translating into language l: $\mathcal{M}_l = m_1, m_2, \ldots, m_{M_l}$. Let us also define the vector of ground-truth difficulty scores for model m and language l as $D_{m,l} = d_{m,l}(x_1), d_{m,l}(x_2), \ldots, d_{m,l}(x_N)$ and the corresponding predictions of a difficulty estimation method as $\hat{D}_{m,l} = \hat{d}_{m,l}(x_1), \hat{d}_{m,l}(x_2), \ldots, \hat{d}_{m,l}(x_n)$. We measure the translation **Difficulty Estimation Correlation (DEC)** by averaging the Kendall's rank correlation coefficients τ_b across models and languages:

$$\text{DEC} = \frac{1}{|\mathcal{L}|} \sum_{l \in \mathcal{L}} \frac{1}{|\mathcal{M}_l|} \sum_{m \in \mathcal{M}_l} \tau_b(\hat{D}_{m,l}, D_{m,l}). \tag{6.1}$$

We refer the reader to Section 2.2 for further details on how Kendall's correlation coefficient τ_b is computed.

Contrasting DEC with standard MT meta-evaluation strategies. The evaluation approach used in Formula 6.1, termed "Group-by-System" by Deutsch et al. (2023) and "System Grouping" by us in Chapter 5, makes DEC fundamentally different from other meta-evaluation strategies used in MT evaluation and Quality Estimation, which, as we recommended in Chapter 5, should rely on the Group-by-Item method instead ("Segment Grouping"). Group-by-Item calculates the correlation between assessments assigned to different translations of the **same source text** (i.e., the same evaluation item), and then averages these correlations across all source texts. The primary benefit of this method is that it mitigates spurious correlations between source text features (e.g., length) and translation quality judgments, as discussed in Section 5.2.

However, we argue that these features are precisely what define a text's translation difficulty. Since our goal is to measure translation difficulty, we define DEC using Group-by-System. This method computes the correlation between human and metric assessments for translations of **different source texts** that were produced by the same MT system, and then averages the correlations across MT systems. By holding the MT system constant, DEC

directly measures the ability of difficulty estimators to identify which source texts were more challenging for that MT system to translate.

This distinction is crucial. Indeed, using Group-by-System to meta-evaluate standard MT or quality estimation metrics would favor those that predict source difficulty rather than purely translation quality. Conversely, using Group-by-Item to evaluate difficulty estimators would be inappropriate, as these estimators assign the same score to all translations of a given source.

6.4 Methods for Difficulty Estimation

In this section, we describe several difficulty estimation methods. We include both common and novel approaches. Specifically, we categorize difficulty estimators into four groups: heuristic-based, learned, LLM-as-a-Judge, and crowd-based.

6.4.1 Heuristic-based estimators

We refer to estimators as heuristic-based if they rely on simple text features. This category includes estimators previously shown to correlate with other measures of difficulty (Mishra et al., 2013; Kocmi and Bojar, 2017; Araghi and Palangkaraya, 2024):

- **Text length** is the number of tokens in a text.[3]
- **Word rarity** is the negative average of the frequencies (estimated from a reference corpus) of the words in a text.[4]
- **Syntactic complexity** is approximated as the height of the dependency tree associated with a text.[5]

[3] We tokenize the text using `spaCy`. Specifically, we use language-specific pipelines: `en_core_web_sm` for English, `ja_core_news_sm` for Japanese, and `spacy_udpipe` for Czech. Text length refers to the number of tokens in a given text.

[4] we compute word frequencies using the `wordfreq` Python library (Speer, 2022).

[5] Similar to Text length, we obtain dependency trees using `spaCy`.

6.4.2 Learned estimators

Learned Machine Translation metrics are often trained to predict the quality of a translation given its source text and, optionally, a reference translation (Rei et al., 2020; Guerreiro et al., 2024; Juraska et al., 2023, 2024). Similarly, neural models can be trained to predict the difficulty of a text. Previous research has explored training similar models for related purposes:

- **PreCOMET** is a suite of source-based regressors based on XLM-RoBERTa (Conneau et al., 2020) that predict the usefulness of a sample for evaluation (Zouhar et al., 2025). Specifically, $\text{PreCOMET}_{\text{diversity}}$ prioritizes samples likely to elicit diverse Machine Translation outputs, while $\text{PreCOMET}_{\text{difficulty}}$ estimates difficulty as defined by item response theory (Santor and Ramsay, 1998).

- **sentinel-src** metric is a regression model based on XLM-RoBERTa. As discussed in Section 5.2.2, sentinel-src estimates translation quality from the source alone – i.e., without accessing the candidate translation. As a result, it learns spurious correlations between features of the source texts and translation quality scores. Such spurious correlations represent the translation difficulty of a given text.

We train two novel sentinel metrics: sentinel-src-24 and sentinel-src-25. They follow the same architecture and training pipeline used for the sentinel-src model introduced in Section 5.2.2. Both models are based on XLM-RoBERTa large as the backbone encoder, followed by a multi-layer feedforward network on top of the `[CLS]` token, and are trained to minimize the Mean Squared Error (MSE) between predicted and human scalar scores. We adopt the same two-stage training approach as sentinel-src. In the first stage, the model is trained on DA annotations, while the second stage fine-tunes it using MQM data. The key differences between our models and sentinel-src lie in the training data recipe used at each stage:

- **Stage 1: DA-based warm-up.** For sentinel-src-24, we extend the DA training data of sentinel-src by including annotations from WMT 21 (Wenzek et al., 2021), as well as DA+SQM annotations from WMT 22 (Kocmi et al., 2022) and WMT 23 (Kocmi et al., 2023), while sentinel-src-25 further includes the ESA annotations from WMT

24. For both model versions, we also incorporate MLQE-PE data (Fomicheva et al., 2022) in the training set for this stage.

- **Stage 2: MQM-based fine-tuning.** In this phase, we expand the MQM training set by adding MQM annotations from WMT 23 (Freitag et al., 2023). Unlike the sentinel-src training pipeline, we do not average the assessments returned by different raters. Instead, we include all available annotations as individual training instances. This applies to WMT 20 and WMT 22 MQM datasets, which include three human scores per translation (Freitag et al., 2021a; Riley et al., 2024). Lastly, the training set of sentinel-src-25 also includes MQM annotations from WMT 24.

In training these models, we treat each pair consisting of a source text segment and its associated human score as an independent training instance. In the mentioned data sets, human scores were assigned to translations, not source texts. Therefore, since there are multiple translations of the same source text, multiple, potentially different human assessments exist for each source text. Our training strategy does not combine these scores in any way. Instead, we include them all as individual training samples for both the DA-based warm-up and the MQM fine-tuning.

Finally, we highlight that our training hyperparameters match those used to train sentinel-src, which are reported in Table 5.1.

6.4.3 LLM-as-a-Judge

LLM-as-a-Judge approaches have seen wide adoption across a range of applications (Zheng et al., 2023; Bavaresco et al., 2025). We investigate the effectiveness of the LLM-as-a-Judge paradigm for the task of difficulty estimation, using GPT-4o (OpenAI et al., 2024) and CommandA (Cohere Team, 2025). We prompt these models to determine the proficiency level required to translate a given text, optionally providing information about the target language, and return a scalar score between 0 and 120 indicating the difficulty level of the given text. We report the prompts used in Table 6.1.

Table 6.1. Prompt used to estimate translation difficulty using LLM-as-a-judge approaches. The blue text in bold is optional, as it is only present in the version of the prompt that includes information about the target language.

LLM-as-a-judge prompt

```
You are given a source text. Your goal is to determine the
approximate proficiency level required to translate this
text into {target_language}, based on a detailed analysis
of its complexity. The final result should be reported
as a single numeric score on a scale of 0 to 120, where
higher numbers correspond to a higher difficulty (i.e.,
more advanced language proficiency requirements). You
should also relate this numeric score to commonly recognized
proficiency levels (e.g., A1, A2, B1, B2, C1, C2). Here
is the expected mapping: 0-20 for A1 (Beginner); 21-40 for
A2 (Elementary); 41-60 for B1 (Intermediate); 61-80 for B2
(Upper Intermediate); 81-100 for C1 (Advanced); 101-120 for
C2 (Mastery).
Instructions: First, examine the text to identify features
affecting the translation into {target_language}, which
affect reading difficulty, including complexity of
vocabulary, grammar, semantic density, and any specialized
knowledge required. Then, provide a brief explanation of
your reasoning for each major factor. Consider whether
the text includes domain-specific terminology, cultural
references, idiomatic expressions, or advanced grammatical
constructions. Finally, assign a numeric score from 0 to
120 and map that score to one of the CEFR levels. Conclude
with a final statement that clearly states your numeric
score and the corresponding proficiency level surrounded by
triple square brackets, for example [[[86, C1 (Advanced)]]].
Analyze following text: {src}
```

6.4.4 Crowd-based Estimators

The methods discussed so far estimate translation difficulty based solely on the source text, and optionally, the target language. However, having defined translation difficulty as the expected quality of a model's translations (Section 6.3), we now introduce difficulty estimators that more closely mimic this definition.

Artificial Crowd. Artificial crowd-based methods first translate a source text and then use reference-less MT metrics to estimate the quality of the resulting translations. Specifically, we translate the source texts from the test set using a diverse set of models to ensure variety in architecture and size:

- **NLLB-moe-54B** is a sparsely-gated mixture-of-experts encoder-decoder translation model (NLLB Team, 2022).
- **Gemma-3-27B-IT** is a multimodal instruction-tuned LLM from the Gemma family (Gemma Team, 2025).
- **Qwen2.5-72B-IT** is the largest instruction-tuned LLM from the Qwen2.5 family (Qwen Team, 2025).
- **CommandA** is a 111B-parameter LLM for real-world enterprise use cases (Cohere Team, 2025).

Then, we evaluate the generated translations using two state-of-the-art, reference-less MT metrics: XCOMET-QE-XXL (Guerreiro et al., 2024) and MetricX-24-Hybrid-QE-XXL (Juraska et al., 2024), hereafter referred to as XCOMET and MetricX, respectively. The final difficulty score for each source text is the average quality score assigned to its translations by one of these metrics. This approach is inspired by the artificial crowd methods for efficient subset selection proposed by Zouhar et al. (2025).

True Crowd. To establish a performance upper bound for Artificial Crowd estimators, we also define True Crowd estimators. Unlike Artificial Crowd, True Crowd estimators use XCOMET and MetricX to score the translations produced by the actual systems whose difficulty we aim to measure – namely, the translations that constitute the WMT24 test set used in our experiments.

