UNIVERSIDADE DE LISBOA

INSTITUTO SUPERIOR TÉCNICO

Robust, Interpretable and Efficient MT Evaluation with Fine-tuned Metrics

Ricardo Costa Dias Rei

Supervisor: Doctor Maria Luísa Torres Ribeiro Marques da Silva Coheur
Co-Supervisor: Doctor Alon Lavie

Thesis approved in public session to obtain the PhD Degree in

Computer Science and Engineering

Jury final classification: Pass with Distinction and Honour

2024

UNIVERSIDADE DE LISBOA

INSTITUTO SUPERIOR TÉCNICO

Robust, Interpretable and Efficient MT Evaluation with Fine-tuned Metrics

Ricardo Costa Dias Rei

Supervisor: Doctor Maria Luísa Torres Ribeiro Marques da Silva Coheur
Co-Supervisor: Doctor Alon Lavie

Thesis approved in public session to obtain the PhD Degree in

Computer Science and Engineering

Jury final classification: Pass with Distinction and Honour

Jury

Chairperson: Doctor Maria Inês Camarate de Campos Lynce de Faria, Instituto Superior Técnico, Universidade de Lisboa

Members of the Committee:
Doctor Derek Fai Wong, Faculty of Science and Technology, University of Macau, China
Doctor Maria Luísa Torres Ribeiro Marques da Silva Coheur, Instituto Superior Técnico, Universidade de Lisboa
Doctor Rui Miguel Carrasqueiro Henriques, Instituto Superior Técnico, Universidade de Lisboa
Doctor George Foster, Google Research, Montreal, Canada

2024

To my wife, family and friends...

Preface

Chapter 3 is inspired by findings from the following papers:

> **R. Rei**, C. Stewart, A. C. Farinha, and A. Lavie. 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, Nov. 2020. Association for Computational Linguistics.
>
> **R. Rei**, C. Stewart, A. C. Farinha, and A. Lavie. Unbabel's participation in the WMT20 metrics shared task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 911–920, Online, Nov. 2020. Association for Computational Linguistics.
>
> **R. Rei**, A. C. Farinha, C. Zerva, D. van Stigt, C. Stewart, P. G Ramos, T. Glushkova, A F.T. Martins and A. Lavie. 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, Nov. 2021. Association for Computational Linguistics.
>
> **R. Rei**, J. G. C. de Souza, D. Alves, C. Zerva, A. C. Farinha, T. Glushkova, A. Lavie, L. Coheur and A. F.T. Martins. COMET-22: Unbabel-IST 2022 Submission for the Metrics Shared Task. In *Proceedings of the Seventh Conference on Machine Translation*, pages 1578–585, Abu Dhabi, United Arab Emirates, Nov. 2022. Association for Computational Linguistics.

Chapter 4 is composed of the two following papers:

> M. Treviso, N. M. Guerreiro, **R. Rei**, A. F. T. Martins. IST-Unbabel 2021 Submission for the Explainable Quality Estimation Shared Task In *Proceedings of the 2nd Workshop on Evaluation and Comparison of NLP Systems*, Punta Cana, Dominican Republic, Nov. 2021. Association for Computational Linguistics.
>
> **R. Rei**, N. M. Guerreiro, M. Treviso, A. Lavie, L. Coheur, A. F. T. Martins. The Inside Story: Towards Better Understanding of Machine Translation Neural Evaluation Metrics. In *Proceedings of the 61th Annual Meeting of the Association for Computational Linguistics*, Toronto, Canada, Jul. 2023. Association for Computational Linguistics.

Chapter 5 have appeared in the paper:

> **R. Rei**, A. C. Farinha, J. G. C. de Souza, P. G. Ramos, A. F. T. Martins, L. Coheur, and A. Lavie. Searching for Cometinho: The Little Metric That Could . In *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, Ghent, Belgium, Jun. 2022. Association for Computational Linguistics.

Chapter 6 have appeared in the paper:

> **R. Rei**, A. C. Farinha, C. Stewart, L. Coheur, and A. Lavie. MT-Telescope: An interactive platform for contrastive evaluation of MT systems. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 73–80, Online, Aug. 2021b. Association for Computational Linguistics.

Chapter 7 is composed of the two following papers:

> T. Glushkova, C. Zerva, **R. Rei**, A. F.T. Martins. Uncertainty-Aware Machine Translation Evaluation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, Punta Cana, Dominican Republic, Nov. 2021. Association for Computational Linguistics.

> P. Fernandes, A. Farinhas, **R. Rei**, J. G. C. de Souza, P. Ogayo, N. Graham, A. F. T. Martins. Quality-Aware Decoding for Neural Machine Translation. In *Accepted at the 2022 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, Seattle, Washington, Jul. 2022. Association for Computational Linguistics.

Acknowledgements

Obtaining a PhD is very much like embarking in an epic journey from which we can draw a lot of parallels with the Lord of the Rings. As I reach the culmination of this grand adventure, I find it fitting to express my deepest gratitude to those who have played pivotal roles.

First and foremost, my heartfelt appreciation goes to Catarina Farinha and Craig Stwart. Their unwavering support and presence throughout my expedition have been reminiscent of the loyal and steadfast Samwise Gamgee, whose unwavering commitment propelled Frodo forward.

Next, I must acknowledge the invaluable guidance and wisdom of Professor Alon Lavie, who assumed the mantle of the venerable Gandalf. Like the wise wizard, Professor Lavie served as my guide, mentor, and counselor, illuminating the path ahead and instilling in me the courage to overcome obstacles.

I owe a special debt of gratitude to Professor Luisa Coheur, whose extraordinary dedication and expertise embody the essence of Lady Galadriel. While orchestrating events behind the scenes, she has selflessly guided and nurtured not only this thesis but also my growth as a scholar.

Last but certainly not least, I am indebted to the esteemed fellowship of the ring, comprised of Professor André Martins, Paulo Dimas, and Professor João Graça. Like Frodo, whose burden was lightened by the support of his companions, I could not have undertaken this journey without their collaboration, encouragement, and invaluable insights.

This epic endeavor has been made possible through the generous support of the P2020 programs MAIA (contract 045909) and the fruitful collaboration between Unbabel and INESC-ID. Their collective contributions have paved the way for groundbreaking research and transformative discoveries.

In closing, I am reminded of the words of J.R.R. Tolkien: 'Even the smallest person can change the course of the future.' To all those mentioned above and countless others who have played a part in my odyssey, I offer my deepest gratitude. Together, we have embarked on a truly remarkable adventure, one that will forever shape my academic and personal journey.

Resumo

Com a crescente necessidade de Tradução Automática (TA) num mundo cada vez mais globalizado, existe também uma crescente necessidade de avaliar constantemente a qualidade das traduções produzidas. Esta avaliação pode ser realizada através de anotadores humanos que realizam avaliações de qualidade ou através da utilização de métricas automáticas. Embora a avaliação humana seja preferível, é cara e demorada. Consequentemente, ao longo da última década, o progresso na TA tem sido principalmente medido utilizando métricas automáticas que avaliam a similaridade lexical em relação a traduções de referência. No entanto, numerosos estudos demonstraram que as métricas baseadas em lexias não se correlacionam bem com os julgamentos humanos, lançando dúvidas sobre a fiabilidade da investigação em TA.

Motivado por estes desafios, o principal objetivo desta tese é melhorar o estado atual da avaliação da TA através do desenvolvimento de novas métricas automáticas que satisfaçam quatro critérios: 1) forte correlação com anotações humanas, 2) robustez em diferentes domínios e pares de línguas, 3) interpretabilidade e 4) eficiência.

Com base nos recentes avanços em processamento de linguagem natural, propomos que uma métrica supervisionada que incorpora o texto a traduzir no processo de avaliação. Para validar esta hipótese, introduzimos o COMET (Crosslingual Optimized Metric for Evaluation of Translation), uma framework de aprendizagem profunda para treino de modelos de avaliação de TA. Os modelos desenvolvidos dentro desta framework são treinados para prever anotações humanas de TA, como *Avaliações Diretas* (AD), *Métricas de Qualidade Multidimensional* (MQM) ou *Taxa de Edição de Tradução Mediada por Humanos* (HTER). Os nossos resultados demonstram que as métricas desenvolvidas dentro da nossa framework alcançam correlações estado da arte com julgamentos humanos em vários domínios e pares de línguas.

No entanto, métricas lexicais ainda têm méritos em termos de interpretabilidade e eficácia. Já métricas como as do COMET, baseadas em aprendizagem profunda, são consideradas "caixas-pretas" lentas. Para melhorar isso, usamos métodos de explicabilidade neuronal para mostrar como essas métricas usam informações de tokens ligadas a erros de tradução, comprovando sua interpretabilidade ao comparar mapas de saliência com anotações MQM. Também realizamos experiências para reduzir o custo computacional e tamanho dos modelos do COMET, mantendo suas correlações de estado da arte com anotações humanas, diminuindo a diferença de desempenho entre métricas lexicais e de redes neurais.

Apesar da robustez das métricas de TA, argumentamos que, ao aplicá-las e relatá-las no nível do sistema, são insuficientes para uma avaliação eficaz. Defendemos uma análise mais detalhada ao nível do segmento para compreender verdadeiramente a qualidade da TA. Para isso, desenvolvemos o MT-TELESCOPE, uma ferramenta de análise comparativa entre sistemas de TA, que expõe fatores de desempenho e analisa fenômenos como entidades mencionadas.

Ao longo dos últimos três anos, o COMET teve um impacto significativo na comunidade de investigação, com vários estudos a validar as nossas descobertas e a demonstrar a sua correlação superior com anotações humanas. Através deste trabalho, enfrentamos a tarefa ambiciosa de revolucionar a avaliação da TA introduzindo novas métricas que se destacam em termos de desempenho, robustez, interpretabilidade e eficiência computacional. Esta tese representa um progresso substancial para alcançar este objetivo.

Palavras-chave: Processamento de Linguagem Natural, Tradução Automática, Avaliação, Qualidade Estimada, COMET

Abstract

With the increasing need for Machine Translation (MT) in a world which is becoming globalized, there is also an increasing need to constantly evaluate the quality of the produced translations. This evaluation can be achieved through human annotators performing quality assessments or by employing automatic metrics. While human evaluation is preferred, it is expensive and time-consuming. Consequently, over the past decade, MT progress has primarily been measured using automatic metrics that assess lexical similarity against reference translations. However, numerous studies have demonstrated that lexical-based metrics do not correlate well with human judgments, casting doubt on the reliability of research in MT.

Motivated by these challenges, the main goal of this thesis is to enhance the current state of MT evaluation by developing new automatic metrics that satisfy four criteria: 1) strong correlation with human judgments, 2) robustness across different domains and language pairs, 3) interpretability, and 4) efficiency.

Based on recent advancements in cross-lingual language modeling, we propose that a supervised metric incorporating the source-language input into the evaluation process will yield more accurate MT evaluation. To validate this hypothesis, we introduce COMET (Crosslingual Optimized Metric for Evaluation of Translation), a neural framework for training multilingual MT evaluation models that serve as metrics. Models developed within the COMET framework are trained to predict human judgments of MT quality, such as *Direct Assessments* (DA), *Multidimensional Quality Metrics* (MQM), or *Human-mediated Translation Edit Rate* (HTER). Our results demonstrate that metrics developed within our framework achieve state-of-the-art correlations with human judgments across various domains and language pairs.

Nevertheless, lexical metrics still possess redeeming qualities in terms of interpretability and lightweight nature. In contrast, fine-tuned neural metrics like COMET are considered "slow black-boxes". To address this, we employ neural explainability methods to reveal that these metrics leverage token-level information directly associated with translation errors. We showcase their effectiveness for interpreting state-of-the-art fine-tuned neural metrics by comparing token-level neural saliency maps with MQM annotations. Additionally, we present several experiments aimed at reducing the computational cost and model size of COMET while maintaining its state-of-the-art correlation with human judgments, thus bridging the performance gap between lexical and model-based metrics.

Notwithstanding the strength of MT metrics, we argue that, when applied and reported at the system level, these are insufficient for effective MT evaluation. We claim that to truly understand the underlying MT quality, we need more fine-grained analysis built around segment-level scoring. To showcase the strength of more fine-grained segment-level analysis we developed MT-TELESCOPE. MT-TELESCOPE is an analysis tool for contrastive MT evaluation that takes system-level comparisons a step further by exposing the underlying factors behind performance and zooms into a fine-grained analysis of translation accuracy down to individual phenomena (e.g named entities).

Over the past three years, COMET has made a significant impact in the research community, with multiple studies validating our findings and demonstrating its superior correlation with human judgments. Through this work, we undertake the ambitious task of revolutionizing MT evaluation by introducing new metrics that excel in terms of performance, robustness, interpretability, and lightweight nature. This thesis represents substantial progress towards achieving this goal.

Keywords: MT Evaluation, Automatic Evaluation, Machine Translation, Quality Estimation, COMET

CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1

Introduction

Our world is becoming a global community and within that community there is a need to communicate and understand each other. Yet, with approximately 8 billion people living in this world, only 1.5 billion people speak English as a second-language, as reported by Statista[1]. Also, even if everyone would share the same second language, the comfort of communicating in one's mother tongue is undeniable. With roughly 6,500 different languages around the planet, translation is the only viable solution to break language barriers and build a united global community.

Translation can be performed by human translators or automatically through the use of Machine Translation (MT). While human translation in most cases results in high-quality translations, that are faithful to the original content of the message, it does not scale well as it is slow and expensive. Thus, the best way to scale translation is through the use of MT. Nonetheless, widespread adoption of MT raises quality concerns as it is less accurate and can lead to miscommunication. To mitigate this issue, it is essential to continuously measure the quality of MT when choosing the appropriate model for deployment or determining which translation to deliver. However, measuring MT quality can lead to a similar problem of translation itself. On one hand, when performed by humans it is slow but accurate and when performed automatically it is fast, scalable, but less reliable/accurate.