Table 6.2. Statistics of the test set released at the WMT 2024 General Machine Translation Shared Task (Kocmi et al., 2024a). "#Source texts" indicates the number of source texts in the test set, and "#Translations" indicates the number of available translations for each source text.

	EN→ES	EN→HI	EN→IS	EN→JA	EN→RU	EN→UK	EN→ZH	EN→CS	CS→UK
#Source texts	634	634	634	634	634	634	634	634	1954
#Translations	14	11	11	13	14	11	13	16	12

Table 6.3. Statistics of the test set released at the WMT 2024 Metrics Shared Task (Freitag et al., 2024). "#Source texts" indicates the number of source texts in the test set, and "#Translations" indicates the number of available translations for each source text.

	EN→DE	EN→ES	JA→ZH
#Source texts	486	622	559
#Translations	19	15	15

Since they rely on the ground-truth translations, True Crowd estimators are not proper difficulty estimators and are effectively equivalent to quality estimators. We employ True Crowd estimators solely to report an upper bound on the performance of Artificial Crowd estimators.

6.5 Experiments

We benchmark the estimators using DEC (Equation 6.1).

6.5.1 Experimental Setup

We measure DEC on the test sets released at the WMT 2024 General MT and Metrics shared tasks (Kocmi et al., 2024a; Freitag et al., 2024). These test sets include a selection of source texts translated into multiple languages by automatic models and human translators. Each translation is paired with quality annotations produced by human annotators following either the Error Span Annotation (ESA, Kocmi et al., 2024b) or the MQM annotation protocols. We report statistics about the considered test sets in Tables 6.2 and 6.3.

We test all methods listed in 6.4 except sentinel-src-25 because its training set includes

the WMT24 test data used in this evaluation.

6.5.2 Results

We present the results in Tables 6.4 and 6.5, with methods organized by category as described in Section 6.4. We also include three distinct oracles to provide the reader with upper-bound performance values. Oracle methods adopt the true human judgments used to derive difficulty scores. We consider three oracles that differ in the type of information they have access to:

- **Oracle (source text + target language + target translation)** assigns to each source text x the true $d_{m,l}(x)$, for each model m and target language l.
- **Oracle (source text + target language)** estimates the difficulty of x by averaging the true $d_{m,l}(x)$ across all models ($\forall m \in \mathcal{M}_l$), meaning that its estimates do not vary across models, but only across target languages.
- **Oracle (source text only)** averages the true $d_{m,l}(x)$ across both models and target languages, assigning the same score to each source text regardless of target language or translator.

Heuristic-based and Learned methods. These estimators base their predictions only on the input text. Consequently, the difficulty scores they assign to each text are the same across all target languages and models.[6] Within this group, all learned estimators outperform the heuristic-based ones. Furthermore, sentinel-src-24 achieves the highest difficulty estimation correlation overall, also higher than sentinel-src introduced in Chapter 5, highlighting the effectiveness of our re-training.

LLM-as-a-Judge. LLM judges are optionally provided with the target language. For both models, the target language information improves performance. This is especially true for CommandA, where the target language information leads to a 0.032 and 0.006 point increase in correlation in Tables 6.4 and 6.5, respectively. However, the overall LLM judges' performance is poor, with scores even lower than the much simpler Text Length heuristic.

[6] I.e., $\forall m_1, m_2 \in \mathcal{M}, l_1, l_2 \in \mathcal{L} : \hat{d}_{m_1,l_1}(x) = \hat{d}_{m_2,l_2}(x)$.

Crowd-based estimators. As expected, the True Crowd methods, which utilize ground-truth translations and thus serve as an upper bound, yield the highest correlation. Given that their performance depends solely on the reference-less metrics employed, this result demonstrates that XCOMET outperforms MetricX on this task.

Instead, Artificial Crowd methods' performance is close to that of sentinel-src-24. However, Artificial Crowd approaches are considerably more resource-intensive than learned methods, as they require both the translation of source texts and a subsequent quality estimation step.

6.5.3 Discussion

Through our evaluation, we find that:

- Heuristic-based estimators, commonly used in previous works, are outperformed by most methods.
- LLM-as-a-Judge approaches are also surpassed by most methods, including the much simpler Text Length heuristic.
- Learned methods – i.e., models explicitly trained to predict text difficulty – outperform other approaches. Artificial Crowd estimators are close, but are considerably more expensive to operate.
- **sentinel-src-24 sets a state-of-the-art in difficulty estimation**, outperforming all other evaluated estimators.[7]

6.5.4 Comparing Human and Machine Translation Difficulty

We now examine whether the texts that models find difficult to translate are also challenging for humans. To do this, we use the difficulty scores $d_{m,l}$ assigned to each source text in the WMT 24 test sets, varying m across human translators and MT models. We measure the Kendall's τ_b between the scores of all pairs of translators, averaging them across all language directions. For consistency, we restrict the analysis to the models and human

[7]We intentionally exclude True Crowd estimators, which, as discussed in Section 6.4.4, are not genuine difficulty estimators and instead serve only as upper bounds for Artificial Crowd.

translators for which we have annotated translations for all language directions, namely, one human translator and the following four models: Unbabel-Tower70B (Alves et al., 2024), IOL-Research (Zhang, 2024), Claude-3.5, and GPT-4 (OpenAI et al., 2024).

The results in Table 6.6 show that the correlations with the human translator are consistently lower (ranging 0.109 to 0.137) than those between Machine Translation models (ranging 0.151 to 0.221). This suggests that human translators may perceive translation difficulty differently from automatic models. Notably, the highest agreement is observed between GPT-4 and Claude-3.5, which might be due to both models being general-purpose LLMs, unlike Unbabel-Tower70B and IOL-Research, which were explicitly trained for Machine Translation.

6.6 Creating Difficult Benchmarks

In this section, we use the top-performing difficulty estimators to create difficult Machine Translation benchmarks. First, we show that the test set employed at the WMT 2024 General MT shared task (Kocmi et al., 2024a) is too easy for current MT models. Then, we define the task of selecting a difficult subset of samples from a given dataset, and evaluate our estimators.

WMT24 is easy. In Figure 6.1, we report the distribution of scores assigned by human annotators to the translations of the WMT 2024 General MT shared task (Kocmi et al., 2024a). Notably, the best model for each language almost always attains 85 to 100 ESA points, and so does at least one system for each input text. This is particularly concerning for the English-to-Spanish translation direction, where top systems made barely any errors. These findings highlight the need to create more difficult benchmarks. To do this, we use difficulty estimators to sample difficult texts from among a larger collection.

6.6.1 Setup

Given a large set of source texts $\mathcal{X}$, we aim to extract a subset $\mathcal{X}' \subseteq \mathcal{X}$ of maximum difficulty of size $|\mathcal{X}'| = B$. In practice, we select B texts with the highest difficulty as

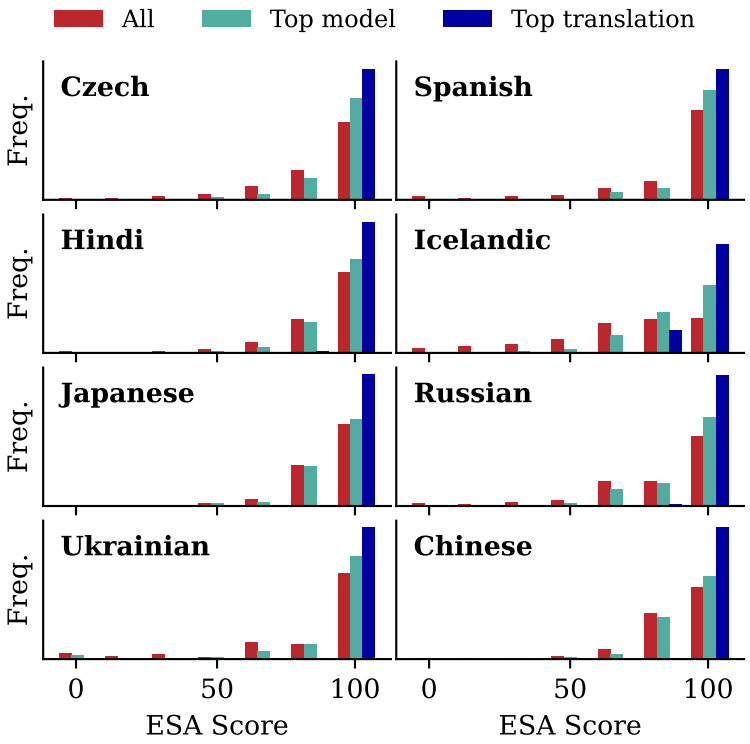


Figure 6.1. Distribution of human scores assigned to the translations of the texts included in the WMT 2024 test set (Kocmi et al., 2024a). We report the scores of all models (ALL), the scores of the top-performing model for each language (Top model), and the scores of the best translation for each input text (Top translation). The chosen bin width is 15 ESA points.

determined by the top-performing difficulty estimators.[8]

We again rely on the WMT 24 test set, which we use as $\mathcal{X}$. Specifically, we employ the ESA-annotated WMT 24 test set, which includes English source texts translated into Chinese, Czech, Hindi, Icelandic, Japanese, Russian, Spanish, and Ukrainian, and Czech source texts translated into Ukrainian. We wish to highlight that since we select B source texts from $\mathcal{X}$, any subset $\mathcal{X}' \subseteq \mathcal{X}$ necessarily contains only English and Czech source texts; target languages are considered solely for evaluation purposes.

Task definition. We assign a single difficulty score $\hat{d}(x)$ to each sample $x \in \mathcal{X}$. For source text-only difficulty estimators, such as heuristics and learned methods, this is straightforward,

[8]We exclude True Crowd estimators because they are not proper difficulty estimators and, as discussed in Section 6.4.4, were used only to offer an upper bound to the performance of Artificial Crowd methods.

as they rely only on the given source text. Instead, for Artificial Crowd methods, we assign to each text x the average quality score of its translations, estimated using XCOMET, averaging across both the MT models employed and the target languages. Finally, we construct $\mathcal{X}'$ by selecting the B most difficult source texts.

Evaluation. One goal of constructing a difficult benchmark is to identify samples where contemporary models still struggle, in order to expose their shortcomings and guide improvements in future iterations. Therefore, we evaluate the usefulness of difficulty estimators based on the average human score obtained by the models' translations on the test set. Specifically, suppose models achieve lower average human scores when measuring the quality of their translations of the subsampled data. In that case, the subsampled data is considered more difficult for such models. As additional information, we also report the proportion of "perfect" outputs (i.e., those that received a full score of 100/100 ESA points from human annotators) among the translations of the sources in the subsampled test set. Formally, given $\mathcal{X}' \subseteq \mathcal{X}$ and a set of models $\mathcal{M}_l$, we report for each subset $\mathcal{X}' \subseteq \mathcal{X}$ with $|\mathcal{X}'| = B$:

$$\text{AvgScore} = \frac{1}{B \cdot |\mathcal{M}_l|} \sum_{\substack{x \in \mathcal{X}' \\ m \in \mathcal{M}_l}} d_{m,l}(x) \tag{6.2}$$

$$\%\text{Perfect} = \frac{1}{B \cdot |\mathcal{M}_l|} \sum_{\substack{x \in \mathcal{X}' \\ m \in \mathcal{M}_l}} \mathbf{1}[d_{m,l}(x) = 100\%] \tag{6.3}$$

6.6.2 Results

We extract several $\mathcal{X}' \subset \mathcal{X}$ by varying the size of the subsample, and report the curves of the Average Score and %Perfect measures in Figure 6.2. First, we wish to highlight that the oracles serve as a performance upper bound only in terms of Average Score, and not in terms of %Perfect, because they are designed to select the sources with the lowest average score, rather than the lowest %Perfect. In this respect, oracle-src selects the sources with the lowest average difficulty score across models and target languages; instead, oracle-tgt selects a different source text for each target language, averaging difficulty scores only across MT models. As we can see from Figure 6.2, text length heuristics and LLM-as-a-Judge-based methods show very close performance to random subset selection, especially in terms of

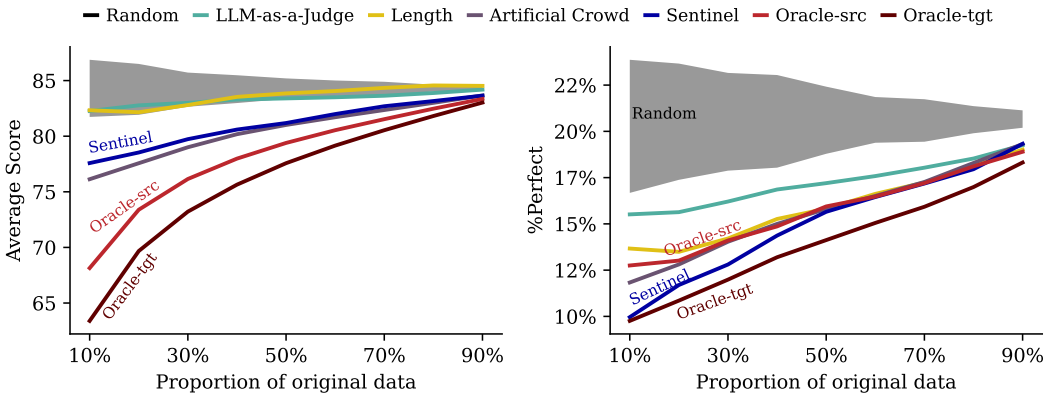


Figure 6.2. Average score and proportion of perfect outputs from the translations of sub-sampled sources. Lower values are better. Results are averaged across all the WMT24 language pairs that we used to simulate the subset selection (see Section 6.6.1). Random selection shows 99% confidence t-test interval from 10 runs.

Average Score. Instead, sentinel-src-24 and Artificial Crowd perform closer to the oracles, even surpassing oracle-src in terms of %Perfect.

Additionally, we also report detailed quantitative results for the scenario where we subsample 25% of the test set in Table 6.7. sentinel-src-24 achieves a lower AvgScore and %Perfect than all other estimators, substantially lower than random selection. Only in AvgScore in the ESA-annotated data Artificial Crowd outperforms sentinel-src-24.

Based on these results, sentinel-src-24 and Artificial Crowd come out as the two best difficulty estimators for difficult benchmark creation. However, we wish to highlight that using Artificial Crowd methods implicitly assumes that the generated test data will be used to evaluate models other than those involved in the difficulty estimation. In fact, because Artificial Crowd relies on an intermediate translation step, it would bias the resulting test set against the models used in its construction. Furthermore, despite its performance being close to sentinel-src-24, Artificial Crowd is significantly more computationally expensive. Thus, we recommend sentinel-src-24 as the go-to difficulty estimator for creating difficult benchmarks.

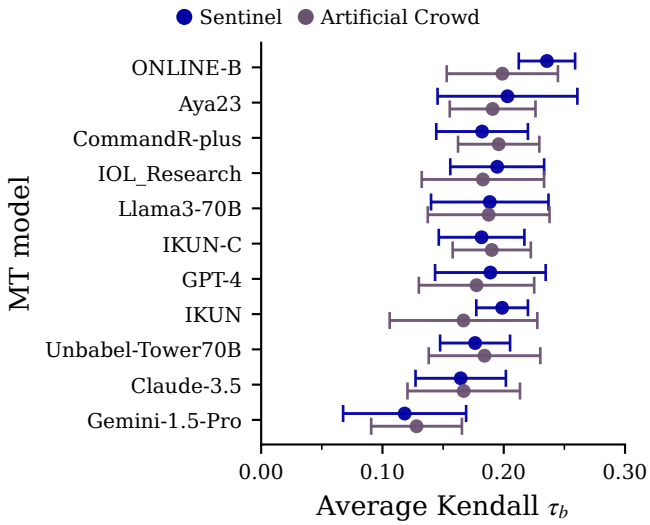


Figure 6.3. Average Kendall τ_b between difficulty estimators and the human judgments assigned to the MT models in the WMT24 test set. For each MT model and difficulty estimator, circles mark the average correlation across en→X language pairs and bars report ± 1 standard deviation.

6.7 Bias toward MT models

Previous sections have shown that sentinel-src-24 can be effectively employed to sample the most difficult source texts out of a larger body of documents. However, an estimator's difficulty scores may select the texts that some MT models find more challenging than others, thereby biasing the resulting benchmark. To investigate this, we measure the alignment between our two best-performing difficulty estimators – i.e., sentinel-src-24 and Artificial Crowd (XCOMET) – and the concept of translation difficulty, as it was defined in Section 6.3, for each MT model in isolation. Specifically, for each model and each en→X translation direction, we measure the Kendall's τ_b correlation between the difficulty scores from the estimators and the human quality judgments assigned to that model's translations. Then, we average these correlations across language pairs, reporting the mean and standard deviation to capture their variability across translation directions. To ensure the averages are robust, this analysis only includes MT models that translated the WMT24 source texts into at least five target languages. We present results in Figure 6.3.

Overall, the mean Kendall's τ_b correlations fall within a narrow range and have comparable standard deviations. This suggests that the notion of difficulty captured by both

estimators aligns relatively uniformly with the MT models' perceived difficulty. However, we note two exceptions: ONLINE-B exhibits a higher correlation with sentinel-src-24 than average, whereas Gemini-1.5-Pro shows a lower correlation with both estimators. While this is a preliminary investigation, this finding suggests that a benchmark created using these estimators could be disproportionately difficult for ONLINE-B and easier for Gemini-1.5-Pro. We leave for future work a deeper investigation into whether this discrepancy stems from a bias in the difficulty estimations or simply from variance, especially given that the correlations were averaged over a limited number of translation directions.

6.8 Conclusion

This chapter advances our ability to measure progress in MT evaluation by exploring new methods for selecting test data samples that remain challenging for modern MT systems and LLMs. To achieve this, we formally defined the task of translation difficulty estimation and introduced the Difficulty Estimation Correlation (DEC), a dedicated measure for evaluating the performance of difficulty estimators. We conducted a comprehensive evaluation of existing and newly proposed estimators, finding that models explicitly trained for the task significantly outperform traditional, heuristic-based methods and LLM-as-a-judge approaches. As a result, we identified sentinel-src-24 as the current state-of-the-art in translation difficulty estimation and further validated its performance in constructing difficult benchmarks. Building on these findings, we developed sentinel-src-25 by incorporating additional data into the training pipeline of sentinel-src-24, and released both models publicly.

Our research equips the MT community with tools for building more challenging benchmarks at scale, enabling more informative assessments of MT systems' true capabilities and better highlighting areas for improvement in the MT field.

Table 6.4. Difficulty Estimation Correlation (DEC) achieved by each method, per language, on the ESA-annotated WMT24. Ranks represent clusters of statistical significance and are computed following Freitag et al. (2024), which leverage the PERM-BOTH hypothesis test introduced by Deutsch et al. (2021).

		Average		CS→UK	EN→CS	EN→ES	EN→HI	EN→IS	EN→JA	EN→RU	EN→UK	EN→ZH
	Method	Rank	DEC	DEC	DEC	DEC	DEC	DEC	DEC	DEC	DEC	DEC
Oracle	Oracle	1	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	Oracle (source text + target lang)	2	0.301	0.298	0.303	0.280	0.271	0.381	0.252	0.320	0.305	0.302
	Oracle (source text only)	3	0.224	0.298	0.246	0.201	0.176	0.238	0.172	0.253	0.213	0.218
Heuristic	Text Length	10	0.121	0.024	0.133	0.129	0.143	0.206	0.078	0.142	0.100	0.132
	Syntactic Complexity	12	0.080	0.018	0.078	0.072	0.112	0.181	0.035	0.090	0.050	0.079
	Word Rarity	16	-0.040	0.016	-0.034	-0.044	-0.065	-0.093	-0.032	-0.043	-0.022	-0.043
Learned	sentinel-src-24	5	0.182	0.167	0.216	0.169	0.173	0.220	0.142	0.204	0.150	0.197
	sentinel-src	6	0.175	0.164	0.205	0.159	0.171	0.223	0.118	0.201	0.141	0.190
	PreCOMET Difficulty	8	0.153	0.137	0.193	0.131	0.139	0.188	0.120	0.166	0.131	0.170
	PreCOMET Diversity	9	0.142	0.059	0.167	0.134	0.129	0.213	0.120	0.159	0.130	0.165
LLM Judge	Command A (tgt-lang)	11	0.104	0.077	0.100	0.098	0.120	0.190	0.068	0.117	0.072	0.097
	GPT-4o (tgt-lang)	12	0.080	0.061	0.067	0.072	0.116	0.179	0.035	0.079	0.037	0.071
	GPT-4o	13	0.077	0.038	0.066	0.072	0.111	0.188	0.029	0.083	0.036	0.071
	Command A	14	0.072	0.045	0.063	0.062	0.103	0.169	0.026	0.079	0.029	0.072
Crowd Based	True (XCOMET)	3	0.221	0.203	0.271	0.194	0.213	0.273	0.173	0.238	0.195	0.233
	True (MetricX)	4	0.207	0.211	0.256	0.184	0.212	0.221	0.176	0.229	0.175	0.203
	Artificial (XCOMET)	6	0.177	0.175	0.192	0.146	0.179	0.240	0.128	0.194	0.160	0.183
	Artificial (MetricX)	7	0.166	0.181	0.174	0.121	0.162	0.247	0.128	0.180	0.136	0.162
	Random	15	0.003	-0.001	0.004	0.004	-0.008	0.010	0.008	0.005	0.008	0.000

Table 6.5. Difficulty Estimation Correlation (DEC) achieved by each method, per language, on the MQM-annotated WMT24. Ranks represent clusters of statistical significance and are computed following Freitag et al. (2024), which leverage the PERM-BOTH hypothesis test introduced by Deutsch et al. (2021).