Automatic MT evaluation is used under two possible scenarios: *model selection*, performed during the development phase, where we are interested in comparing experiments and systems, and *model monitoring*, which concerns to assessing the quality of a translation produced by an MT system, after deployment, in an online fashion. For model selection we rely on *Reference-based MT metrics*. Reference-based MT metrics (which we will call just MT metrics through this document) produce a system-level score by comparing, for several test samples, the output of the MT model with a human-generated reference (the expected output). The system-level score is then used to rank different MT models developed during the *MT Development Phase* (Figure 1.1). On the other hand, model monitoring is performed through the use of *Quality Estimation (QE)* (also known as *Reference-free MT metrics*) and since it is performed after deployment, in an online fashion, it is constraint by the absence of an expected output. Therefore, QE metrics traditionally rely on machine learning to estimate the quality of an MT output. Also, in practical terms, while MT metrics are concerned with overall system performance, QE metrics are concerned with segment-level performance.

[1]The most spoken languages worldwide. Retrieved from `https://www.statista.com/statistics/266808/the-most-spoken-languages-worldwide/` at 02-06-2023

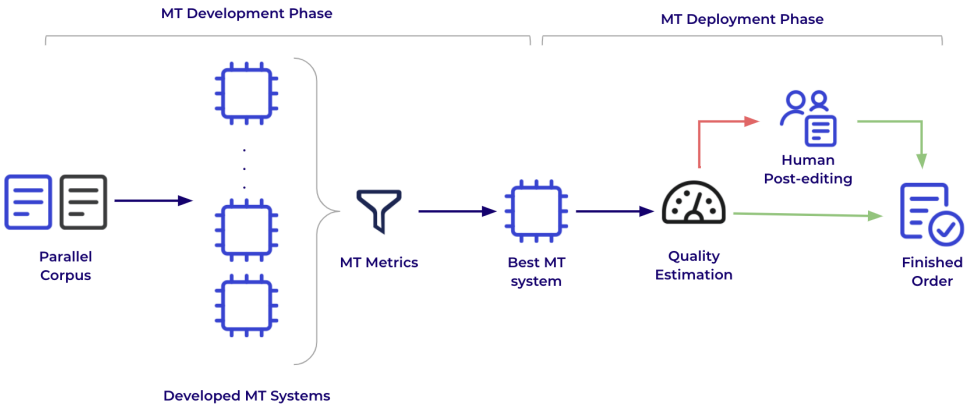


Figure 1.1: Illustration of the MT development and MT deployment phases. In the development phases several models are trained with different architectures and/or hyper-parameters. Then, all models are tested using one or more MT metrics and the best performing model is selected for deployment. During the deployment phase, for each translation request, a QE model is used to assess the translation quality. If the quality of the MT output is insufficient we send that translation to human post-editing before delivering the final order.

Modern Neural Machine Translation (NMT) result in a much higher quality of translation than previous statistical approaches and they often deviate from monotonic lexical transfer between languages. For this reason, it has become increasingly evident that we can no longer rely on traditional lexical-based MT metrics (e.g. BLEU (Papineni et al., 2002), METEOR (Lavie and Denkowski, 2009)) to provide accurate decisions of which MT system is better (Mathur et al., 2020a; Kocmi et al., 2021). Nonetheless, the MT community still relies largely on these outdated metrics and, for many years, no new widely-adopted standard has emerged. In 2019, the WMT News Translation shared Task received a total of 153 MT system submissions (Barrault et al., 2019). The Metrics Shared Task of the same year saw only 24 submissions, almost half of which were participants from the Quality Estimation Shared Task, adapted as metrics (Ma et al., 2019). The findings of the above-mentioned task highlight two major challenges to MT evaluation which we seek to address herein. Namely, that **metrics struggle to accurately correlate with human judgements at segment-level** and **fail to adequately differentiate the highest performing MT systems.**

This growing disparity between the quality of NMT systems and the limitations of traditional lexical metrics has motivated the need for improving the way automatic MT evaluation is conducted today. In light of this, we have identified several desiderata that we aim to address in this thesis.

Firstly, it is **crucial for MT metrics to exhibit a strong correlation with human judgments**. Human evaluation, when conducted properly, such as through the utilization of the Multidimensional Quality Metric (MQM) framework (Lommel et al., 2014), provides extremely informative insights. Therefore, automatic metrics should strive to capture the nuances and qualities that align with human annotations.

Secondly, **should be robust to a wide range of languages and domains**. MT is a multilingual and multi-domain task, and the evaluation metrics should be versatile enough to accommodate the diversity of translation outputs.

Thirdly, **MT metrics should be interpretable**. The black-box nature of many existing metrics hampers our ability to understand and trust their output. Therefore, we argue that interpretability is an important feature of any metric.

Furthermore, the **efficiency and computational speed** of an MT metric are important factors to take into account. One of the remaining redeeming qualities of lexical metrics is that they are incredibly light-weight. In certain MT applications, where thousands of translations need to be scored (e.g. outputs of multiple systems or different hypotheses of the same system, as in Minimum Bayes Risk (MBR) decoding (Kumar and Byrne, 2004)), inference speed becomes a crucial factor.

1.1 Contributions

Taking into consideration the desiderata presented above, the main contributions of this thesis can be summarized as follows:

1. We introduce COMET (**C**rosslingual **O**ptimized **M**etrics for **E**valuation of **T**ranslation)(Rei et al., 2020a), a framework for training highly multilingual and adaptable MT evaluation models that can function as metrics. Our framework takes advantage of recent breakthroughs in cross-lingual language modeling (Artetxe and Schwenk, 2019; Devlin et al., 2019; Conneau and Lample, 2019; Conneau et al., 2020) to generate prediction estimates of human judgments such as Direct Assessments (DA) (Graham et al., 2013), Human-mediated Translation Edit Rate (HTER)(Snover et al., 2006), and metrics compliant with MQM (Lommel et al., 2014). A distinct feature of COMET-based metrics is that, in contrast to previous mainstream metrics, the source-language input is incorporated into the evaluation process. Drawing inspiration from (Takahashi et al., 2020), we demonstrate the value of using a multilingual embedding space, which allows us to leverage information from all three inputs (source, reference, and translation). We show that incorporating source information enhances the performance and robustness of our MT evaluation models.

   Since its publication, COMET has made a significant impact in the community, and its metrics have been widely adopted by MT practitioners. We have continuously improved and kept COMET metrics state-of-the-art through our participation in several WMT shared tasks (Rei et al., 2020b, 2021a, 2022b,c).

2. While COMET metrics have demonstrated significant improvements in correlating with human judgments compared to traditional lexical metrics, they are "black boxes" that provide a single sentence-level score without revealing the underlying decision-making process. A second contribution of this thesis sheds light on the inner workings of these metrics. In our study entitled "The Inside Story" (Rei et al., 2023), we reveal that these metrics leverage token-level information that can be directly attributed to translation errors, as assessed through the comparison of token-level neural saliency maps with MQM annotations and synthetically-generated critical translation errors.

3. We introduce MT-TELESCOPE (Rei et al., 2021b), an analysis tool designed for comparing two MT systems side-by-side under different circumstances. MT-TELESCOPE can be used in conjunction with COMET to enable robust MT comparison. While MT metrics possess their strengths, when applied and reported at the system level, they can only provide a

general indication of the superiority of one system over another, often relying on a single score that may be limited to an arithmetic mean of segment-level score predictions. With MT-TELESCOPE, we aim to simplify the process of comparing MT systems for researchers and industry practitioners. The tool offers easy access to state-of-the-art MT evaluation metrics, statistical tests like bootstrap resampling (Koehn, 2004), dynamic filters to select specific phenomena within your test set, and a visual interface with plots to compare systems side-by-side on a segment-by-segment basis.

4. Finally, despite the high correlations with human judgment, the computational heaviness of COMET metrics, which are built on top of pre-trained language models, limits their usage in scenarios where speed and efficiency are crucial. In an attempt to reduce the computational cost of COMET and make it more efficient, we introduce several techniques based on pruning and knowledge distillation to create more compact and faster versions of COMET, which we dub COMETINHO's (Rei et al., 2022a).

To summarize, with Contribution 1 we introduce a novel approach that improves the current state of MT evaluation by achieving high correlations with human judgments across languages and domains. Contribution 2 focuses on improving the interpretability of COMET metrics, aiming to foster their adoption within the MT community. Contribution 3 promotes the widespread adoption of good practices in MT evaluation. Lastly, Contribution 4 addresses the issue of efficiency, an area where lexical metrics still excel over neural metrics. **Collectively, these contributions advance the field of MT evaluation, offering new evaluation methodologies, interpretability insights, best practices, and improved efficiency.**

In addition to the aforementioned contributions, I have had the privilege of actively collaborating on the following research projects:

- Explainable Quality Estimation (Treviso et al., 2021): In 2021, the Eval4NLP workshop organized the first shared task on Explainable QE. In our submission to this shared task (Treviso et al., 2021), we experimented with several explainability methods to extract the relevance of input tokens from sentence-level QE models built on top of multilingual pre-trained transformers. We showed that these attention methods, combined with gradient methods, can be effectively used to extract explanations for sentence-level results. This work served as the foundation for our paper on interpretable metrics (Rei et al., 2023) and received a best paper award at that workshop[2].

- Quality-Aware Decoding (Fernandes et al., 2022): Despite the progress in MT quality estimation and evaluation in recent years, decoding in NMT mostly centers around finding the most probable translation according to the model (MAP decoding), approximated with beam search, and is oblivious to quality considerations. In this work, we bridge the gap between quality estimation and decoding in NMT by leveraging recent breakthroughs in QE and MT metrics. Through various inference methods such as N-best reranking and MBR decoding, we propose quality-aware decoding for NMT. Our results demonstrate that quality-aware decoding consistently outperforms MAP-based decoding, not only on neural metrics such as COMET and BLEURT, but also on human evaluation based on MQM.

- Uncertainty Quantification in MT Evaluation (Glushkova et al., 2021; Zerva et al., 2022b): The COMET framework relies on point estimates, which provide limited information about

[2]https://eval4nlp.github.io/2021/awards.html

the quality of a given translation. In a series of works, we addressed this limitation by introducing uncertainty-aware MT evaluation and analyzing the trustworthiness of predicted quality. Firstly, we combined the COMET framework with two uncertainty estimation methods, Monte Carlo dropout and deep ensembles, enabling us to obtain quality scores along with confidence intervals, thus adding an extra layer of interpretability to COMET. Secondly, we focused on enhancing the COMET metric by incorporating an uncertainty prediction output. We explored different training objectives, including heteroscedastic regression, divergence minimization, and direct uncertainty prediction, to target various sources of aleatoric and epistemic uncertainty. Through our experiments, we achieved improved results in uncertainty prediction and demonstrated the ability of these predictors to address specific causes of uncertainty in MT evaluation.

- WMT 2021/2022 Metrics Shared Task (Freitag et al., 2021b, 2022): The metrics shared task has been a key component of the Conference on Machine Translation (WMT) since 2008, serving as a way to validate the use of automatic MT evaluation metrics and driving the development of new metrics. Since 2021, I have been involved in the organization of the WMT Metrics shared task and have contributed to several important modifications in the meta-evaluation of MT metrics. These modifications included transitioning from crowd-sourced DA to expert-based MQM annotations (Lommel et al., 2014), evaluating metrics across different domains (News, TED talks, Social media, e-Commerce, and Customer Support), and introducing a new subtask where participants could submit challenge sets targeting potential issues in MT metrics.

1.2 Document Overview

The subsequent sections of this document are organized as follows: In Section 2, we present the related work. Section 3 describes COMET, which serves as the main building block of this thesis. In Section 4, we discuss our work on understanding how neural MT metrics leverage token-level information to score sentences and, thus, making them more interpretable. Section 5 provides a description of MT-TELESCOPE. In Section 6, we outline a set of optimizations aimed at making COMET more computational efficient. Section 7 covers additional contributions. Finally, in Section 8, we present the main conclusions drawn from this work, along with considerations for future research.

Chapter 2

Related Work

As discussed in Chapter 1, the evaluation of Machine Translation (MT) can be conducted through automated methods, utilizing MT metrics that compare a hypothesis with its corresponding reference, or by employing a Quality Estimation (QE) system that estimates the quality of a translation by comparing it with the source text. Alternatively, human evaluation can also be employed.

In this chapter, we will begin by providing a brief overview of the various types of human evaluation metrics that will be utilized in this thesis. Subsequently, we will delve into an exploration of the different automatic metrics and analysis tools available.

2.1 Human Evaluation of Machine Translation

Automated metrics are typically evaluated by measuring their correlations with human annotations. Therefore, it is essential to comprehend the underlying distinctions among various human evaluation schemes in order to better interpret metric results. In this regard, we will start by explaining three common types of human evaluation that are frequently used as ground truth.

2.1.1 Human Translation Error Rate

A straightforward approach to assess the quality of an MT output is to have a professional translator correct it by making only the necessary changes. The corrected version of the original translation is referred to as a Post-Edited Translation (PE). In the case of high-quality MT outputs, the resulting PE will be very similar or identical to the original translation. However, when the MT output is of poor quality, the professional translator will need to make extensive modifications to improve it. Human Translation Error Rate (HTER) (Snover et al., 2006) is a metric defined as follows:

$$\text{HTER} = \frac{\text{number of edits}}{\text{number of words in the final PE}} \tag{1}$$

Intuitively, a larger number of edit operations leads to a higher HTER value, indicating a greater amount of editing effort. Conversely, lower HTER values indicate good translations where

minimal changes were required. Figure 2.1 illustrates the distinction between an MT output, a Post-Edited MT, and a human reference.

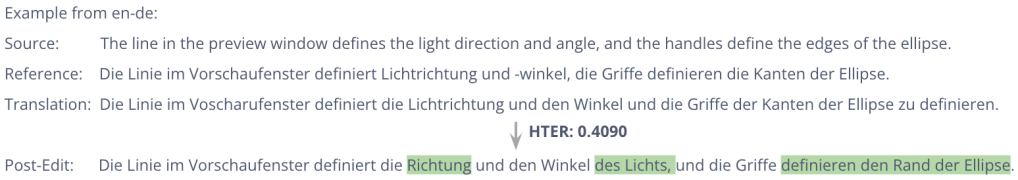


Figure 2.1: Example of a source with respective reference, translation and post-edited translation (PE). Note that the difference between the MT output and the resulting PE is 9 words (9 edit operations). This results in a $\frac{9}{22} = 0.4090$ HTER

2.1.2 Direct Assessments

The Conference on Machine Translation (WMT) organizes annual MT shared tasks, where participating systems are evaluated using direct estimates of quality, also known as *Direct Assessments* (DA) (Graham et al., 2013). A DA involves assigning a score between 0 and 100 to reflect the adequacy of a given translation. In WMT shared tasks, DA scores are typically collected using the APPRAISE TOOL (Federmann, 2010, 2018). Annotators are presented with the original source, the candidate translation, and a specific question: *"How accurately does the above candidate text convey the original semantics of the source text?"*. To provide their answer, annotators utilize a sliding bar ranging from 0 to 100, where 0 represents a completely inaccurate translation and 100 represents a perfect one. Figure 2.2 showcases a screenshot of the tool.