		Average		EN→DE	EN→ES	JA→ZH
	Method	Rank	DEC	DEC	DEC	DEC
Oracle	Oracle	1	1.000	1.000	1.000	1.000
	Oracle (source text + target lang)	2	0.430	0.505	0.280	0.503
	Oracle (source text only)	3	0.404	0.488	0.221	0.503
Heuristic	Text Length	7	0.222	0.262	0.147	0.256
	Syntactic Complexity	10	0.170	0.158	0.073	0.278
	Word Rarity	14	-0.052	-0.114	-0.043	0.001
Learned	sentinel-src-24	5	0.246	0.278	0.168	0.291
	sentinel-src	6	0.235	0.273	0.165	0.268
	PreCOMET Difficulty	10	0.169	0.219	0.129	0.159
	PreCOMET Diversity	10	0.167	0.241	0.143	0.117
LLM Judge	Command A (tgt-lang)	11	0.120	0.122	0.088	0.150
	Command A	11	0.114	0.117	0.060	0.165
	GPT-4o (tgt-lang)	12	0.090	0.096	0.064	0.110
	GPT-4o	12	0.090	0.111	0.049	0.109
Crowd Based	True (XCOMET)	4	0.278	0.309	0.208	0.315
	True (MetricX)	5	0.248	0.268	0.192	0.284
	Artificial (XCOMET)	8	0.207	0.243	0.159	0.220
	Artificial (MetricX)	9	0.185	0.209	0.145	0.201
	Random	13	0.002	0.003	0.004	0.000

Table 6.6. Average Kendall τ_b correlation between the difficulty scores assigned to a human translator and four MT models and LLMs.

	IOL	GPT-4	Claude3.5	Tower70B
Human	0.137	0.137	0.127	0.109
Tower70B	0.176	0.158	0.151	
Claude3.5	0.178	0.221		
GPT-4	0.202			

Table 6.7. Comparison of methods for selecting the most difficult 25% samples from the ESA and MQM test sets, evaluated using the average human score on the selected subset and the proportion of model outputs in the selected subset that achieve a perfect human score (100/100 for the ESA-annotated test sets, and 0 for the MQM-annotated ones. Results are calculated per language pair and then averaged. Lower is better.

	ESA		MQM	
Method	**AvgScore**	**%Perfect**	**AvgScore**	**%Perfect**
Oracle (source text only)	74.9	13.3%	-6.6	32.7%
Oracle (source text + target lang)	71.6	11.4%	-6.8	30.5%
sentinel-src-24	79.1	**12.1%**	**-5.1**	**39.6%**
Artificial Crowd (XCOMET-QE-XXL)	**78.3**	13.3%	-4.4	43.8%
Text Length	82.7	14.1%	-4.5	43.6%
LLM-as-a-Judge (Command A)	83.0	16.1%	-3.1	51.1%
Random	84.4	21.0%	-2.5	58.8%
Full test set	84.4	20.7%	-2.5	57.7%

Chapter 7

The 2025 Conference on Machine Translation

In the last two chapters, we discussed two of our most relevant contributions to the field of Machine Translation: the introduction of Sentinel metrics as tools to probe the reliability of the MT meta-evaluation (Chapter 5), and the introduction of the Translation Difficulty Estimation as the task of estimating the difficulty of translating a given text (Chapter 6).

This chapter is dedicated to illustrating briefly the impact that these research contributions had on the 2025 edition of the Conference on Machine Translation (WMT25), and more generally, on the Machine Translation field.

7.1 The Impact of Sentinel Metrics

Sentinel metrics had already been included as baselines in the WMT24 Metrics Shared Task (Freitag et al., 2024), and they have been included again as baselines in the WMT25 Automated Translation Quality Evaluation Systems Shared Task (Lavie et al., 2025).[1] Furthermore, the segment-level MT meta-evaluation, as conducted at WMT, now adopts exclusively the *Segment Grouping* strategy, as we recommend in Section 5.2.

[1]The WMT Metrics Shared Task was renamed into "Automated Translation Quality Evaluation Systems" Shared Task in 2025.

7.2 The Impact of Translation Difficulty Estimation

We employed our best-performing translation difficulty estimator, sentinel-src-25, to select challenging source texts for inclusion in the test set of the 2025 Conference on Machine Translation (Kocmi et al., 2025c). In previous WMT editions, source texts were collected via random sampling. In contrast, this year we introduced a difficulty-based sub-sampling procedure aimed at increasing the overall difficulty of the final test set.

First, we collected public domain or open-licensed source data from a range of domains, focusing on the most recent materials to minimize potential overlap with pretraining and fine-tuning data of the evaluated systems. Specifically, we gathered data from six domains – news, literary, speech, social, educational, and dialogue – and three source languages – Czech, English, and Japanese.

We then applied the translation difficulty estimation method introduced in Chapter 6 to identify the source documents most difficult to translate. The difficulty of each collected source document was estimated using sentinel-src-25, which, as described in Chapter 6, is an improved version of `sentinel-src` introduced in Chapter 5. For each document, we estimated the difficulty of individual paragraphs and averaged these scores to obtain document-level difficulty estimates. We retained only the most difficult documents for inclusion in the test sets, targeting approximately 9,000 words per domain, distributed across 60 to 100 segments.[2] To ensure that this process did not introduce ill-formed or garbled texts, we manually validated the selected sources and discarded any problematic ones. We applied the difficulty-based sub-sampling only to the news, speech, and social domains, as the amount of source data available for the other domains was insufficient for the procedure.

We present the results of the difficulty-based sampling in Figure 7.1, which shows the distribution of ESA scores assigned to the translations in the WMT24 and WMT25 test sets. The data used to compute these statistics include only translation directions present in both WMT editions and only translations with English as the source language. As evident from the figure, the frequency of high-quality translations decreased substantially. All histograms are lower in WMT25 than in WMT24. This is particularly noticeable for the "All" histograms, whose highest bar in WMT25 corresponds to the 80-90 ESA bucket, rather

[2] A segment corresponds to one or more paragraphs, where each paragraph is defined by a line break in the source document.

than the 90-100 one. Moreover, the decline in translation quality is also considerable for the top systems, as both the "Top System" and "Top Translation" bars are substantially lower in WMT25.

We emphasize that this comparison is not entirely fair, as different WMT editions may involve different annotators and annotations guidelines. In fact, the ESA slider marks (refer to Figure 2.3 for an illustration of how slider marks are used in the ESA interface) were updated in WMT25 to reflect the increased quality of translations produced by modern systems:

- 0% ESA indicates *No meaning preserved* in WMT24, but *Broken/poor* translations in WMT25;
- 33% ESA indicates *Some meaning preserved* in WMT24, but *Flawed* translations in WMT25;
- 66% ESA indicates *Most meaning preserved* in WMT24, but *Good* translations in WMT25;
- 100% ESA indicates *Perfect* translations both in WMT24 and WMT25.

In light of this, we present the histogram comparison only as an indication of the potential effects of our difficulty-based sampling, and caution should be exercised when interpreting these results. Nonetheless, since the only ESA label that remained unchanged corresponds to *Perfect* translations, we also compute the percentage of perfect translations in both test sets (again considering only the language directions with English source included both in WMT24 and WMT25). We find that 18.74% of translations are perfect in WMT24, while only 3.60% are perfect in WMT25. This result further confirms the effectiveness of the difficulty-based sampling, which enabled the WMT25 General Machine Translation task to evaluate systems on instances where they still fail, providing a more informative assessment of their true capabilities and better highlighting areas for improvement.