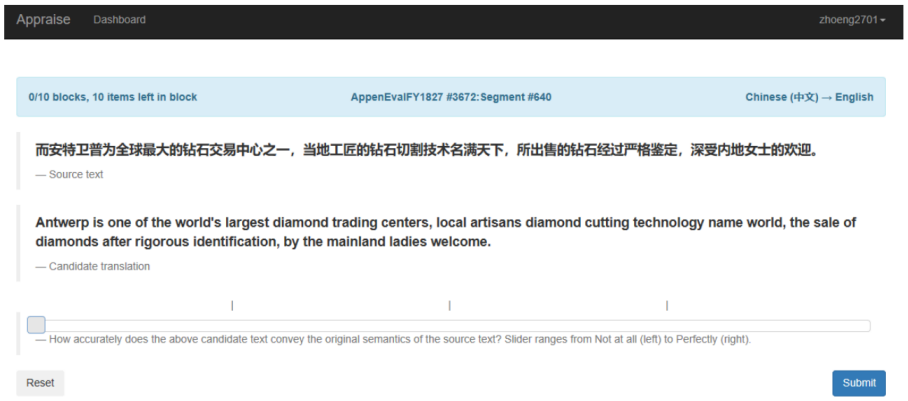


Figure 2.2: Example of a DA annotation performed on APPRAISE TOOL.

One of the main advantages of DA evaluation is its relative speed and the minimal training required for annotators. This makes DA evaluation well-suited for large-scale campaigns of MT evaluation, such as the annual workshops and conferences on Machine Translation (WMT). Moreover, the simplicity of the scoring process, utilizing a sliding bar, makes it accessible to a wide range of annotators, even those without specialized training in translation or linguistics.

2.1.3 Multidimensional Quality Metrics

Multidimensional Quality Metrics (MQM) is a versatile framework that offers a hierarchy of translation errors, which can be customized to suit specific applications. An evaluation schema based on MQM (Lommel et al., 2014) requires explicit error annotation and is often preferred over simpler evaluation schemas, such as DA (Freitag et al., 2021a).

When using MQM for evaluation, annotators are instructed to identify and highlight errors in the text. For each error, they are required to select a category and assign a severity level (minor, major, or critical). Each severity level has an associated weight. Typically, the final score for a segment is derived solely from the severity levels of errors, disregarding assigned categories. However, it is worth noting that the option to compute the score directly from severities also exists, offering flexibility in the evaluation process. A commonly used formula to compute the final score is as follows:

$$\text{MQM score} = 100 - \frac{I_{\text{Minor}} + 5 \times I_{\text{Major}} + 10 \times I_{\text{Crit.}}}{\text{Sentence Length} \times 100} \tag{2}$$

In the above equation, I_{Minor} represents the number of minor errors, I_{Major} represents the number of major errors, and $I_{\text{Crit.}}$ represents the number of critical errors. The MQM score is derived by subtracting the weighted error count from 100 and normalizing it based on the sentence length.

Due to the higher level of detail provided by MQM annotations, it has been embraced by translation companies like Unbabel[1] as a means to evaluate not only machine translation but also human translation.

Furthermore, MQM has recently piqued the interest of major technology companies, including Google, which has started investigating its application in machine translation. In their implementation, they annotate only *minor* and *major* errors, while introducing a "Non-translation" category with a predefined weight of 25. This category is assigned to translations that are so poor that the annotator can barely comprehend them. Additionally, Google does not normalize the score based on segment length. To mitigate the impact of long segments, they have imposed a maximum of five errors per segment, instructing annotators to select the five most severe errors when segments contain more errors. Consequently, the Google MQM equation is as follows:

$$\text{MQM score} = I_{\text{Minor}+5\times I_{\text{Major}}+25\times I_{\text{Non-translation}}} \tag{3}$$

2.2 Automatic Evaluation of Machine Translation

In this section, we will explore the current state of automatic evaluation for machine translation. Firstly, we will discuss the different approaches to reference-based evaluation (Section 2.2.1), and then we will delve into the current state-of-the-art reference-free evaluation (Section 2.2.2).

[1] https://help.unbabel.com/hc/en-us/articles/360004004013-How-does-Unbabel-assess-quality

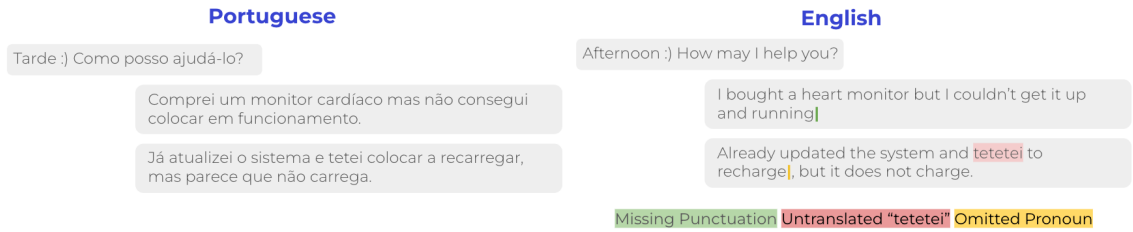


Figure 2.3: Example of MQM annotations performed on a customer support chat. This image reflects the kind of Human Translation Evaluations typically performed at Unbabel for quality audits. Text marked in green represent minor errors. Text marked in red represent critical errors. Finally, text marked in yellow represent major errors.

2.2.1 Machine Translation Metrics (Reference-based)

Reference-based machine translation metrics can be categorized into three groups: *n-gram-based* metrics, *embedding-based and unsupervised neural* metrics, and *neural fine-tuned* methods.

2.2.1.1 n-gram-based Metrics

n-gram-based metrics establish an alignment between the n-grams in a given translation hypothesis and their respective human reference. An n-gram match occurs when a sequence of n subsequent words/characters in the hypothesis aligns successfully with the reference. In n-gram-based metrics, good translations are characterized by a high number of aligned n-grams, indicating both adequacy and fluency. The choice of n determines the level of fluency considered in the evaluation. Based on the number and size of aligned n-grams, n-gram-based metrics calculate a score that is typically based on precision (fraction of n-grams in the hypothesis that also appear in the reference), recall (fraction of n-grams in the reference that also appear in the hypothesis), or a combination of both. The most widely used n-gram-based metric is BLEU (Papineni et al., 2002).

BLEU was one of the first machine translation metrics developed, and it has since become the standard method for evaluating translation quality in the MT community. The BLEU score is based on precision of word n-grams. To address the bias towards longer translations in precision-based metrics, BLEU incorporates a brevity penalty. In its early stages, BLEU showed correlations above 0.9 (on a scale from -1 to 1) with human judgments conducted on a scale from 1 (very bad) to 5 (very good). These high correlations were promising, as they allowed researchers to assess and compare the quality of their MT systems accurately without relying solely on human annotations. However, there was still room for improvement. BLEU considered only exact matches, neglecting partial matches in terms of surface forms, stemmed forms, and meanings. Other metrics followed in the footsteps of BLEU by calculating word overlaps between n-grams.

NIST (Doddington, 2002) modifies BLEU by assigning weights to n-gram matches based on their frequency in the test set. This idea was later employed in embedding-based metrics like YISI (Lo, 2019) and BERTSCORE (Zhang et al., 2020).

The ROUGE (Lin, 2004) family of metrics, originally designed for text summarization, focuses on n-gram recall instead of n-gram precision. They introduce the concept of skip-bigrams, allowing for gaps between matching words and providing more flexibility for partial matches compared to BLEU.

METEOR (Banerjee and Lavie, 2005) combines word precision and recall by computing an F-measure with a strong emphasis on recall. It relaxes the reliance on higher-order n-grams and incorporates support for morphological variants and synonyms using external resources such as WordNet (Fellbaum, 1998). METEOR-NEXT Denkowski and Lavie (2010) extends the previous version by incorporating paraphrase tables, which take word synonym matching a step further by considering entire sentence matches.

These variations and extensions to the original BLEU metric reflect ongoing efforts to enhance the evaluation of machine translation by accounting for different linguistic aspects and improving the assessment of translation quality.

Similarly to the aforementioned metrics, TER (Snover et al., 2006) calculates an alignment between the translation and the reference using the Levenshtein distance (Levenshtein, 1966). While HTER requires a post-edit human translation, TER can be computed using an independent reference translation.

For Asian languages like Chinese and Japanese, where white space is not used as a word delimiter, specialized word tokenizers are necessary to work with the aforementioned metrics. Additionally, when using words as n-gram units, matches between morphological variants are limited, which is particularly important for morphologically rich languages like Russian. To address this, metric developers have started exploring character-level metrics such as CHARTER Wang et al. (2016) (which operates similarly to TER at the character level) and CHRF Popović (2015).

CHRF Popović (2015) calculates an F1 score based on character-level 6-grams. This straightforward approach has shown competitive results compared to METEOR without the need for external resources. Moreover, since it relies solely on character sequences, this metric is suitable for various scripts without depending on language-specific tokenization. CHRF has been included in all editions of the WMT Metrics shared task since its proposal and consistently outperforms BLEU (Bojar et al., 2016, 2017; Ma et al., 2018, 2019; Mathur et al., 2020b).

2.2.1.2 Embedding-based and Unsupervised Neural Metrics

In recent years, word embeddings (Mikolov et al., 2013; Pennington et al., 2014; Peters et al., 2018; Devlin et al., 2019) have become popular for creating soft alignments between translation hypotheses and references. Metrics like YISI-* (Lo, 2019) have shown superior performance compared to lexical metrics since their introduction in the WMT18 metrics shared task (Ma et al., 2018).

The YISI- family of metrics consists of YISI-0, YISI-1, and YISI-2. Among these metrics, YISI-1 has been identified as the best performing one (Ma et al., 2018; Lo, 2019). YISI-1 measures the quality between a candidate translation and reference translations by computing cosine similarity between words and aggregating the scores to produce an n-gram-based similarity. Optionally, YISI-1 can include shallow semantic features such as semantic role labeling. YISI-0 is a lexical version of YISI-1 that evaluates lexical similarity using the longest common character substring accuracy between a translation hypothesis and a reference. On the other hand, YISI-2 is similar to YISI-1 but does not rely on references. It utilizes a multilingual embedding space to measure semantic similarity between a given translation hypothesis and a source sentence. The YISI- metrics have been submitted to the WMT Metrics shared task since 2018 and were the best performing metrics for WMT18 and WMT19.

Similar to YISI-*, BERTSCORE (Zhang et al., 2020) utilizes contextual embeddings from pre-trained transformers to create soft alignments using cosine similarity between translation and reference embeddings. Based on the alignment matrix, BERTSCORE provides precision, recall, and F1 scores. In experiments for MT, the authors reported higher correlations at the system level compared to YISI-1 on the WMT18 Metrics benchmark, although these correlations were not statistically significant for most language pairs.

In addition to YISI and BERTSCORE, which can be classified as **unsupervised neural metrics** due to their utilization of embeddings from neural language models like BERT, there are other embedding-based metrics that have demonstrated improvements over lexical counterparts. Examples of such metrics include METEOR-VECTOR (Servan et al., 2016), BLEU2VEC (Tättar and Fishel, 2017), and MOVERSCORE (Zhao et al., 2019).

Another notable example of an unsupervised neural metric is PRISM (Thompson and Post, 2020). Unlike the previously mentioned metrics, PRISM does not rely on a general-purpose language model. Instead, it leverages a multilingual machine translation (MT) model, which can be viewed as a paraphraser when used to translate from a source language to the same target language. By scoring the probability of a translation given its reference and vice versa, PRISM presents a novel approach to assessing the quality of machine translation.

In addition to its strong correlation with human judgments, PRISM offers the advantage of interpretability. The metric examines the log probabilities of individual tokens, providing insights into the preferences and biases of the paraphraser in relation to a specific translation. This interpretability allows users to gain a deeper understanding of the metric's assessment by analyzing the specific signals derived from the log probabilities of each token.

Throughout this thesis, we will frequently refer to BERTSCORE and PRISM as our baselines to represent the category of evaluation metrics discussed in this section.

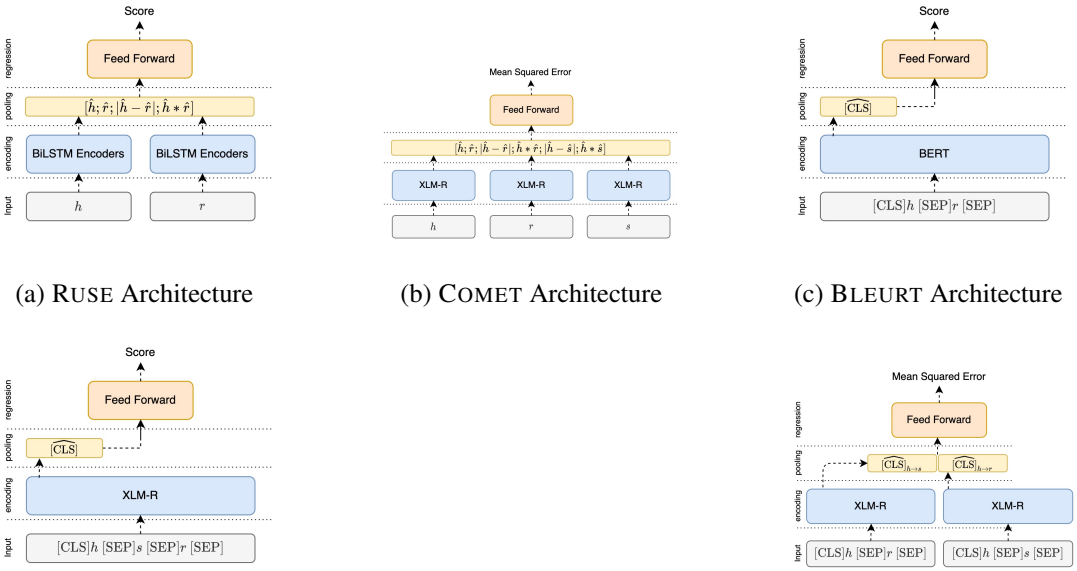
2.2.1.3 Neural Fine-tuned Metrics

Neural fine-tuned metrics take a different approach compared to lexical and embedding-based metrics. Instead of directly measuring similarity between a translation hypothesis and a reference, these metrics use supervised learning to mimic human perception of translation quality.