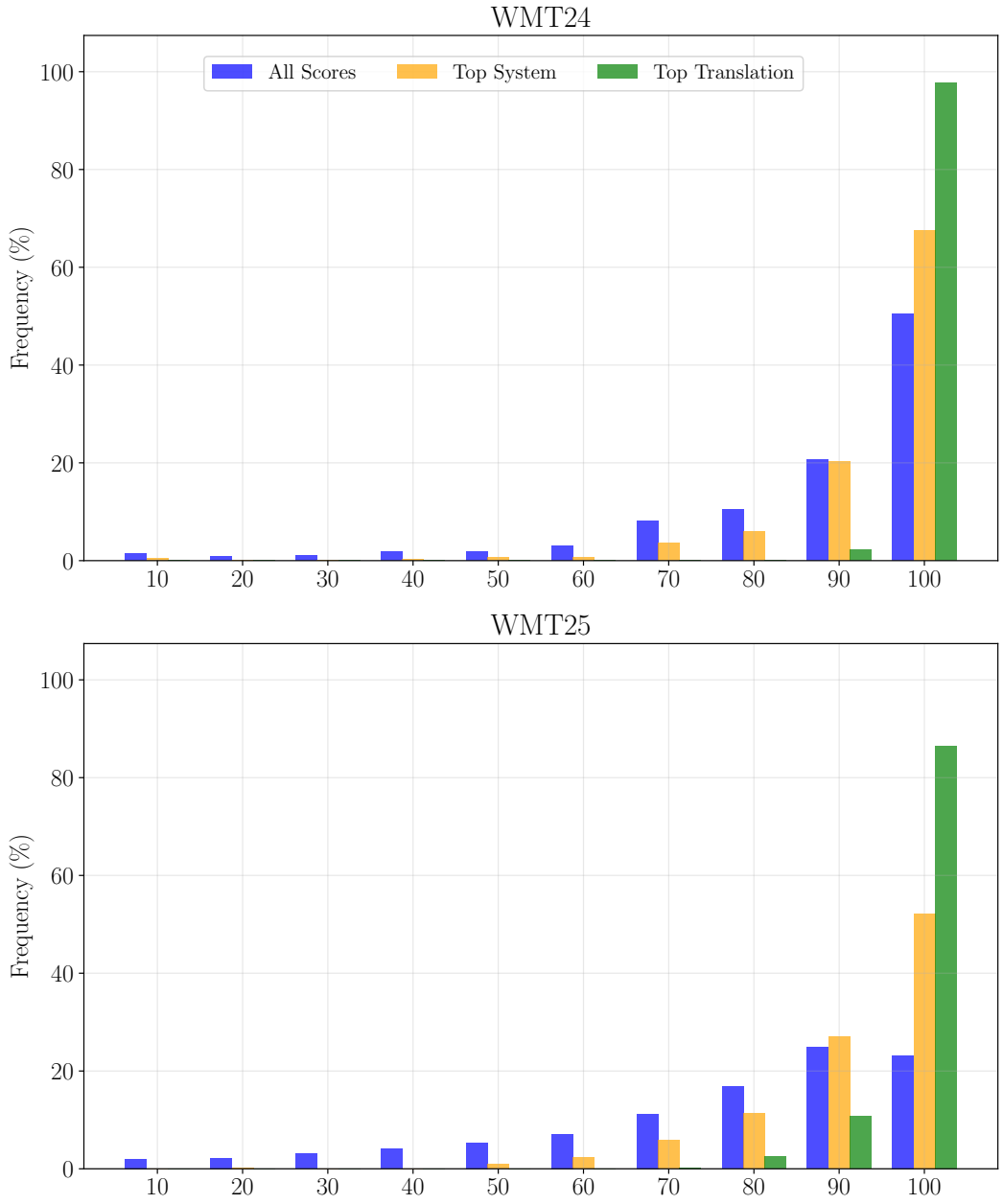


Figure 7.1. Distribution of the ESA scores assigned to translations of the WMT24 and WMT25 test sets. We included only language directions present in both test sets and only translations with English as the source language. "All" includes all translations, "Top system" includes translations pproduced by the best system for each translation direction, and "Top translation" includes the best translation for each source segment. The score buckets have a width of 10.

Chapter 8

Conclusions

In this final chapter, we present the closing remarks of this thesis. We begin by reviewing and summarizing the contributions to the field of Machine Translation introduced and developed throughout this document. Then, we discuss potential directions for future research, outlining several open problems whose resolution we believe to be of utmost importance for the field.

8.1 Summary of Contributions

This thesis has investigated the limits of MT evaluation, aiming to improve both how we measure progress in MT and our ability to interpret evaluation results.

Chapter 3 addressed the lack of interpretability in neural MT evaluation. We introduced the MaTESe metrics – neural metrics to tag translation spans containing errors and assign them a severity level – to mitigate the problem of *Error Attribution*, one of the main challenges we identified as causing the lack of interpretability in automatic MT evaluation (Section 1.1). We then addressed the problem of unclear metric *Performance* in Chapter 4, where we measured the capabilities of MT metrics using Precision, Recall, F-score, and Re-Ranking Precision (Section 4.2). Specifically, we designed evaluation scenarios that approximate two common applications: data filtering and translation re-ranking. This approach provided intuitive insights into metric capabilities beyond their usual correlation with human judgments.

Chapter 5 continued our contributions to MT meta-evaluation. By introducing the concept of sentinel metrics, we discovered that some widely used meta-evaluation strategies

were confounded by spurious correlations between text features and human judgment, potentially steering the research community toward optimizing systems and metrics for misleading objectives. Furthermore, by measuring the correlation between sentinel metrics and state-of-the-art MT metrics, we showed that metrics may be learning to base their evaluations on the same spurious correlations that compromise the meta-evaluation (Section 5.2.4). Our findings from Chapter 5 have influenced standard practices in MT meta-evaluation: sentinel metrics have been adopted as baselines for the WMT24 Metrics Shared Task (Freitag et al., 2024) and the WMT25 Automated Translation Quality Evaluation Systems Shared Task (Lavie et al., 2025). Moreover, our work established *Segment Grouping* as the recommended strategy for measuring correlations between metric scores and human judgments.

In Chapter 6, we explored how to improve MT evaluation along another dimension – focusing not on evaluation techniques but on the selection of test data. To enable the construction of test sets that challenge modern MT systems and LLMs, we introduced Translation Difficulty Estimation: the task of automatically identifying difficult-to-translate texts. We conducted a large-scale investigation to determine which approaches – from heuristic techniques to dedicated systems and LLM-as-a-Judge methods – are most effective for difficulty estimation, thereby establishing the first state of the art for this task. Furthermore, we released our best difficulty estimators, sentinel-src-24 and sentinel-src-25, and conducted additional analyses to assess their effectiveness as tools for constructing challenging MT benchmarks (Section 6.6).

Finally, in Chapter 7, we discussed the impact of our research on the 2025 edition of the Conference on Machine Translation, illustrating how we used sentinel-src-25 to sample difficult-to-translate texts for inclusion in the test set of the General Machine Translation Shared Task. We also showed that the scores assigned by human annotators to participants' translations suggest that this test set was more challenging than the WMT24 one, potentially increasing the benchmark's utility in discriminating between MT systems, assessing their true capabilities, and highlighting areas for improvement.

8.2 Future Work

This thesis has advanced MT evaluation in terms of accuracy and interpretability of results. However, many challenges remain unsolved. In a related work on MT evaluation (Proietti et al., 2025a), we showed that existing benchmarks suggest that automatic metrics have approached human parity, with automatic metrics ranking higher than human annotators. Contextually, we argued that rather than demonstrating true human parity, these results likely reflect limitations in the current MT evaluation and meta-evaluation frameworks, which may lack the granularity needed to capture further progress, particularly now that systems generate high-quality, difficult-to-distinguish translations. In this respect, we identified three main issues, each pointing to promising directions for future research.

The Difficulty of Test Data. Current test data might be too easy, limiting its ability to distinguish between MT systems and, consequently, hindering the evaluation of MT metrics. To address this, we introduced difficulty-based sampling as an alternative to random sampling for test set creation in Chapter 6. However, our work is only a starting point for research in MT difficulty estimation, and several research questions remain open. For example, in Section 6.7, we conducted a preliminary investigation into whether our difficulty estimation process biases the resulting test set in favor of or against specific MT systems. Future research should expand this analysis to ensure that difficult benchmarks remain fair, especially given that MT system developers are already integrating difficulty estimators into training pipelines to improve performance on challenging texts (Kocmi et al., 2025b).

Meta-evaluation. MT meta-evaluation is still not perfect. For example, we demonstrated that Pairwise Accuracy with Tie Calibration (Deutsch et al., 2023) is unfair to metrics that produce discrete outputs (Chapter 5). To date, MT meta-evaluation lacks a fair measure for jointly evaluating discrete and continuous metrics at the segment level. This not only complicates the evaluation of MT metrics but also makes it difficult to compare the performance of automatic MT evaluation to that of human annotators. Indeed, most metrics produce continuous assessments, whereas human judgments are inherently discrete.

Annotation Quality. Human annotation campaigns can produce low-quality annotations that yield low inter-annotator agreement, insufficient for reliably measuring progress. This is a longstanding challenge in Machine Translation, but it has become increasingly critical as translation quality has improved. At current performance levels, most translations differ only in subtle nuances, making consistent human evaluation more difficult. MT meta-evaluation is also affected by low inter-annotator agreement: when annotators disagree, it becomes difficult to determine whether our metrics are genuinely improving or merely aligning more closely with the preferences of a particular annotator. Future research should explore how to produce higher-quality, more reliable annotations of translation quality or how to reduce annotation costs without compromising reliability, to enable the collection of larger datasets and thereby reduce annotation variance.

Bibliography

Sina Ahmadi, Micha David Hess, Elena Álvarez-Mellado, Alessia Battisti, Cui Ding, Anne Göhring, Yingqiang Gao, Zifan Jiang, Andrianos Michail, Peshmerge Morad, Joel Niklaus, Maria Christina Panagiotopoulou, Stefano Perrella, Juri Opitz, Anastassia Shaitarova, and Rico Sennrich. 2025. ConLoan: A contrastive multilingual dataset for evaluating loanwords. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 30070–30090, Vienna, Austria. Association for Computational Linguistics.

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

Ana Sofia Vieira de Jesus Almeida. 2017. Difficulty estimation of machine translation. MSc Thesis.

Duarte Miguel Alves, José Pombal, Nuno M Guerreiro, Pedro Henrique Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, Pierre Colombo, José G. C. de Souza, and Andre Martins. 2024. Tower: An open multilingual

large language model for translation-related tasks. In *Proceedings of the First Conference on Language Modeling*.

Rohan Anil, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, Eric Chu, Jonathan H. Clark, Laurent El Shafey, Yanping Huang, Kathy Meier-Hellstern, Gaurav Mishra, Erica Moreira, Mark Omernick, Kevin Robinson, Sebastian Ruder, Yi Tay, Kefan Xiao, Yuanzhong Xu, Yujing Zhang, Gustavo Hernandez Abrego, Junwhan Ahn, Jacob Austin, Paul Barham, Jan Botha, James Bradbury, Siddhartha Brahma, Kevin Brooks, Michele Catasta, Yong Cheng, Colin Cherry, Christopher A. Choquette-Choo, Aakanksha Chowdhery, Clément Crepy, Shachi Dave, Mostafa Dehghani, Sunipa Dev, Jacob Devlin, Mark Díaz, Nan Du, Ethan Dyer, Vlad Feinberg, Fangxiaoyu Feng, Vlad Fienber, Markus Freitag, Xavier Garcia, Sebastian Gehrmann, Lucas Gonzalez, Guy Gur-Ari, Steven Hand, Hadi Hashemi, Le Hou, Joshua Howland, Andrea Hu, Jeffrey Hui, Jeremy Hurwitz, Michael Isard, Abe Ittycheriah, Matthew Jagielski, Wenhao Jia, Kathleen Kenealy, Maxim Krikun, Sneha Kudugunta, Chang Lan, Katherine Lee, Benjamin Lee, Eric Li, Music Li, Wei Li, YaGuang Li, Jian Li, Hyeontaek Lim, Hanzhao Lin, Zhongtao Liu, Frederick Liu, Marcello Maggioni, Aroma Mahendru, Joshua Maynez, Vedant Misra, Maysam Moussalem, Zachary Nado, John Nham, Eric Ni, Andrew Nystrom, Alicia Parrish, Marie Pellat, Martin Polacek, Alex Polozov, Reiner Pope, Siyuan Qiao, Emily Reif, Bryan Richter, Parker Riley, Alex Castro Ros, Aurko Roy, Brennan Saeta, Rajkumar Samuel, Renee Shelby, Ambrose Slone, Daniel Smilkov, David R. So, Daniel Sohn, Simon Tokumine, Dasha Valter, Vijay Vasudevan, Kiran Vodrahalli, Xuezhi Wang, Pidong Wang, Zirui Wang, Tao Wang, John Wieting, Yuhuai Wu, Kelvin Xu, Yunhan Xu, Linting Xue, Pengcheng Yin, Jiahui Yu, Qiao Zhang, Steven Zheng, Ce Zheng, Weikang Zhou, Denny Zhou, Slav Petrov, and Yonghui Wu. 2023. Palm 2 technical report.

Sahar Araghi and Alfons Palangkaraya. 2024. The link between translation difficulty and the quality of machine translation: A literature review and empirical investigation. *Language Resources and Evaluation*, 58(4):1093–1114.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL*

Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

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

Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019. Findings of the 2019 conference on machine translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy. Association for Computational Linguistics.

Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. 2020. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, 58:82–115.

Anna Bavaresco, Raffaella Bernardi, Leonardo Bertolazzi, Desmond Elliott, Raquel Fernández, Albert Gatt, Esam Ghaleb, Mario Giulianelli, Michael Hanna, Alexander Koller, Andre Martins, Philipp Mondorf, Vera Neplenbroek, Sandro Pezzelle, Barbara Plank, David Schlangen, Alessandro Suglia, Aditya K Surikuchi, Ece Takmaz, and Alberto Testoni. 2025. LLMs instead of human judges? a large scale empirical study across 20 NLP evaluation tasks. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 238–255, Vienna, Austria. Association for Computational Linguistics.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, page 41–48, New York, NY, USA. Association for Computing Machinery.

Frederic Blain, Chrysoula Zerva, Ricardo Rei, Nuno M. Guerreiro, Diptesh Kanojia, José G. C. de Souza, Beatriz Silva, Tânia Vaz, Yan Jingxuan, Fatemeh Azadi, Constantin Orasan, and André Martins. 2023. Findings of the WMT 2023 shared task on quality estimation. In *Proceedings of the Eighth Conference on Machine Translation*, pages 629–653, Singapore. Association for Computational Linguistics.

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

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

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Raphael Rubino, Lucia Specia, and Marco Turchi. 2017a. Findings of the 2017 conference on machine translation (WMT17). In *Proceedings of the Second Conference on Machine Translation*, pages 169–214, Copenhagen, Denmark. Association for Computational Linguistics.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Martin Popel, Matt Post, Raphael

Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin Verspoor, and Marcos Zampieri. 2016a. Findings of the 2016 conference on machine translation. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 131–198, Berlin, Germany. Association for Computational Linguistics.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Barry Haddow, Matthias Huck, Chris Hokamp, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Carolina Scarton, Lucia Specia, and Marco Turchi. 2015. Findings of the 2015 workshop on statistical machine translation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 1–46, Lisbon, Portugal. Association for Computational Linguistics.

Ondřej Bojar, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, and Christof Monz. 2018. Findings of the 2018 conference on machine translation (WMT18). In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 272–303, Belgium, Brussels. Association for Computational Linguistics.

Ondřej Bojar, Yvette Graham, and Amir Kamran. 2017b. Results of the WMT17 metrics shared task. In *Proceedings of the Second Conference on Machine Translation*, pages 489–513, Copenhagen, Denmark. Association for Computational Linguistics.

Ondřej Bojar, Yvette Graham, Amir Kamran, and Miloš Stanojević. 2016b. Results of the WMT16 metrics shared task. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 199–231, Berlin, Germany. Association for Computational Linguistics.

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

Zewen Chi, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-Ling Mao, Heyan Huang, and Ming Zhou. 2021. InfoXLM: An information-theoretic framework for cross-lingual language model pre-training. In *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.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

Hyung Won Chung, Thibault Fevry, Henry Tsai, Melvin Johnson, and Sebastian Ruder. 2021. Rethinking embedding coupling in pre-trained language models. In *Proceedings of the International Conference on Learning Representations*.

Cohere Team. 2025. Command a: An enterprise-ready large language model.

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

Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint, arXiv:2207.04672*.

Daniel Deutsch, Rotem Dror, and Dan Roth. 2021. A statistical analysis of summarization evaluation metrics using resampling methods. *Transactions of the Association for Computational Linguistics*, 9:1132–1146.

Daniel Deutsch, George Foster, and Markus Freitag. 2023. Ties matter: Meta-evaluating modern metrics with pairwise accuracy and tie calibration. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12914–12929, Singapore. Association for Computational Linguistics.

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

Shachar Don-Yehiya, Leshem Choshen, and Omri Abend. 2022. PreQuEL: Quality estimation of machine translation outputs in advance. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11170–11183, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning.

Sören Dreano, Derek Molloy, and Noel Murphy. 2023. Embed_Llama: Using LLM embeddings for the metrics shared task. In *Proceedings of the Eighth Conference on Machine Translation*, pages 738–745, Singapore. Association for Computational Linguistics.

Bryan Eikema and Wilker Aziz. 2020. Is MAP decoding all you need? the inadequacy of the mode in neural machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4506–4520, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Bryan Eikema and Wilker Aziz. 2022. Sampling-based approximations to minimum Bayes risk decoding for neural machine translation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10978–10993, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Achilles Fang. 1959. *Some Reflections on the Difficulty of Translation*, pages 111–134. Harvard University Press, Cambridge, MA and London, England.

António Farinhas, José de Souza, and Andre Martins. 2023. An empirical study of translation hypothesis ensembling with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11956–11970, Singapore. Association for Computational Linguistics.

Christian Federmann. 2012. Appraise: An open-source toolkit for manual evaluation of machine translation output. The Prague Bulletin of Mathematical Linguistics (PBML).

Zhaopeng Feng, Jiayuan Su, Jiamei Zheng, Jiahan Ren, Yan Zhang, Jian Wu, Hongwei Wang, and Zuozhu Liu. 2025. M-MAD: Multidimensional multi-agent debate for advanced machine translation evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7084–7107, Vienna, Austria. Association for Computational Linguistics.

Patrick Fernandes, Daniel Deutsch, Mara Finkelstein, Parker Riley, André Martins, Graham Neubig, Ankush Garg, Jonathan Clark, Markus Freitag, and Orhan Firat. 2023. The devil is in the errors: Leveraging large language models for fine-grained machine translation evaluation. In *Proceedings of the Eighth Conference on Machine Translation*, pages 1066–1083, Singapore. Association for Computational Linguistics.

Patrick Fernandes, António Farinhas, Ricardo Rei, José G. C. de Souza, Perez Ogayo, Graham Neubig, and Andre Martins. 2022. Quality-aware decoding for neural machine translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1396–1412, Seattle, United States. Association for Computational Linguistics.

Mara Finkelstein and Markus Freitag. 2024. MBR and QE finetuning: Training-time distillation of the best and most expensive decoding methods. In *Proceedings of The Twelfth International Conference on Learning Representations*.

Marina Fomicheva, Shuo Sun, Erick Fonseca, Chrysoula Zerva, Frédéric Blain, Vishrav Chaudhary, Francisco Guzmán, Nina Lopatina, Lucia Specia, and André F. T. Martins. 2022. MLQE-PE: A multilingual quality estimation and post-editing dataset. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4963–4974, Marseille, France. European Language Resources Association.

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

Markus Freitag, David Grangier, Qijun Tan, and Bowen Liang. 2022a. High quality rather than high model probability: Minimum Bayes risk decoding with neural metrics. *Transactions of the Association for Computational Linguistics*, 10:811–825.

Markus Freitag, Nitika Mathur, Daniel Deutsch, Chi-Kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Frederic Blain, Tom Kocmi, Jiayi Wang, David Ifeoluwa Adelani, Marianna Buchicchio, Chrysoula Zerva, and Alon Lavie. 2024. Are LLMs breaking MT metrics? results of the WMT24 metrics shared task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 47–81, Miami, Florida, USA. Association for Computational Linguistics.

Markus Freitag, Nitika Mathur, Chi-kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Tom Kocmi, Frederic Blain, Daniel Deutsch, Craig Stewart, Chrysoula Zerva, Sheila Castilho, Alon Lavie, and George Foster. 2023. Results of WMT23 metrics shared task: Metrics might be guilty but references are not innocent. In *Proceedings of the Eighth Conference on Machine Translation*, pages 578–628, Singapore. Association for Computational Linguistics.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André F. T. Martins. 2022b. Results of WMT22 metrics shared task: Stop using BLEU – neural metrics are better and more robust. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 46–68, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, George Foster, Alon Lavie, and Ondřej Bojar. 2021b. Results of the WMT21 metrics shared task: Evaluating metrics with expert-based human evaluations on TED and news domain. In *Proceedings*

of the Sixth Conference on Machine Translation, pages 733–774, Online. Association for Computational Linguistics.

Naman Goyal, Jingfei Du, Myle Ott, Giri Anantharaman, and Alexis Conneau. 2021. Larger-scale transformers for multilingual masked language modeling. In *Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021)*, pages 29–33, Online. Association for Computational Linguistics.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2013. Continuous measurement scales in human evaluation of machine translation. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 33–41, Sofia, Bulgaria. Association for Computational Linguistics.

Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André F. T. Martins. 2024. xCOMET: Transparent machine translation evaluation through fine-grained error detection. *Transactions of the Association for Computational Linguistics*, 12:979–995.

Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2018. A survey of methods for explaining black box models. *ACM Comput. Surv.*, 51(5).

Caglar Gulcehre, Tom Le Paine, Srivatsan Srinivasan, Ksenia Konyushkova, Lotte Weerts, Abhishek Sharma, Aditya Siddhant, Alex Ahern, Miaosen Wang, Chenjie Gu, Wolfgang Macherey, Arnaud Doucet, Orhan Firat, and Nando de Freitas. 2023. Reinforced self-training (rest) for language modeling. *arXiv preprint, arXiv:2308.08998*.

Sandra Hale and Stuart Campbell. 2002. The interaction between text difficulty and translation accuracy. *Babel*, 48:14–33.

Lifeng Han, Irina Sorokina, Gleb Erofeev, and Serge Gladkoff. 2021. cushLEPOR: customising hLEPOR metric using optuna for higher agreement with human judgments or pre-trained language model LaBSE. In *Proceedings of the Sixth Conference on Machine Translation*, pages 1014–1023, Online. Association for Computational Linguistics.