Some popular neural fine-tuned metrics include RUSE (Shimanaka et al., 2018a,b), BLEURT (Sellam et al., 2020), COMET(Rei et al., 2020a), C-SPEC(Takahashi et al., 2020), ROBLEURT (Wan et al., 2021) and UNITE (Wan et al., 2022).

RUSE utilizes three pre-trained Bidirectional Long Short-Term Memory neural networks (BiLSTM) to encode both the translation hypothesis and the reference. These BiLSTMs are trained on various tasks to extract different sentence embeddings, which are then concatenated to form feature vectors. The semantic relation between the feature vectors is captured using element-wise product and absolute difference operations. The concatenated feature vector is fed into a feed-forward regressor for quality assessment. In experiments, RUSE showed improvements in segment-level performance for the WMT16 Metrics compared to sentence-level BLEU.

BLEURT utilizes the BERT model to jointly encode the translation hypothesis and the reference, with the embedding of the `[CLS]` token serving as the input feature for the feed-forward regressor. The training process of BLEURT consists of two stages. In the warm-up stage, the metric predicts other machine translation (MT) metrics using a large-scale synthetic corpus. This



(a) RUSE Architecture (b) COMET Architecture (c) BLEURT Architecture

(d) ROBLEURT Architecture (e) C-SPEC Architecture

Figure 2.4: Neural architecture difference between RUSE, BLEURT, COMET, C-SPEC and ROBLEURT. All These metrics are made of the following main blocks; *input layer*, *encoding layer*, *pooling layer* and *regression layer*. The *input layer* receives the translation hypothesis (h) along with the corresponding reference (r) and, in COMET, C-SPEC and ROBLEURT, the source (s). Then prepares that input for the *encoding layer* where a pre-trained model is used to extract features. Those features are then passed to a pooling layer that creates an overall representation of all inputs and passes it to the *regression layer* that will produce a quality assessment. As we will see in section 3 COMET can also be trained for a translation ranking task where the quality assessment is given directly after the pooling layer.

stage helps the model learn from a wide range of data. In the fine-tuning stage, BLEURT is further trained using human quality assessment data, which provides more targeted and specific feedback.

In terms of performance, BLEURT has demonstrated superior results compared to other embedding-based metrics like BERTSCORE and MOVERSCORE in terms of correlations with human judgments. Additionally, the warm-up stage in the training process has shown to enhance the metric's robustness against quality drifts in machine translation, ensuring consistent performance over different translation scenarios.

C-SPEC is similar to BLEURT in terms of architecture but incorporates the source segment as an additional input. The translation hypothesis is encoded twice: once with the source segment and again with the reference segment. The resulting [CLS] token embeddings are concatenated and passed to a regression layer. For the WMT 2021 shared task, Takahashi et al. (2021) proposed training the model with pseudo negative examples, which are translations with induced critical errors.

ROBLEURT is an extension of BLEURT that introduces the source segment as part of the input and replaces the BERT model with XLM-R. During the warm-up stage, ROBLEURT leverages the COMET metric (as discussed in Section 2) to generate synthetic labels for training. Building on the advancements of ROBLEURT, the authors presented UNITE Wan et al. (2022), which introduces a novel training objective that combines reference-free and reference-based evaluation.

UNITE is jointly trained to estimate translation quality using three different inputs: (hypothesis, source), (hypothesis, reference), and (hypothesis, source, reference). Each input configuration provides a distinct perspective on the quality assessment. During inference, users have the flexibility to choose which input to utilize or can even ensemble the three output scores. This flexibility enables users to tailor the metric according to their specific evaluation needs.

These neural fine-tuned metrics have dominated the WMT Metrics task since 2020. Metrics like BLEURT, COMET, UNITE have formed distinct clusters of winning metrics in recent years (Freitag et al., 2021b, 2022). Figure 2.4 illustrates the differences in neural architectures for these metrics.

2.2.2 Quality Estimation (Reference-free Metrics)

In the field of machine translation evaluation, Quality Estimation (QE) has historically been regarded as a technique for predicting the quality of machine translations without relying on a reference translation. QE serves as a proxy for MT evaluation (Specia et al., 2009). Initially, when QE was introduced as a regression task (Specia et al., 2009), the majority of MT metrics focused on lexical aspects (as discussed in Section 2.2.1.1). However, with recent advancements in QE, particularly driven by large cross-lingual pre-trained models, these systems have emerged as competitive alternatives to traditional MT metrics that assess translation quality using reference translations (Rei et al., 2021a). Furthermore, the architecture of QE systems and neural fine-tuned metrics, such as COMET, share many similarities due to their reliance on large pre-trained models. The key distinction between these two types of systems lies in the presence or absence of a reference translation. To clearly differentiate between these approaches in this thesis, we will refer to the task of QE as reference-free evaluation/metrics.

Most machine translation (MT) metrics rely on a reference translation to compare and evaluate a hypothesis translation. However, the process of creating reliable reference translations is both

time-consuming and expensive, often requiring the expertise of professional translators. As a result, reference-free metrics have emerged as an alternative to traditional metrics, offering a way to save time and reduce costs.

The Conference in Machine Translation (WMT) annually hosts a shared task on quality estimation, which consists of various subtasks. These subtasks include sentence-level HTER (Human Translation Error Rate) prediction, word-level OK/BAD tagging, and more recently, sentence-level DA (Direct Assessment) prediction. The primary subtasks for many years were sentence-level HTER prediction and word-level OK/BAD tagging. In word-level OK/BAD tagging, the goal is to predict the edit operations that are performed during the post-editing process. This subtask is often performed alongside HTER prediction. Figure 2.5 provides an illustration of the objectives of these two subtasks.

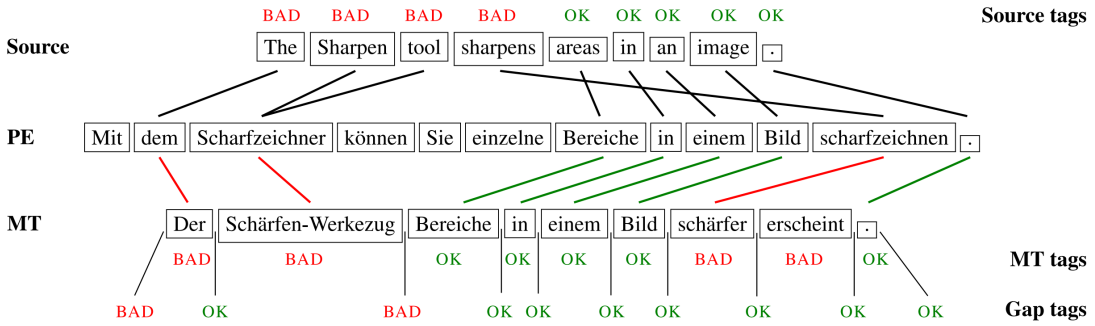


Figure 2.5: Example of taken from (Kepler et al., 2019) where an english source sentence (top), a German translation (bottom) and its post-edition (middle) are shown. We can also observe the different type of word-level quality tags. The HTER sentence score for this segment is given by the number of edit operations (8) normalized by the length of the post-edition (12), which results in $8/12 = 66.7\%$

In more recent years, particularly in 2020, the primary task of the QE shared task shifted from HTER prediction and word-level tags to DA prediction, aligning with the direction taken by metrics.

Two prominent open-source toolkits for QE are OpenKiwi (Kepler et al., 2019) and TransQuest (Ranasinghe et al., 2020). OpenKiwi emerged as the winner of the QE shared task in 2019 (Fonseca et al., 2019), while TransQuest claimed the top spot in 2020 (Specia et al., 2020). Interestingly, the system architectures employed by both toolkits remained consistent across these two years. They utilized a cross-lingual pretrained transformer model, XLM-RoBERTa, which jointly encoded the source and target (MT hypothesis) texts. These encoded representations were then leveraged to make predictions at either the word-level or segment-level. At the segment-level, these models shared fundamental similarities with BLEURT, with the distinction that they relied solely on the source text instead of utilizing the reference translation.

2.2.3 Natural Language Generation Analysis Tools

While the MT metrics discussed earlier provide a general assessment of the quality of developed systems, without fine-grained evaluation methods such as MQM (section 2.1.3), it is difficult to pinpoint particular weaknesses in the tested systems. In this section, we will introduce several

analysis tools that assist MT practitioners in gaining a deeper understanding of the underlying quality of their systems.

`ComparEval` (Klejch et al., 2015) was developed to assist MT developers in evaluating different systems and settings. It consists of three key components:

- *Evaluation mechanism*: This component computes MT evaluation scores according to several evaluation metrics and statistical tests.
- *Back-end engine*: The back-end engine monitors and stores previous translation outputs, such as outputs from previous MT systems.
- *Evaluation panel*: The evaluation panel is a graphical interface built on top of the other two components. It displays the performance of different systems being evaluated, both at the system-level and the segment-level.

Figure 2.6 illustrates the system-level evaluation panel, where multiple systems can be compared based on their performance across lexical metrics, such as BLEU, uni-gram F-Measure, Precision, and Recall.

Figure 2.7 demonstrates the segment-level evaluation panel. Here, users can compare two systems based on their lexical similarity to the reference translation. It's important to note that for segment-level analysis, users need to specify the two systems they want to compare and the lexical metric they want to use. The segments are then sorted according to the differences in the chosen metric.

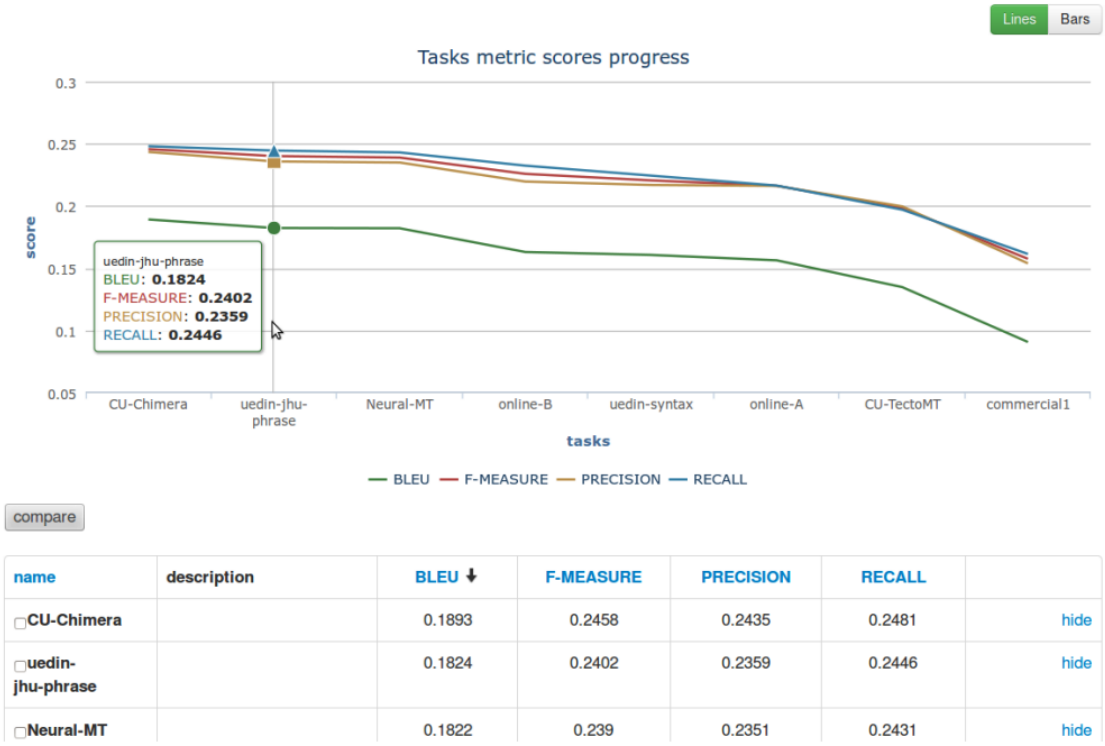
Another similar but more recent tool is `Compare-MT` (Neubig et al., 2019). It builds upon the features of `ComparEval` by incorporating additional metrics such as METEOR and ROUGE. It also introduces a *bucketed analysis* feature, which categorizes words or sentences into buckets and calculates relevant statistics for each bucket. This analysis enables users to answer questions such as, "On what types of sentences can one system outperform the other?" The sentence type can be defined based on factors like length or quality, as determined by the available metrics.

Figure 2.8 illustrates the results of a sentence-length *bucketed analysis* comparing a Phrase-Based Machine Translation (PBMT) system with a Neural Machine Translation (NMT) system. From this analysis, we can observe that the PBMT system performs better on longer sentences, while its overall performance is lower compared to the NMT system.

While `ComparEval` and `Compare-MT` primarily focus on machine translation, `VizSeq` (Wang et al., 2019) is designed to handle various natural language generation tasks, including those involving images, audio, and video as the source. Unlike previous tools that heavily rely on lexical metrics, `VizSeq` incorporates embedding-based metrics such as BERTSCORE.

2.2.4 Computationally Efficient Evaluation

Not much research has focused on improving the computational efficiency of neural fine-tuned metrics. To the best of our knowledge, only one study by Pu et al. (Pu et al., 2021) has addressed this issue. The authors aimed to explore the trade-off between multilinguality and model capacity for machine translation evaluation. They trained several smaller versions of RemBERT (Chung et al., 2021), which had 3, 6, and 12 layers. All these models were trained on the same data and



name	description	BLEU ↓	F-MEASURE	PRECISION	RECALL	
CU-Chimera		0.1893	0.2458	0.2435	0.2481	hide
uedin-jhu-phrase		0.1824	0.2402	0.2359	0.2446	hide
Neural-MT		0.1822	0.239	0.2351	0.2431	hide

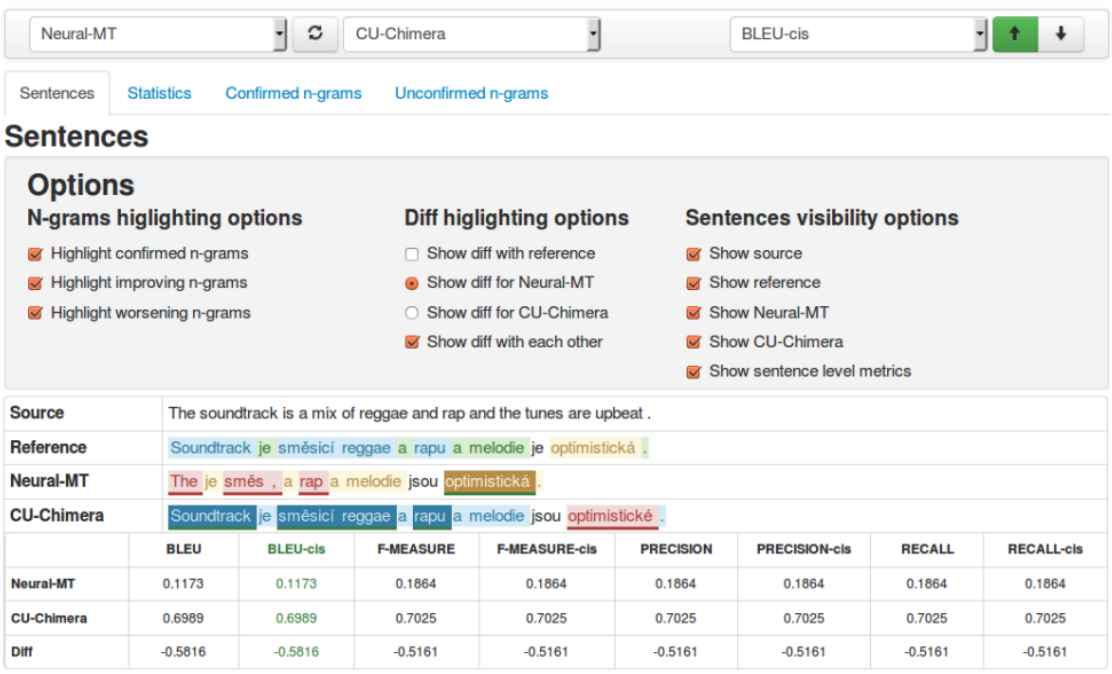
Figure 2.6: System-level *evaluation panel* from `ComparEval` (Klejch et al., 2015). In this panel we can observe 8 different systems being compared according to BLEU, n-gram F-Measure, Precision and Recall for a given testset.

supported 104 languages. The results indicated that compared to the original RemBERT model with 32 layers, the smaller versions performed significantly worse, particularly for zero-shot language pairs. To mitigate this problem, the authors proposed using synthetic data for low-resource languages and employing a 1-to-N distillation approach. This approach involved distilling the knowledge of a large teacher model into several students specialized in language families (e.g., Germanic or Romance languages). The resulting models achieved 92.6% of the teacher's performance while utilizing only one-third of its parameters.

2.2.5 Explainable Quality Estimation

In recent years, quality estimation (QE) systems such as OPENKIWI and TRANSQUEST have achieved remarkable performances (Fonseca et al., 2019; Specia et al., 2020, 2021). However, these systems, built on large pretrained language models, trade efficiency and interpretability for improved performance. The lack of interpretability undermines user trust in these advanced technologies, leading to the neglect of these high-quality systems by many users (Leiter et al., 2022).

To address these limitations, in 2021, the 2nd edition of the Workshop on "Evaluation & Comparison of NLP Systems" (Eval4NLP 2021) (Gao et al., 2021) organized a shared task on Explainable QE (Fomicheva et al., 2021). The primary goal of this task was to provide a sentence-level score indicating the overall quality of a translation and to explain this score by identifying the specific words that were considered errors. In this first edition, the authors introduced a new dataset where human annotators were asked to explain DA annotations by highlighting the errors present



	BLEU	BLEU-cis	F-MEASURE	F-MEASURE-cis	PRECISION	PRECISION-cis	RECALL	RECALL-cis
Neural-MT	0.1173	0.1173	0.1864	0.1864	0.1864	0.1864	0.1864	0.1864
CU-Chimera	0.6989	0.6989	0.7025	0.7025	0.7025	0.7025	0.7025	0.7025
Diff	-0.5816	-0.5816	-0.5161	-0.5161	-0.5161	-0.5161	-0.5161	-0.5161

Figure 2.7: Segment-level *evaluation panel* from `ComparEval` (Klejch et al., 2015). In this panel we can observe the lexical differences between the reference, and two systems: *Neural-MT* and *CU-Chimera*.

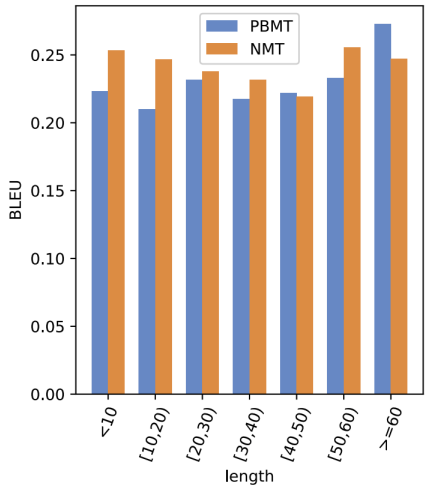


Figure 2.8: Sentence length *bucketed analysis* from `Compare-MT` (Neubig et al., 2019). In this plot we are comparing a Phrase-Based Machine Translation (PBMT) system against a Neural Machine Translation (NMT) system for different buckets defined according to sentence length.

in the translation. Later, in the WMT 2022 QE shared task (Zerva et al., 2022a), the organizers introduced an explainable QE subtask using MQM annotations, which naturally provided error spans.

Given the parallels between QE and MT Metrics and our interest in interpretability for this thesis, we actively participated in both editions of this shared task. Our work on explainable

QE is presented in Chapter 7 and serves as the foundation for Chapter 4, where we apply the findings from QE to our best-performing neural fine-tuned metrics. As we will see, while our work mainly focuses on attention and gradient methods, other participants followed simpler approaches that only leverage the underlying representations produced by the encoder models behind the QE system. An example of such an approach is the work by Tao et al. (2022), which, similar to BERTSCORE, creates alignments between the source and translation. Each word then receives a score that is the inverse of the cosine similarity. Intuitively, words that do not align well with the source sentence correspond to translation errors. This simple approach was a top-performing method for 5 out of 9 language pairs in the WMT 2022 QE shared task Zerva et al. (2022a).

Chapter 3

COMET: A Neural Framework for MT Evaluation

As previously mentioned, in this chapter, we present COMET (**C**rosslingual **O**ptimized **M**etric for **E**valuation of **T**ranslation), a framework for training highly multilingual and adaptable MT evaluation models that can function as metrics. Our framework takes advantage of recent breakthroughs in cross-lingual language modeling (Artetxe and Schwenk, 2019; Devlin et al., 2019; Conneau and Lample, 2019; Conneau et al., 2020) to generate prediction estimates of human judgments such as Direct Assessments (DA) (Graham et al., 2013), Human-mediated Translation Edit Rate (HTER) (Snover et al., 2006) and metrics compliant with the *Multidimensional Quality Metric* framework (Lommel et al., 2014).

To illustrate the effectiveness and flexibility of the COMET framework, train several models that estimate different types of human judgements. Through our experiments, we showcase promising progress in terms of improved correlation with human evaluation and robustness across multiple dimensions, such as high-quality MT, diverse domains, and various languages.

The work presented in this section builds upon our initial research (Rei et al., 2020a) published in the Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), as well as findings from our participation in the WMT Metrics shared task from 2020 to 2022 (Rei et al., 2020b, 2021a, 2022b).

3.1 Model Architectures

As we have seen in Section 2.1, human judgements of MT quality usually come in the form of segment-level scores, such as DA, MQM and HTER (see Section 2.1). For DA, it is common practice to convert scores into relative rankings (DARR) when the number of annotations per segment is limited (Bojar et al., 2017; Ma et al., 2018, 2019). This means that, for two MT hypotheses h_i and h_j of the same source s, if the DA score assigned to h_i is higher than the score assigned to h_j, h_i is regarded as a "better" hypothesis.[1] To encompass these differences, our framework supports two distinct architectures: The **Estimator model** and the **Translation Ranking model**. The fundamental difference between them is the training objective. While the

[1] In the WMT Metrics Shared Task, if the difference between the DA scores is not higher than 25 points, those segments are excluded from the DARR data.

Estimator is trained to regress directly on a quality score, the Translation Ranking model is trained to minimize the distance between a "better" hypothesis and both its corresponding reference and its original source. Both models are composed of a cross-lingual language model and a pooling layer.

3.1.1 Cross-lingual language model

The primary building block of all the models in our framework is a pretrained, cross-lingual language model such as multilingual BERT (Devlin et al., 2019), XLM (Conneau and Lample, 2019) or XLM-RoBERTa (Conneau et al., 2020)[2]. These models contain several transformer encoder layers that are trained to reconstruct masked tokens by uncovering the relationship between those tokens and the surrounding ones. When trained with data from multiple languages this pretrained objective has been found to be highly effective in cross-lingual tasks such as document classification and natural language inference (Conneau et al., 2020), generalizing well to unseen languages and scripts (Pires et al., 2019). For our preliminary experiments, we rely on XLM-RoBERTa (base) as our encoder model. Later on, we end up increasing the encoder size and replace the base model with the large one for better performance and generalization.

Given an input sequence $x = [x_0, x_1, ..., x_n]$, the encoder produces an embedding $\boldsymbol{e}_j^{(\ell)}$ for each token x_j and each layer $\ell \in \{0, 1, ..., k\}$. In our framework, we apply this process to the source, MT hypothesis, and reference in order to map them into a shared feature space.

3.1.2 Pooling Layer

The embeddings generated by the last layer of the pretrained encoders are usually used for fine-tuning models to new tasks. However, (Tenney et al., 2019) showed that different layers within the network can capture linguistic information that is relevant for different downstream tasks. In the case of MT evaluation, Zhang et al. (2020) showed that different layers can achieve different levels of correlation and that utilizing only the last layer often results in inferior performance. In this work, we used the approach described in (Peters et al., 2018) and pool information from the most important encoder layers into a single embedding for each token, $\boldsymbol{e}_j$, by using a layer-wise attention mechanism. This embedding is then computed as:

$$\boldsymbol{e}_{x_j} = \mu E_{x_j}^{\top} \boldsymbol{\alpha} \tag{1}$$

where μ is a trainable weight coefficient, $\boldsymbol{E}_{x_j} = [\boldsymbol{e}_{x_j}^{(0)}, \boldsymbol{e}_{x_j}^{(1)}, \dots \boldsymbol{e}_{x_j}^{(k)}]$ corresponds to the vector of layer embeddings for token x_j, and $\boldsymbol{\alpha} = \text{softmax}([\alpha^{(1)}, \alpha^{(2)}, \dots, \alpha^{(k)}])$ is a vector corresponding to the layer-wise trainable weights. In order to avoid overfitting to the information contained in any single layer, we used layer dropout (Kondratyuk and Straka, 2019), in which with a probability p the weight $\alpha^{(i)}$ is set to $-\infty$.

Finally, as in (Reimers and Gurevych, 2019), we apply average pooling to the resulting token embeddings to derive a sentence embedding for each segment.

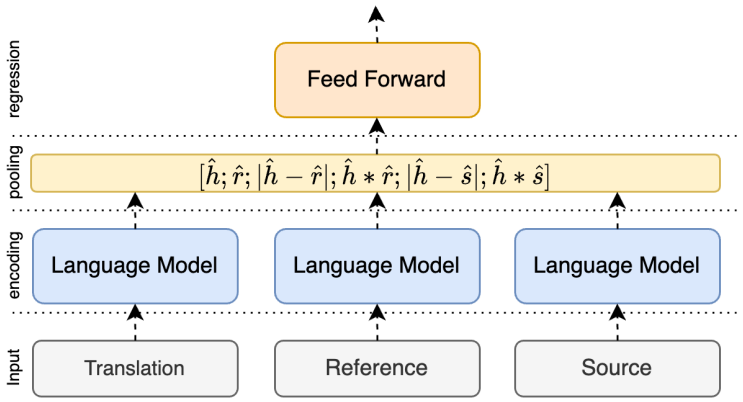


Figure 3.1: Estimator model architecture. The source, hypothesis and reference are independently encoded using a pre-trained cross-lingual language model. The resulting word embeddings are then passed through a pooling layer to create a sentence embedding for each segment. Finally, the resulting sentence embeddings are combined and concatenated into one single vector that is passed to a feed-forward regression module. The entire model is trained by minimizing the Mean Squared Error (MSE).

3.1.3 Estimator Model

Given a d-dimensional sentence embedding for the source, the hypothesis, and the reference, we adopt the approach proposed in RUSE (Shimanaka et al., 2018b) and extract the following combined features:

- Element-wise source product: $\boldsymbol{h} \odot \boldsymbol{s}$
- Element-wise reference product: $\boldsymbol{h} \odot \boldsymbol{r}$
- Absolute element-wise source difference: $|\boldsymbol{h} - \boldsymbol{s}|$
- Absolute element-wise reference difference: $|\boldsymbol{h} - \boldsymbol{r}|$

These combined features are then concatenated to the reference embedding $\boldsymbol{r}$ and hypothesis embedding $\boldsymbol{h}$ into a single vector $\boldsymbol{x} = [\boldsymbol{h}; \boldsymbol{r}; \boldsymbol{h} \odot \boldsymbol{s}; \boldsymbol{h} \odot \boldsymbol{r}; |\boldsymbol{h} - \boldsymbol{s}|; |\boldsymbol{h} - \boldsymbol{r}|]$ that serves as input to a feed-forward regressor. The strength of these features is in highlighting the differences between embeddings in the semantic feature space.

The model is then trained to minimize the MSE between the predicted scores and quality assessments (DA, HTER or MQM). Figure 3.1 illustrates the proposed architecture.