Hany Hassan, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mu Li, Shujie Liu, Tie-Yan Liu, Renqian Luo, Arul Menezes, Tao Qin, Frank Seide, Xu Tan, Fei Tian, Lijun Wu, Shuangzhi Wu, Yingce Xia, Dongdong Zhang, Zhirui Zhang, and Ming Zhou. 2018. Achieving human parity on automatic chinese to english news translation.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023a. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023b. DeBERTav3: Improving deBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. In *Proceedings of The Eleventh International Conference on Learning Representations*.

Zhiwei He, Xing Wang, Wenxiang Jiao, Zhuosheng Zhang, Rui Wang, Shuming Shi, and Zhaopeng Tu. 2024. Improving machine translation with human feedback: An exploration of quality estimation as a reward model. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8164–8180, Mexico City, Mexico. Association for Computational Linguistics.

Dan Hendrycks and Kevin Gimpel. 2017. Bridging nonlinearities and stochastic regularizers with gaussian error linear units.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55.

Juraj Juraska, Daniel Deutsch, Mara Finkelstein, and Markus Freitag. 2024. MetricX-24: The Google submission to the WMT 2024 metrics shared task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 492–504, Miami, Florida, USA. Association for Computational Linguistics.

Juraj Juraska, Mara Finkelstein, Daniel Deutsch, Aditya Siddhant, Mehdi Mirzazadeh, and Markus Freitag. 2023. MetricX-23: The Google submission to the WMT 2023 metrics

shared task. In *Proceedings of the Eighth Conference on Machine Translation*, pages 756–767, Singapore. Association for Computational Linguistics.

Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Petrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucińska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju yeong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Noveen Sachdeva, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Põder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan

Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Kat Black, Nabila Babar, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry, Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. 2025. Gemma 3 technical report.

Marzena Karpinska, Nishant Raj, Katherine Thai, Yixiao Song, Ankita Gupta, and Mohit Iyyer. 2022. DEMETR: Diagnosing evaluation metrics for translation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9540–9561, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

M. G. Kendall. 1945. The treatment of ties in ranking problems. *Biometrika*, 33:239–51.

Tom Kocmi, Sweta Agrawal, Ekaterina Artemova, Eleftherios Avramidis, Eleftheria Briakou, Pinzhen Chen, Marzieh Fadaee, Markus Freitag, Roman Grundkiewicz, Yupeng Hou, Philipp Koehn, Julia Kreutzer, Saab Mansour, Stefano Perrella, Lorenzo Proietti, Parker Riley, Eduardo Sánchez, Patricia Schmidtova, Mariya Shmatova, and Vilém Zouhar. 2025a. Findings of the wmt25 multilingual instruction shared task: Persistent hurdles in reasoning, generation, and evaluation. In *Proceedings of the Tenth Conference on Machine Translation (WMT 2025)*, pages 462–483, Suzhou, China. Association for Computational Linguistics.

Tom Kocmi, Arkady Arkhangorodsky, Alexandre Berard, Phil Blunsom, Samuel Cahyawijaya, Théo Dehaze, Marzieh Fadaee, Nicholas Frosst, Matthias Galle, Aidan Gomez, Nithya Govindarajan, Wei-Yin Ko, Julia Kreutzer, Kelly Marchisio, Ahmet Üstün, Sebastian Vincent, and Ivan Zhang. 2025b. Command-a-translate: Raising the bar of machine translation with difficulty filtering. In *Proceedings of the Tenth Conference on Machine*

Translation (WMT 2025), pages 789–799, Suzhou, China. Association for Computational Linguistics.

Tom Kocmi, Ekaterina Artemova, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Konstantin Dranch, Anton Dvorkovich, Sergey Dukanov, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Howard Lakougna, Jessica Lundin, Christof Monz, Kenton Murray, Masaaki Nagata, Stefano Perrella, Lorenzo Proietti, Martin Popel, Maja Popović, Parker Riley, Mariya Shmatova, Steinþór Steingrímsson, Lisa Yankovskaya, and Vilém Zouhar. 2025c. Findings of the wmt25 general machine translation shared task: Time to stop evaluating on easy test sets. In *Proceedings of the Tenth Conference on Machine Translation (WMT 2025)*, pages 355–413, Suzhou, China. Association for Computational Linguistics.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Benjamin Marie, Christof Monz, Kenton Murray, Masaaki Nagata, Martin Popel, Maja Popović, Mariya Shmatova, Steinthór Steingrímsson, and Vilém Zouhar. 2024a. Findings of the WMT24 general machine translation shared task: The LLM era is here but MT is not solved yet. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1–46, Miami, Florida, USA. Association for Computational Linguistics.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Philipp Koehn, Benjamin Marie, Christof Monz, Makoto Morishita, Kenton Murray, Masaaki Nagata, Toshiaki Nakazawa, Martin Popel, Maja Popović, Mariya Shmatova, and Jun Suzuki. 2023. Findings of the 2023 conference on machine translation (WMT23): LLMs are here but not quite there yet. In *Proceedings of the Eighth Conference on Machine Translation*, pages 1–42, Singapore. Association for Computational Linguistics.

Tom Kocmi, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Thamme Gowda, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Rebecca Knowles, Philipp Koehn, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki

Nakazawa, Michal Novák, Martin Popel, and Maja Popović. 2022. Findings of the 2022 conference on machine translation (WMT22). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 1–45, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Tom Kocmi and Ondřej Bojar. 2017. Curriculum learning and minibatch bucketing in neural machine translation. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 379–386, Varna, Bulgaria. INCOMA Ltd.

Tom Kocmi and Christian Federmann. 2023a. GEMBA-MQM: Detecting translation quality error spans with GPT-4. In *Proceedings of the Eighth Conference on Machine Translation*, pages 768–775, Singapore. Association for Computational Linguistics.

Tom Kocmi and Christian Federmann. 2023b. Large language models are state-of-the-art evaluators of translation quality. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 193–203, Tampere, Finland. European Association for Machine Translation.

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

Tom Kocmi, Vilém Zouhar, Eleftherios Avramidis, Roman Grundkiewicz, Marzena Karpinska, Maja Popović, Mrinmaya Sachan, and Mariya Shmatova. 2024b. Error span annotation: A balanced approach for human evaluation of machine translation. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1440–1453, Miami, Florida, USA. Association for Computational Linguistics.

Tom Kocmi, Vilém Zouhar, Christian Federmann, and Matt Post. 2024c. Navigating the metrics maze: Reconciling score magnitudes and accuracies. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1:*

Long Papers), pages 1999–2014, Bangkok, Thailand. Association for Computational Linguistics.

Tom Kocmi, Vilém Zouhar, Eleftherios Avramidis, Roman Grundkiewicz, Marzena Karpinska, Maja Popović, Mrinmaya Sachan, and Mariya Shmatova. 2024d. Error span annotation: A balanced approach for human evaluation of machine translation. *arXiv preprint, arXiv:2406.11580*.

Mateusz Krubiński, Erfan Ghadery, Marie-Francine Moens, and Pavel Pecina. 2021a. Just ask! evaluating machine translation by asking and answering questions. In *Proceedings of the Sixth Conference on Machine Translation*, pages 495–506, Online. Association for Computational Linguistics.

Mateusz Krubiński, Erfan Ghadery, Marie-Francine Moens, and Pavel Pecina. 2021b. MTEQA at WMT21 metrics shared task. In *Proceedings of the Sixth Conference on Machine Translation*, pages 1024–1029, Online. Association for Computational Linguistics.

Shankar Kumar and William Byrne. 2004. Minimum Bayes-risk decoding for statistical machine translation. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 169–176, Boston, Massachusetts, USA. Association for Computational Linguistics.

John P. Lalor, Hao Wu, Tsendsuren Munkhdalai, and Hong Yu. 2018. Understanding deep learning performance through an examination of test set difficulty: A psychometric case study. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4711–4716, Brussels, Belgium. Association for Computational Linguistics.

Samuel Läubli, Rico Sennrich, and Martin Volk. 2018. Has machine translation achieved human parity? a case for document-level evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4791–4796, Brussels, Belgium. Association for Computational Linguistics.

Alon Lavie, Greg Hanneman, Sweta Agrawal, Diptesh Kanojia, Chi-kiu Lo, Vilém Zouhar, Frederic Blain, Chrysoula Zerva, Eleftherios Avramidis, Sourabh Dattatray Deoghare,

Archchana Sindhujan, Jiayi Wang, David Ifeoluwa Adelani, Brian Thompson, Tom Kocmi, Markus Freitag, and Daniel Deutsch. 2025. Findings of the wmt25 shared task on automated translation evaluation systems: Linguistic diversity is challenging and references still help. In *Proceedings of the Tenth Conference on Machine Translation (WMT 2025)*, pages 414–461, Suzhou, China. Association for Computational Linguistics.

Zheng Wei Lim, Trevor Cohn, Charles Kemp, and Ekaterina Vylomova. 2023. Predicting human translation difficulty using automatic word alignment. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 11590–11601, Toronto, Canada. Association for Computational Linguistics.

Zheng Wei Lim, Ekaterina Vylomova, Charles Kemp, and Trevor Cohn. 2024. Predicting human translation difficulty with neural machine translation. *Transactions of the Association for Computational Linguistics*, 12:1479–1496.

Liyuan Liu, Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Jiawei Han. 2021. On the variance of the adaptive learning rate and beyond.

Xuebo Liu, Houtim Lai, Derek F. Wong, and Lidia S. Chao. 2020a. Norm-based curriculum learning for neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 427–436, Online. Association for Computational Linguistics.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020b. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

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

Arle Lommel, Hans Uszkoreit, and Aljoscha Burchardt. 2014b. Multidimensional quality metrics (MQM): A framework for declaring and describing translation quality metrics. *Tradumàtica*, 0(12):0455–463.

Qingsong Ma, Ondřej Bojar, and Yvette Graham. 2018. Results of the WMT18 metrics shared task: Both characters and embeddings achieve good performance. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 671–688, Belgium, Brussels. Association for Computational Linguistics.

Qingsong Ma, Johnny Wei, Ondřej Bojar, and Yvette Graham. 2019. Results of the WMT19 metrics shared task: Segment-level and strong MT systems pose big challenges. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 62–90, Florence, Italy. Association for Computational Linguistics.

Matouš Macháček and Ondřej Bojar. 2013. Results of the WMT13 metrics shared task. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 45–51, Sofia, Bulgaria. Association for Computational Linguistics.

Matouš Macháček and Ondřej Bojar. 2014. Results of the WMT14 metrics shared task. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 293–301, Baltimore, Maryland, USA. Association for Computational Linguistics.

Federico Martelli, Stefano Perrella, Niccolò Campolungo, Tina Munda, Svetla Koeva, Carole Tiberius, and Roberto Navigli. 2024. DiBiMT: A Gold Evaluation Benchmark for Studying Lexical Ambiguity in Machine Translation. *Computational Linguistics*, pages 1–79.