3.1.4 Translation Ranking Model

Our Translation Ranking model receives as input a tuple $\chi = (s, h^+, h^-, r)$ where h^+ denotes an hypothesis that was ranked higher than another hypothesis h^-. We then pass χ through our cross-lingual language model and pooling layer to obtain a sentence embedding for each segment in the

[2]We used only masked language models but this could be done with any kind of language model

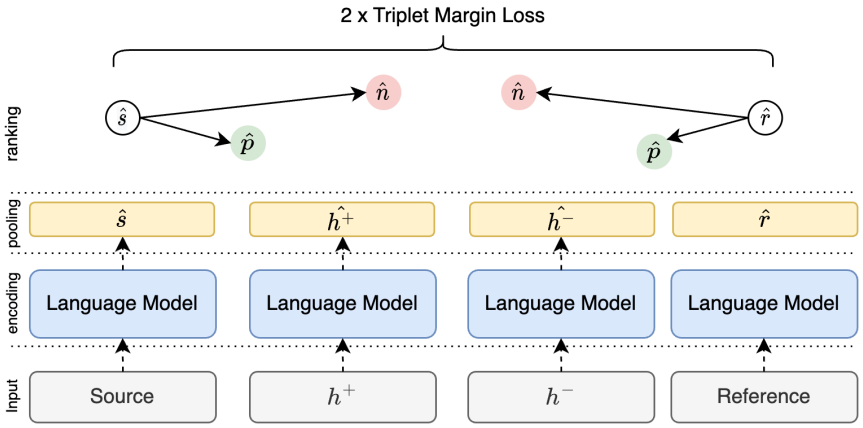


Figure 3.2: Translation Ranking model architecture. This architecture receives 4 segments: the source, the reference, a "better" hypothesis, and a "worse" one. These segments are independently encoded using a pre-trained cross-lingual language model and a pooling layer on top. Finally, using the Triplet Margin Loss (Schroff et al., 2015) we optimize the resulting embedding space to minimize the distance between the "better" hypothesis and the "anchors" (source and reference).

χ. Finally, using the embeddings $\{\boldsymbol{s}, \boldsymbol{h}^+, \boldsymbol{h}^-, \boldsymbol{r}\}$, we compute the Triplet Margin Loss (Schroff et al., 2015) in relation to the source and reference:

$$L(\chi) = L(\boldsymbol{s}, \boldsymbol{h}^+, \boldsymbol{h}^-) + L(\boldsymbol{r}, \boldsymbol{h}^+, \boldsymbol{h}^-) \tag{2}$$

where:

$$L(\boldsymbol{s}, \boldsymbol{h}^+, \boldsymbol{h}^-) = \max\{0, d(\boldsymbol{s}, \boldsymbol{h}^+) \ - d(\boldsymbol{s}, \boldsymbol{h}^-) + \epsilon\} \tag{3}$$

$$L(\boldsymbol{r}, \boldsymbol{h}^+, \boldsymbol{h}^-) = \max\{0, d(\boldsymbol{r}, \boldsymbol{h}^+) \ - d(\boldsymbol{r}, \boldsymbol{h}^-) + \epsilon\} \tag{4}$$

$d(\boldsymbol{u}, \boldsymbol{v})$ denotes the euclidean distance between $\boldsymbol{u}$ and $\boldsymbol{v}$ and ϵ is a margin. Thus, during training the model optimizes the embedding space so the distance between the anchors ($\boldsymbol{s}$ and $\boldsymbol{r}$) and the "worse" hypothesis $\boldsymbol{h}^-$ is greater by at least ϵ than the distance between the anchors and "better" hypothesis $\boldsymbol{h}^+$. Figure 3.2 illustrates the proposed architecture.

During inference, the described model receives a triplet $(s, \hat{h}, r)$ with only one hypothesis. The quality score assigned to $\hat{h}$ is the harmonic mean between the distance to the source $d(\boldsymbol{s}, \hat{\boldsymbol{h}})$ and the distance to the reference $d(\boldsymbol{r}, \hat{\boldsymbol{h}})$:

$$f(s, \hat{h}, r) = \frac{2 \times d(\boldsymbol{r}, \hat{\boldsymbol{h}}) \times d(\boldsymbol{s}, \hat{\boldsymbol{h}})}{d(\boldsymbol{r}, \hat{\boldsymbol{h}}) + d(\boldsymbol{s}, \hat{\boldsymbol{h}})} \tag{5}$$

Finally, we convert the resulting distance into a similarity score bounded between 0 and 1 as follows:

$$\hat{f}(s, \hat{h}, r) = \frac{1}{1 + f(s, \hat{h}, r)} \tag{6}$$

3.2 Corpora

To demonstrate the effectiveness of our described model architectures (Section 3.1), we train four MT evaluation models where each model targets a different type of human judgment. Next we present the corpora used to train each of these models:

3.2.1 HTER Corpora

To train a regressor that predicts HTER, we merge two publicly available corpora: QT21 (Specia et al., 2017) and APE-QUEST (Ive et al., 2020).

The QT21 corpus contains industry-generated sentences from the information technology or life sciences domains. It consists of a total of 173K tuples, including the source sentence, the corresponding human-generated reference, the MT hypothesis (from either a phrase-based statistical MT or a NMT system), and the post-edited MT (PE). The language pairs covered in this corpus are English→German, English→Latvian, English→Czech, and German→English.

The APE-QUEST corpus (Ive et al., 2020) comprises an additional 31K tuples for the language pairs English→Dutch, French, Portuguese in the legal domain. The MT output in this corpus was produced using neural MT systems.

For both corpus HTER score is obtained by computing TER between the MT hypothesis and the corresponding PE. Finally, after computing the HTER for each MT, we built a training dataset $D = \{s_i, h_i, r_i, y_i\}_{i=1}^{N}$, where s_i denotes the source text, h_i denotes the MT hypothesis, r_i the reference translation, and y_i the HTER score for the hypothesis h_i. In this manner, we seek to learn a regression $f(s, h, r) \rightarrow y$ that predicts the human-effort required to correct the hypothesis by looking at the source, hypothesis, and reference (but not the post-edited hypothesis).

3.2.2 MQM Corpora

For our MQM regressor, we utilized the MQM annotations from (Freitag et al., 2021a), which pertain to WMT2020, and these were combined with the TED Talk MQM annotations from the 2021 Metrics shared task (Freitag et al., 2021b). The resulting training dataset $D = \{s_i, h_i, r_i, y_i\}_{n=1}^{N}$ consists of the source text (s_i), the MT hypothesis (h_i) (a translation output from a WMT 2020/21 MT submission), the reference translation (r_i), and the MQM score (y_i) for each hypothesis h_i. This corpus comprises 56.5K tuples and covers three language pairs: English→Russian, English→German, and Chinese→English, as well as two domains: News and TED Talks.

All annotations for English→German and Chinese→English were collected by Google using a similar annotation process. However, for English→Russian, the annotations were collected by Unbabel, following slightly different guidelines that include critical errors (this difference is further explained in Section 2.1.3). To maintain consistency, we decided not to apply sentence length normalization for the English→Russian portion.

Furthermore, the annotations from the 2021 Metrics shared task also include annotations for the News domain. However, considering that we already had a substantial number of annotations for this domain from Freitag et al. (2021a), we made the decision to use the 2021 News domain as our development set. This choice allowed us to maintain consistency in our evaluation and effectively utilize the available data for training and testing our models.

3.2.3 DA and DARR Corpora

Since 2017, the organizers of the WMT News Translation Shared Task (Barrault et al., 2019) have been collecting human judgments in the form of DA (Graham et al., 2013). It is common practice to then transform those annotations into DARR (Ma et al., 2019). In this chapter, we utilize the annotations from 2017 to 2019, either in their original DA format, where the DA score is used for regression, or in the form of DARR, where our objective is to learn a discriminative function $f(s, h, r)$ such that the score assigned to a "better" hypothesis (h_i^+) is strictly higher than the score assigned to a "worse" hypothesis (h_i^-) ($f(s_i, h_i^+, r_i) > f(s_i, h_i^-, r_i)$).

The resulting DARR dataset $D = \{s_i, h_i^+, h_i^-, r_i\}_{i=1}^N$ contains 854K instances, covering 24 high and low-resource language pairs, including Chinese→English (zh-en) and English→Gujarati (en-gu). On the other hand, the DA dataset $D = \{s_i, h_i, r_i, y_i\}_{n=1}^N$, where y_i represents a Z-normalized DA score[3], consists of 698K tuples, covering the same 24 language pairs[4].

3.3 Experiments

Using the corpora described in the previous sections, we conducted experiments to train and evaluate different versions of our Estimator and Translation Ranking models. Specifically, we trained three versions of the Estimator model, as described in Section 3.1.3: one that performs regression on HTER (COMET-HTER), another that performs regression on MQM (COMET-MQM), and a third version that performs regression on DA (COMET-DA). Additionally, we trained the Translation Ranking model, as described in Section 3.1.4, using the WMT DARR corpus from 2017 and 2019 (COMET-DARR). In this section, we provide details of the training setup for these models and the corresponding evaluation setup.

3.3.1 Training Setup

The Estimators (COMET-HTER/MQM/DA) share the same training setup and hyperparameters (detailed in the Appendices). During training, we load the pretrained encoder and initialize both the pooling layer and the feed-forward regressor. The layer-wise scalars $\boldsymbol{\alpha}$ from the pooling layer are initially set to zero, while the weights from the feed-forward regressor are initialized randomly. To prevent catastrophic forgetting and improve generalization, we employ gradual unfreezing (Howard and Ruder, 2018) during training. This means that the encoder model is frozen for the first 30% of the first epoch, and then the entire model (except the embedding layer) is fine-tuned with a constant learning rate of $1e{-}5$. Contrarely, since COMET-DARR model does not have any additional parameters, it is fine-tuned from the beginning.

[3] In WMT shared tasks, DA scores are typically normalized using the mean and standard deviation of each annotator

[4] It is worth noting that DARR data is derived from DA using pairwise comparisons, which explains the larger size of DARR compared to the original DA dataset, even after excluding comparisons that fall within a 25-point difference.

Additionally, all models are trained using the AdamW optimizer (Loshchilov and Hutter, 2018) with a batch size of 32 and using 2021 data for validation and early stopping. As we have said previously, for the estimator models we used the News MQM annotations from (Freitag et al., 2021b) while for COMET-DARR we used the relative-ranks from that same year and languages.

3.3.2 Evaluation Setup

To address the low segment-level correlations exhibited by MT Metrics (Ma et al., 2019) while maintaining robustness across different language pairs and domains, our evaluation setup focuses on segment-level correlations using MQM annotations from the WMT 2022 Metrics task (Freitag et al., 2022). We consider four different domains: News, eCommerce, Social Media, and Customer Support. However, since these annotations only cover high-resource language pairs (English→German, English→Russian, Chinese→English) that include English, as a secondary evaluation, we use the DARR from WMT 2021[5]. This allows us to provide a comprehensive evaluation that covers different domains and includes challenging language pairs such as Hindi↔Bengali (`hi-bn` and `bn-hi`), Zulu↔Xhosa (`zu-xh` and `xh-zu`), English→Hausa (`en-ha`), English→Icelandic (`en-is`), and English→Japanese (`en-ja`).

For the WMT22 MQM data, we report the Pearson correlation coefficient (ρ) and Kendall's Tau (τ) according to the Perm-Both hypothesis test (Deutsch et al., 2021), using 500 re-sampling runs with a significance level (p) set to 0.05.

Regarding the WMT21 DARR test data, since we have pairwise comparisons, we use a modified version of Kendall's Tau, denoted as $\hat{\tau}$, which is defined as follows:

$$\hat{\tau} = \frac{\textit{Concordant} - \textit{Discordant}}{\textit{Concordant} + \textit{Discordant}} \tag{7}$$

Here, *Concordant* represents the number of times a metric assigns a higher score to the "better" hypothesis h^+, while *Discordant* represents the number of times a metric assigns a higher score to the "worse" hypothesis h^-. It also includes cases where the scores assigned to both hypotheses are the same. Notably, the DARR data excludes segments where the difference between the DA scores of two hypotheses with the same source is not higher than 25 points, ensuring that there are no ties in the data.

In addition to our proposed models, we include several baselines for comparison. We use two of the most commonly used lexical metrics, BLEU (Papineni et al., 2002) and CHRF (Popović, 2015), as well as two unsupervised neural metrics, BERTSCORE[6] (Zhang et al., 2020) and PRISM (Thompson and Post, 2020). These baselines provide a reference for the performance of our models.

Table 3.1: Segment-level correlations for WMT 2022 MQM annotations over News, eComerce, Social media, and Customer Support domains (Freitag et al., 2022). The metrics are Pearson (ρ) and Kendall Tau (τ). Results in bold indicate which metrics are top-performing for that specific language pair, domain and metric according to Perm-Both hypothesis test (Deutsch et al., 2021), using 500 re-sampling runs, and setting $p = 0.05$.

						BERTSCORE			COMET Models			
			BLEU	CHRF	PRISM	P	R	F1	HTER	MQM	DA	DARR
English→German	News	ρ	0.220	0.260	0.464	0.335	0.337	0.345	0.437	**0.502**	**0.507**	0.214
		τ	0.167	0.202	0.293	0.238	0.245	0.247	0.278	0.344	**0.361**	0.169
	eCom	ρ	0.173	0.222	0.306	0.266	0.241	0.261	0.366	**0.395**	0.339	0.159
		τ	0.179	0.212	0.290	0.241	0.237	0.246	0.326	**0.343**	0.321	0.253
	Social	ρ	0.172	0.220	0.265	0.245	0.256	0.256	0.300	0.333	**0.362**	0.158
		τ	0.130	0.168	0.212	0.188	0.198	0.197	0.244	0.270	**0.297**	0.156
	CS	ρ	0.228	0.285	0.226	0.271	0.252	0.270	0.281	**0.373**	**0.359**	0.153
		τ	0.201	0.257	0.207	0.238	0.225	0.235	0.230	**0.298**	**0.307**	0.160
English→Russian	News	ρ	0.169	0.230	**0.417**	0.310	0.314	0.321	0.343	**0.424**	**0.433**	0.154
		τ	0.125	0.164	0.310	0.222	0.229	0.231	0.273	**0.338**	**0.349**	0.139
	eCom	ρ	0.249	0.287	0.408	0.349	0.350	0.359	0.462	**0.540**	0.444	0.201
		τ	0.202	0.221	0.300	0.267	0.266	0.274	0.344	**0.366**	0.348	0.206
	Social	ρ	0.213	0.143	0.207	0.196	0.186	0.195	0.330	**0.481**	0.371	0.165
		τ	0.152	0.132	0.234	0.200	0.194	0.200	0.281	**0.372**	0.317	0.111
	CS	ρ	0.155	0.185	0.234	0.215	0.190	0.207	0.310	**0.333**	**0.328**	0.145
		τ	0.140	0.175	0.201	0.194	0.172	0.188	0.276	**0.285**	**0.296**	0.203
Chinese→English	News	ρ	0.097	0.078	0.288	0.200	0.212	0.215	0.372	**0.509**	0.377	0.047
		τ	0.046	0.042	0.203	0.138	0.145	0.148	0.277	**0.346**	0.304	0.025
	eCom	ρ	0.220	0.230	0.312	0.267	0.284	0.285	0.379	**0.436**	0.410	0.103
		τ	0.174	0.187	0.260	0.223	0.239	0.239	0.300	**0.330**	**0.337**	0.079
	Social	ρ	0.161	0.177	0.287	0.227	0.253	0.248	0.362	**0.501**	0.375	0.144
		τ	0.162	0.190	0.273	0.224	0.249	0.244	0.286	**0.320**	**0.324**	0.184
	CS	ρ	0.160	0.206	0.265	0.171	0.154	0.167	0.342	0.375	0.354	0.152
		τ	0.125	0.160	0.218	0.146	0.136	0.145	**0.285**	**0.291**	**0.293**	0.173
	AVG.	ρ	0.185	0.210	0.307	0.254	0.252	0.261	0.357	**0.433**	0.388	0.149
		τ	0.150	0.176	0.250	0.210	0.211	0.216	0.283	**0.325**	**0.321**	0.155

3.4 Preliminary Results

In this section, we will begin by presenting the performance of COMET models trained on different types of human assessments and compare these neural fine-tuned metrics with other "unsupervised" metrics across different domains and languages.

[5]We did not use all DARR from WMT 2021. Since we are already evaluating high-resource language pairs with MQM data we used only low-resource and mid-resource language pairs and we excluded all directions →English because they were annotated by mechanical turkers and the quality of those assessments is known to be low (Freitag et al., 2021a,b).

[6]For a fair comparison we used XLM-R base instead of the default encoder, mBERT.

3.4.1 Strong correlations across multiple domains

Table 3.1 presents the results for all four domains across three language pairs. We compare our COMET models against popular lexical metrics (BLEU and CHRF), as well as more recent unsupervised neural metrics: BERTSCORE and PRISM. Notably, our estimator models consistently outperform all other metrics, often by significant margins. While the MQM model exhibits the highest performance, it is noteworthy that the DA model also performs exceptionally well, with correlations close to those of the MQM model. This is surprising considering that DA annotations are known to be noisy and have poor correlation with MQM data (Freitag et al., 2021a).

3.4.2 Robustness to low-resource language pairs

Table 3.2 presents the results for low/mid resource language pairs. Similar to the MQM results, we compare our COMET models against BLEU and CHRF, BERTSCORE and PRISM. However, note that PRISM does not support most of these languages, so we can only report PRISM results for Bengali and Japanese. Comparing the results with the MQM evaluations, we observe that the estimator model trained on MQM performs the poorest among the COMET models, showing performances similar to CHRF. We attribute this lack of robustness to the limited number of language pairs seen during training, as publicly available MQM data only exists for three language pairs. On the other hand, the DA model, which has been trained on 24 language pairs, demonstrates excellent generalization and outperforms all other metrics on average.

Table 3.2: Segment-level correlations for WMT 2021 DARR over mid and low-resource language pairs. The correlation metric used is the WMT Kendall ($\hat{\tau}$) (Equation 7). * Because PRISM does not support all languages the average result is not directly comparable with other metrics in this table.

				BERTSCORE			COMET Models			
	BLEU	CHRF	PRISM	P	R	F1	HTER	MQM	DA	DARR
ZU-XH	0.381	**0.530**	-	0.444	0.481	0.469	**0.528**	0.403	0.512	0.510
XH-ZU	0.187	**0.301**	-	0.263	0.292	0.284	0.285	0.241	0.273	0.222
BN-HI	0.070	0.071	-	**0.134**	0.105	0.112	0.124	0.100	0.125	0.086
HI-BN	0.246	0.327	0.577	0.393	0.395	0.403	**0.479**	0.450	0.462	0.430
EN-JA	0.315	0.371	0.442	0.348	0.429	0.420	0.451	0.370	**0.497**	0.474
EN-HA	0.124	**0.186**	-	0.135	0.162	0.155	**0.178**	0.111	0.173	0.165
EN-IS	0.279	0.373	-	0.354	0.377	0.373	0.396	0.385	**0.434**	0.400
AVG.	0.229	0.308	0.510*	0.296	0.320	0.317	0.349	0.294	**0.354**	0.327

3.5 The `wmt22-comet-da` Metric

Based on the results we have just presented, we believe that employing a DA estimator is the most promising approach to achieve our goal of developing a **single metric that exhibits strong correlations with human judgments and demonstrates robustness across various domains and language pairs**. Unlike MQM data, which is limited in availability, DAs are abundant in the literature and easy to collect. This abundance opens up possibilities for training an estimator with

more diverse data, encompassing various domains and languages. However, we acknowledge that MQM data offers greater richness, and we recognize the potential of aligning the evaluation metric around MQM for future advancements, particularly if we can generate error spans with categories and severities.

For the WMT 2022 shared task (Rei et al., 2022b), one of our focus was on developing a robust DA estimator. To accomplish this, we concatenated all publicly available DA up to 2021, while reserving the 2021 DARR and 2022 MQM data for testing purposes. Therefore, the final DA corpus for this model consisted of the concatenation of DA from 2017, 2018, 2019, 2020, and the MLQE-PE dataset (Fomicheva et al., 2022). Although the MLQE-PE corpus did not provide explicit references, we utilized the post-edit translations as references. Notably, the MLQE-PE dataset encompassed several low-resource languages, including Nepali, Sinhalese, Pashto, and Khmer. The resulting corpus comprised 1027155 tuples, covering 36 language pairs (languages distribution is shown in Appendix A.1).

In terms of domain, the resulting corpus mainly consists of News and Wikipedia articles. However, due to the nature of these domains, it encompasses a wide range of topics, enhancing the diversity of the data and contributing to the robustness of our estimator model.

Furthermore, we employed XLM-R Large as our encoder model, replacing XLM-R Base. During the fine-tuning process, we made slight adjustments to the hyperparameters, particularly using a lower learning rate for the encoder and implementing layer-wise learning rate decay (see Appendix A.3 for detailed hyperparameters). These modifications were determined to be the best-performing ones based on the MQM annotations from the 2021 News domain, which we used for development.

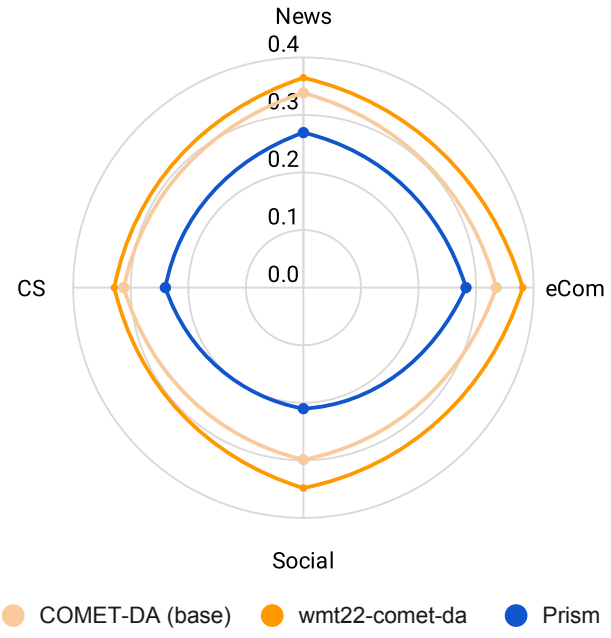


Figure 3.3: Kendall Tau (τ) correlations across the different WMT 2022 shared task domains for `wmt22-comet-da`, our initial model trained on top of XLM-R Base, and PRISM, a strong unsupervised neural baseline.

Figure 3.3 shows the average τ for each domain of WMT 2022 MQM data. We observe that, compared to the previous DA model, our results improved across all domains. For comparison, we also show the results of PRISM, the best-performing unsupervised metric from Table 3.1.

Figure 3.4 contrasts the results of our best model from Table 3.2 and CHRF, the best baseline, with `wmt22-comet-da` on low/mid-resource language pairs. Overall, we observe improvements in all languages, and when compared to CHRF, we can see significant differences for Hindi→Bengali, English→Icelandic, and English→Japanese.

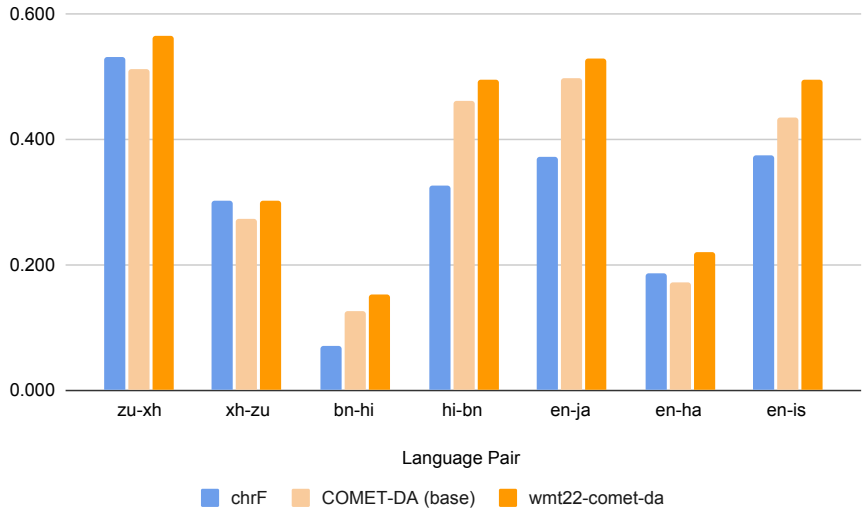


Figure 3.4: Kendall Tau ($\hat{\tau}$) correlations for mid/low-resource language pairs for `wmt22-comet-da`, our initial model trained on top of XLM-R Base, and CHRF, a strong lexical baseline known for its effectiveness on languages with uncommon tokenization.

3.5.1 Robustness to High-Quality MT

For our analysis, we utilized the WMT 2022 News test set and evaluated a subset of the data using the top-performing MT systems for English→German and Chinese→English. We compared our approach against BERTSCORE F1, PRISM, and CHRF. The results are presented in Figure 3.5.

In the case of Chinese→English, we observed that CHRF experienced a drop from 0.195 to 0.149 τ when considering only the top-4 MT systems. This observation supports the findings of (Ma et al., 2019), where lexical metrics demonstrated a significant decrease in performance when considering only the top-4 systems[7].

Interestingly, the neural unsupervised metrics, BERTSCORE F1 and PRISM, demonstrated stable correlations with PRISM even exhibiting an increase in performance when evaluating the top-4 systems. Notably, when examining COMET (`wmt22-comet-da`), we observed its robust behavior across different data cuts. This indicates its ability to differentiate between two translations, even when they may have minimal quality differences.

[7]It is important to note that the analysis conducted in (Ma et al., 2019) differs from our analysis in this section. They employed system-level Pearson ($\hat{\rho}$), which we consider to be unreliable due to the limited number of data points. Therefore, we focused on segment-level correlations, which provide a larger number of data points for analysis, even

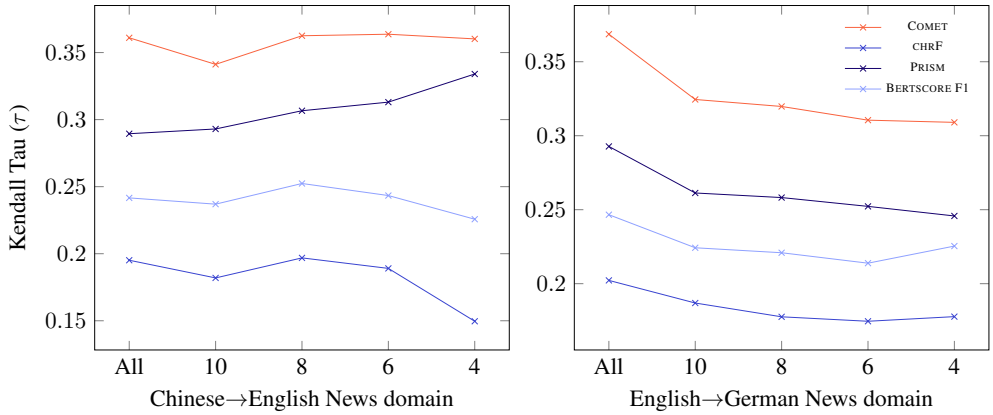


Figure 3.5: Kendall Tau τ performance over the top (10, 8, 6, and 4) performing systems on WMT 22 News Testset. For COMET we used the `wmt22-comet-da` model.

3.6 Comparison to other Neural Fine-tune Metrics

Since the introduction of COMET, several other neural fine-tuned metrics have emerged in the literature. In this section, we will focus on analyzing two popular metrics: BLEURT[8] and UNITE[9]. As discussed in Section 2.2.1.3, the key distinction between these metrics and COMET estimators lies in their approach to encoding. Unlike COMET, which utilizes combined features (as described in Section 3.1.3), BLEURT and UNITE employ joint sentence encoding and leverage full input attention to generate a `[CLS]` embedding that incorporates information from the entire input. Notably, both BLEURT and UNITE use different data for training. While both metrics use fewer DA annotations compared to our most recent model, they augment the data they use with synthetic examples of perfect and imperfect translations.

Table 3.3 presents the correlations achieved by these metrics on the MQM annotations for high-resource language pairs from the WMT 22 Metrics task, while Table 3.4 shows the results for low/mid resource language pairs. UNITE can be used with different input combinations: `[mt; src]` (SRC), `[mt; ref]` (REF), and `[mt; src; ref]` (SRC+REF). We provide results for all of these combinations, as well as the average those scores. Observing the results, we can see that these metrics achieve similar performance, with UNITE showing better correlations on low/mid resource language pairs, particularly when using SRC+REF, while COMET seem to perform better, on average, on the MQM data for high-resource languages.