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

Nitika Mathur, Johnny Wei, Markus Freitag, Qingsong Ma, and Ondřej Bojar. 2020b. Results of the WMT20 metrics shared task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 688–725, Online. Association for Computational Linguistics.

Abhijit Mishra, Pushpak Bhattacharyya, and Michael Carl. 2013. Automatically predicting sentence translation difficulty. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 346–351, Sofia, Bulgaria. Association for Computational Linguistics.

Subhajit Naskar, Daniel Deutsch, and Markus Freitag. 2023. Quality estimation using minimum Bayes risk. In *Proceedings of the Eighth Conference on Machine Translation*, pages 806–811, Singapore. Association for Computational Linguistics.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan,

Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report.

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

Pennsylvania, USA. Association for Computational Linguistics.

Stefano Perrella, Lorenzo Proietti, Pere-Lluís Huguet Cabot, Edoardo Barba, and Roberto Navigli. 2024a. Beyond correlation: Interpretable evaluation of machine translation metrics. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20689–20714, Miami, Florida, USA. Association for Computational Linguistics.

Stefano Perrella, Lorenzo Proietti, Alessandro Scirè, Edoardo Barba, and Roberto Navigli. 2024b. Guardians of the machine translation meta-evaluation: Sentinel metrics fall in! In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16216–16244, Bangkok, Thailand. Association for Computational Linguistics.

Stefano Perrella, Lorenzo Proietti, Alessandro Scirè, Niccolò Campolungo, and Roberto Navigli. 2022. MaTESe: Machine translation evaluation as a sequence tagging problem. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 569–577, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Jan-Thorsten Peter, David Vilar, Daniel Deutsch, Mara Finkelstein, Juraj Juraska, and Markus Freitag. 2023. There's no data like better data: Using QE metrics for MT data filtering. In *Proceedings of the Eighth Conference on Machine Translation*, pages 561–577, Singapore. Association for Computational Linguistics.

Emmanouil Antonios Platanios, Otilia Stretcu, Graham Neubig, Barnabas Poczos, and Tom Mitchell. 2019. Competence-based curriculum learning for neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1162–1172, Minneapolis, Minnesota. Association for Computational Linguistics.

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

Lorenzo Proietti, Stefano Perrella, and Roberto Navigli. 2025a. Has machine translation evaluation achieved human parity? the human reference and the limits of progress. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 790–813, Vienna, Austria. Association for Computational Linguistics.

Lorenzo Proietti, Stefano Perrella, Simone Tedeschi, Giulia Vulpis, Leonardo Lavalle, Andrea Sanchietti, Andrea Ferrari, and Roberto Navigli. 2024. Analyzing homonymy disambiguation capabilities of pretrained language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 924–938, Torino, Italia. ELRA and ICCL.

Lorenzo Proietti, Stefano Perrella, Vilém Zouhar, Roberto Navigli, and Tom Kocmi. 2025b. Estimating machine translation difficulty. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 24261–24285, Suzhou, China. Association for Computational Linguistics.

Amy Pu, Hyung Won Chung, Ankur Parikh, Sebastian Gehrmann, and Thibault Sellam. 2021. Learning compact metrics for MT. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 751–762, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Miguel Ramos, Patrick Fernandes, António Farinhas, and Andre Martins. 2024. Aligning neural machine translation models: Human feedback in training and inference. In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)*, pages 258–274, Sheffield, UK. European Association for Machine Translation (EAMT).

Lance Ramshaw and Mitch Marcus. 1995. Text chunking using transformation-based learning. In *Third Workshop on Very Large Corpora*.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022a. COMET-22:

Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

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

Ricardo Rei, Nuno M. Guerreiro, JosÃ© Pombal, Daan van Stigt, Marcos Treviso, Luisa Coheur, José G. C. de Souza, and André Martins. 2023. Scaling up CometKiwi: Unbabel-IST 2023 submission for the quality estimation shared task. In *Proceedings of the Eighth Conference on Machine Translation*, pages 841–848, Singapore. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. 2022b. CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Parker Riley, Daniel Deutsch, George Foster, Viresh Ratnakar, Ali Dabirmoghaddam, and Markus Freitag. 2024. Finding replicable human evaluations via stable ranking probability. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4908–4919, Mexico City, Mexico. Association for Computational Linguistics.

Pedro Rodriguez, Joe Barrow, Alexander Miserlis Hoyle, John P. Lalor, Robin Jia, and

Jordan Boyd-Graber. 2021. Evaluation examples are not equally informative: How should that change NLP leaderboards? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4486–4503, Online. Association for Computational Linguistics.

Ananya Sai B, Tanay Dixit, Vignesh Nagarajan, Anoop Kunchukuttan, Pratyush Kumar, Mitesh M. Khapra, and Raj Dabre. 2023. IndicMT eval: A dataset to meta-evaluate machine translation metrics for Indian languages. In *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.

Keisuke Sakaguchi, Matt Post, and Benjamin Van Durme. 2014. Efficient elicitation of annotations for human evaluation of machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 1–11, Baltimore, Maryland, USA. Association for Computational Linguistics.

Darcy A Santor and James O Ramsay. 1998. Progress in the technology of measurement: Applications of item response models. *Psychological assessment*, 10(4):345.

Roy Schwartz, Gabriel Stanovsky, Swabha Swayamdipta, Jesse Dodge, and Noah A. Smith. 2020. The right tool for the job: Matching model and instance complexities. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6640–6651, Online. Association for Computational Linguistics.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

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

Petru Soviany, Radu Tudor Ionescu, Paolo Rota, and Nicu Sebe. 2022. Curriculum learning: A survey. *Int. J. Comput. Vision*, 130(6):1526–1565.

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

Robyn Speer. 2022. rspeer/wordfreq: V3.0.

Miloš Stanojević, Amir Kamran, Philipp Koehn, and Ondřej Bojar. 2015. Results of the WMT15 metrics shared task. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 256–273, Lisbon, Portugal. Association for Computational Linguistics.

Michal Stefanik, Vít Novotný, and Petr Sojka. 2021. Regressive ensemble for machine translation quality evaluation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 1041–1048, Online. Association for Computational Linguistics.

Kosuke Takahashi, Yoichi Ishibashi, Katsuhito Sudoh, and Satoshi Nakamura. 2021. Multilingual machine translation evaluation metrics fine-tuned on pseudo-negative examples for WMT 2021 metrics task. In *Proceedings of the Sixth Conference on Machine Translation*, pages 1049–1052, Online. Association for Computational Linguistics.

Surat Teerapittayanon, Bradley McDanel, and H.T. Kung. 2016. Branchynet: Fast inference via early exiting from deep neural networks. In *2016 23rd International Conference on Pattern Recognition (ICPR)*, pages 2464–2469.

Brian Thompson, Nitika Mathur, Daniel Deutsch, and Huda Khayrallah. 2024. Improving statistical significance in human evaluation of automatic metrics via soft pairwise accuracy. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1222–1234, Miami, Florida, USA. Association for Computational Linguistics.

Brian Thompson and Matt Post. 2020. Automatic machine translation evaluation in many languages via zero-shot paraphrasing. In *Proceedings of the 2020 Conference on Empirical*

Methods in Natural Language Processing (EMNLP), pages 90–121, Online. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models.

Bram Vanroy, Orphee de clercq, and Lieve Macken. 2019. Correlating process and product data to get an insight into translation difficulty. *Perspectives*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

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

Xin Wang, Yudong Chen, and Wenwu Zhu. 2022. A survey on curriculum learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9):4555–4576.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*.

Guillaume Wenzek, Vishrav Chaudhary, Angela Fan, Sahir Gomez, Naman Goyal, Somya Jain, Douwe Kiela, Tristan Thrush, and Francisco Guzmán. 2021. Findings of the WMT 2021 shared task on large-scale multilingual machine translation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 89–99, Online. Association for Computational Linguistics.

Minghao Wu, Jiahao Xu, Yulin Yuan, Gholamreza Haffari, Longyue Wan, Weihua Luo, and Kaifu Zhang. 2025. (perhaps) beyond human translation: Harnessing multi-agent

collaboration for translating ultra-long literary texts. *Transactions of the Association for Computational Linguistics*, 13:901–922.

Haoran Xu, Amr Sharaf, Yunmo Chen, Weiting Tan, Lingfeng Shen, Benjamin Van Durme, Kenton Murray, and Young Jin Kim. 2024. Contrastive preference optimization: Pushing the boundaries of llm performance in machine translation. *arXiv preprint, arXiv:2401.08417*.

Wenda Xu, Danqing Wang, Liangming Pan, Zhenqiao Song, Markus Freitag, William Wang, and Lei Li. 2023. INSTRUCTSCORE: Towards explainable text generation evaluation with automatic feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5967–5994, Singapore. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *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.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025. Qwen2.5 technical report.

Runzhe Zhan, Xuebo Liu, Derek F. Wong, and Lidia S. Chao. 2021a. Difficulty-aware machine translation evaluation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 26–32, Online. Association for Computational Linguistics.

Runzhe Zhan, Xuebo Liu, Derek F. Wong, and Lidia S. Chao. 2021b. Variance-aware

machine translation test sets. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*.

Shijie Zhang, Renhao Li, Songsheng Wang, Philipp Koehn, Min Yang, and Derek F. Wong. 2025. HiMATE: A hierarchical multi-agent framework for machine translation evaluation. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 11121–11145, Suzhou, China. Association for Computational Linguistics.

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

Wenbo Zhang. 2024. IOL research machine translation systems for WMT24 general machine translation shared task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 147–154, Miami, Florida, USA. Association for Computational Linguistics.

Xuan Zhang, Gaurav Kumar, Huda Khayrallah, Kenton Murray, Jeremy Gwinnup, Marianna J Martindale, Paul McNamee, Kevin Duh, and Marine Carpuat. 2018. An empirical exploration of curriculum learning for neural machine translation.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. 2019. MoverScore: Text generation evaluating with contextualized embeddings and earth mover distance. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 563–578, Hong Kong, China. Association for Computational Linguistics.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Vilém Zouhar, Shuoyang Ding, Anna Currey, Tatyana Badeka, Jenyuan Wang, and Brian Thompson. 2024. Fine-tuned machine translation metrics struggle in unseen domains.

In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 488–500, Bangkok, Thailand. Association for Computational Linguistics.

Vilém Zouhar, Peng Cui, and Mrinmaya Sachan. 2025. How to select datapoints for efficient human evaluation of nlg models?