Apart from performance, considering the architecture, COMET models offer two advantages: Firstly, they can handle longer inputs since they do not have to fit more than one sentence into the 512 max positional embeddings of XLM-R/RemBERT. Secondly, they allow caching of source and reference embeddings, as discussed in Chapter 5, making them more efficient for tasks such as Minimum Bayes Risk (MBR) decoding[10] or scoring multiple systems over the same test set. Additionally, compared to both BLEURT and COMET, UNITE is much slower when ensembling the results from SRC, REF, and SRC+REF, as it requires three forward passes without result reuse.

when considering only the top 4 performing systems.

[8] We used the latest checkpoint BLEURT-20: `https://tinyurl.com/6jee7ts8`

[9] UNITE was reportedly built using our framework, further demonstrating the versatility and adaptability of the COMET framework. We used the UNITE-MUP checkpoint: `https://tinyurl.com/jdjzsbt6`

[10] In Section 7.2 we present work where we test COMET with MBR decoding

Table 3.3: Comparison between different neural fine-tuned metrics on segment-level correlations for WMT 2022 MQM annotations over News, eComerce, Social media, and Customer Support domains (Freitag et al., 2022). The correlation metrics are Pearson (ρ) and Kendall Tau (τ). Results in bold indicate which metrics are top-performing for that specific language pair, domain and metric according to Perm-Both hypothesis test (Deutsch et al., 2021), using 500 re-sampling runs, and setting $p = 0.05$.

				UNITE				COMET
			BLEURT	SRC	REF	SRC+REF	Avg.	wmt22-comet-da
English→German	News	ρ	**0.568**	0.516	0.557	0.553	**0.568**	**0.581**
		τ	**0.380**	0.329	**0.374**	0.371	**0.373**	0.369
	eCom	ρ	**0.444**	0.329	0.404	0.411	0.409	**0.442**
		τ	0.347	0.321	0.366	**0.372**	**0.374**	**0.378**
	Social	ρ	0.430	0.314	0.428	0.422	0.421	**0.461**
		τ	**0.328**	0.248	**0.328**	0.324	0.322	**0.330**
	CS	ρ	**0.467**	0.247	0.457	**0.463**	0.447	0.445
		τ	**0.338**	0.168	0.329	**0.335**	0.308	0.324
English→Russian	News	ρ	0.498	0.492	0.484	0.499	**0.515**	**0.516**
		τ	0.379	0.370	0.375	0.379	**0.392**	**0.391**
	eCom	ρ	**0.553**	0.436	0.534	**0.537**	0.534	**0.539**
		τ	**0.417**	0.350	0.401	0.402	0.399	**0.409**
	Social	ρ	0.398	0.361	0.379	0.379	0.398	**0.417**
		τ	0.348	0.321	0.328	0.331	0.345	**0.366**
	CS	ρ	0.384	0.386	**0.476**	**0.476**	**0.480**	0.431
		τ	0.324	0.303	**0.361**	**0.361**	**0.368**	0.338
Chinese→English	News	ρ	**0.462**	0.375	0.394	0.399	0.418	0.423
		τ	**0.336**	0.295	0.313	0.319	0.331	**0.335**
	eCom	ρ	**0.456**	0.319	0.401	0.410	0.410	0.440
		τ	**0.367**	0.272	0.344	0.351	0.349	0.358
	Social	ρ	**0.420**	0.336	0.385	0.381	0.393	0.410
		τ	**0.360**	0.299	0.349	0.349	0.351	0.353
	CS	ρ	0.363	0.321	0.390	0.386	**0.401**	**0.398**
		τ	0.301	0.253	**0.317**	0.308	**0.318**	**0.322**
	AVG.	ρ	0.454	0.369	0.441	0.443	0.449	**0.459**
		τ	0.352	0.294	0.349	0.350	0.353	**0.356**

However, one disadvantage we found with COMET estimator models is that they are difficult to scale, particularly in the era of large language models. Preliminary experiments replacing XLM-R Large (560M parameters) with XLM-R XL and XXL (3.5B and 10.7B parameters respectively) proved challenging to converge during training, and successful convergence was only achieved by reducing the sentence embedding size with a linear projection that reduces the embedding dimension before passing it to the feed-forward layer. However, this compression results in some information loss, and the resulting model is not better than the one trained with XLM-R Large. This limitation makes it more attractive to develop large-scale metrics using an architecture closer to BLEURT and UNITE and, potentially, distil those metrics into smaller models following COMET architecture.

Table 3.4: Comparison between different neural fine-tuned metrics on segment-level correlations for WMT 2021 DARR over mid and low-resource language pairs. The correlation metric used is the WMT Kendall ($\hat{\tau}$) (Equation 7).

		UNITE				COMET
	BLEURT	SRC	REF	SRC+REF	Avg.	wmt22-comet-da
ZU-XH	0.563	0.484	**0.592**	0.587	0.559	0.566
XH-ZU	**0.364**	0.285	0.360	**0.365**	0.348	0.302
BN-HI	**0.178**	0.134	0.170	0.170	0.167	0.153
HI-BN	0.499	0.512	0.495	0.516	**0.529**	0.495
EN-JA	0.482	0.465	0.529	0.531	**0.543**	0.528
EN-HA	0.186	0.221	0.249	0.257	**0.260**	0.220
EN-IS	0.469	0.436	0.489	0.489	0.488	**0.494**
AVG.	0.392	0.363	0.412	**0.417**	0.413	0.394

In addition to the correlations discussed above, independent studies have also explored the strengths and weaknesses of these three metrics. Amrhein and Sennrich (2022) showed that COMET estimator models were not able to detect errors in numbers and some named entities. This finding was corroborated by Alves et al. (2022), who found similar limitations in BLEURT, while UNITE demonstrated more robustness to these phenomena. Furthermore, Yan et al. (2023) compared these three metrics, along with other embedding-based ones, and observed the presence of *universal adversarial translations* in BLEURT, while COMET and UNITE appeared to be more robust to such phenomena. The authors hypothesized that the presence of such translations in BLEURT could be attributed to the data augmentation techniques used, which might introduce undesirable biases to the model.

3.7 System-level Results

Until now, our focus has primarily been on segment-level correlations, where traditional metrics have shown particularly low performance. However, the main purpose of metrics is not to assess the quality of individual translations, but rather to determine which MT system performs best for a given test set.

To evaluate system-level performance, we adopted a similar setup as described in Section 3.3.2. Instead of Kendall Tau, we report the pairwise accuracy proposed in (Kocmi et al., 2021) (◎). This measure, similar to $\hat{\tau}$ (Eq. 7), quantifies how often a metric agrees with human annotators on determining which system performs better on a given test set.

Table 3.5 presents the system-level results on WMT MQM annotations, comparing the neural fine-tuned metrics discussed in the previous section with CHRF, BERTSCORE (F1), and PRISM. From the results we observe that neural fine-tuned metrics consistently outperform lexical and unsupervised metrics at the system level. Notably, BLEURT, UNITE, and COMET achieve similar performances.

Table 3.5: System-level results for WMT 2022 MQM annotations over News, eComerce, Social media, and Customer Support domains (Freitag et al., 2022). Performance is measured in Pearson (ρ) and Pairwise Accuracy ($\circledcirc$) (Kocmi et al., 2021). Results in bold indicate which metrics are top-performing for that specific language pair and domain according to Perm-Both hypothesis test (Deutsch et al., 2021), using 100 re-sampling runs, and setting $p = 0.05$.

						Neural Fine-tuned		
			chrF	Prism	BERTScore	BLEURT	UniTE	COMET
English→German	News	ρ	0.414	0.506	0.466	0.791	0.819	**0.892**
		$\circledcirc$	0.619	0.686	0.657	0.771	0.743	**0.800**
	Social	ρ	0.659	0.615	0.663	**0.938**	0.930	**0.939**
		$\circledcirc$	0.667	0.676	0.714	**0.924**	0.848	0.838
	eCom	ρ	0.739	0.636	0.727	0.825	0.823	**0.911**
		$\circledcirc$	0.733	0.705	0.733	**0.810**	0.743	**0.800**
	CS	ρ	**0.946**	0.931	0.917	0.917	0.886	**0.944**
		$\circledcirc$	**0.848**	**0.819**	**0.848**	**0.810**	0.705	0.752
English→Russian	News	ρ	**0.907**	0.832	**0.920**	0.550	0.669	0.402
		$\circledcirc$	**0.810**	**0.829**	**0.829**	0.743	0.762	0.733
	Social	ρ	0.835	0.802	0.848	**0.957**	**0.958**	0.919
		$\circledcirc$	0.752	0.762	0.752	**0.933**	**0.924**	0.886
	eCom	ρ	0.795	0.703	0.808	**0.970**	**0.958**	0.950
		$\circledcirc$	0.819	0.800	**0.886**	**0.924**	**0.905**	**0.895**
	CS	ρ	0.817	0.701	0.783	**0.873**	**0.875**	**0.892**
		$\circledcirc$	0.762	0.714	0.762	**0.800**	**0.819**	**0.790**
Chinese→English	News	ρ	0.403	0.669	0.492	0.792	**0.922**	**0.899**
		$\circledcirc$	0.619	**0.743**	0.676	**0.762**	0.724	0.705
	Social	ρ	0.750	0.880	0.806	0.896	**0.963**	0.943
		$\circledcirc$	0.752	0.857	0.829	**0.924**	**0.895**	0.867
	eCom	ρ	0.396	0.497	0.450	0.760	0.766	**0.793**
		$\circledcirc$	0.533	0.610	0.562	**0.705**	**0.686**	**0.686**
	CS	ρ	0.742	0.487	0.482	0.736	**0.896**	**0.904**
		$\circledcirc$	**0.695**	0.667	**0.705**	**0.705**	**0.695**	0.638
	Avg.	ρ	0.700	0.688	0.697	0.834	**0.872**	0.866
		$\circledcirc$	0.717	0.739	0.746	**0.817**	0.787	0.783

3.8 How Far Can We Go Without References?

Throughout the years of developing the COMET framework, we have built several reference-free models to participate in the "QE-as-a-metric" subtask of the Metrics task (Rei et al., 2020b, 2021a) or directly in the QE shared task (Zerva et al., 2021; Rei et al., 2022c). Surprisingly, some of these models have demonstrated excellent performance in metrics shared tasks (Mathur et al., 2020b; Freitag et al., 2021b, 2022) and even won the QE shared task in 2022 (Zerva et al., 2022a). Furthermore, an independent study conducted by Microsoft revealed that a reference-free COMET model ranked as the second-best metric in terms of system accuracy ($\circledcirc$) across 101 different languages and 232 translation directions.

Our best reference-free model follows a similar architecture to OPENKIWI (Section 2.2.2), but utilizes the same data and hyperparameters as the `wmt22-comet-da` metric. Consequently, we named this model `wmt22-cometkiwi-da`. It is worth noting that running UNITE on SRC inputs shares the same architecture as `wmt22-cometkiwi-da`. By referring to Table 3.3 and Table 3.4, we can already gain insights into the competitiveness of a reference-free evaluation compared to state-of-the-art MT Metrics. An ensemble of such models was used to secure first place in the WMT 2022 QE shared task (Rei et al., 2022c). In this thesis, we scale that model from 560M parameters to 3.5B and 10.7B parameters using larger versions of XLM-R (Goyal et al., 2021).

Table 3.6: Performance of reference-free models of different scale, ranging from 560M parameters to 10.7B, measured by segment-level correlations for WMT 2022 MQM annotations over News, eComerce, Social media, and Customer Support domains (Freitag et al., 2022). We used our best reference-base metric `wmt22-comet-da` as baseline. The correlation metrics are Pearson (ρ) and Kendall Tau (τ). Results in bold indicate which metrics are top-performing for that specific language pair, domain and metric according to Perm-Both hypothesis test (Deutsch et al., 2021), using 500 re-sampling runs, and setting $p = 0.05$.

			COMET	COMETKIWI		
			wmt22-comet-da (580M)	large (560M)	xl (3.5B)	xxl (10.7B)
English→German	News	ρ	**0.581**	0.546	0.542	0.548
		τ	0.369	0.308	0.334	0.330
	Social	ρ	**0.442**	0.430	0.431	**0.464**
		τ	**0.378**	**0.374**	0.358	0.364
	eCom	ρ	**0.461**	0.324	0.368	0.423
		τ	**0.330**	0.236	0.276	0.313
	CS	ρ	0.445	0.301	0.365	0.388
		τ	0.324	0.154	0.206	0.217
English→Russian	News	ρ	0.516	**0.538**	**0.543**	0.531
		τ	0.391	0.401	0.410	**0.420**
	Social	ρ	**0.539**	0.517	0.537	**0.552**
		τ	0.409	0.372	0.418	**0.438**
	eCom	ρ	**0.417**	0.344	0.392	**0.415**
		τ	**0.366**	0.323	0.356	**0.370**
	CS	ρ	**0.431**	0.324	0.413	**0.422**
		τ	0.338	0.288	0.339	**0.368**
Chinese→English	News	ρ	**0.423**	0.401	0.437	**0.434**
		τ	**0.335**	0.326	**0.341**	0.333
	Social	ρ	**0.440**	0.372	0.355	0.399
		τ	**0.358**	0.316	0.301	0.324
	eCom	ρ	**0.410**	0.378	0.360	**0.408**
		τ	**0.353**	0.344	0.327	**0.353**
	CS	ρ	**0.398**	0.355	0.334	0.372
		τ	**0.322**	0.293	0.262	0.298
AVG.		ρ	**0.459**	0.403	0.423	0.446
		τ	**0.356**	0.311	0.327	0.344

Table 3.7: Performance of reference-free models of different scale, ranging from 560M parameters to 10.7B, measured by segment-level correlations for WMT 2021 DARR over mid and low-resource language pairs. The correlation metric used is the WMT Kendall ($\hat{\tau}$) (Equation 7). Results are averaged over 500 re-sampling runs

	COMET	COMETKIWI		
	`wmt22-comet-da (580M)`	`large (560M)`	`xl (3.5B)`	`xxl (10.7B)`
ZU-XH	**0.566**	0.455	0.516	**0.559**
XH-ZU	0.302	0.248	0.321	**0.365**
BN-HI	**0.153**	0.137	**0.144**	0.129
HI-BN	0.495	**0.529**	0.479	0.506
EN-JA	0.528	0.520	0.507	**0.539**
EN-HA	0.220	0.159	**0.256**	**0.253**
EN-IS	0.494	0.454	0.469	**0.510**
AVG.	0.394	0.357	0.385	**0.409**

Reference-free evaluation is competitive with reference-based evaluation even for the same compute budget. From Table 3.6, we can observe that the average τ of `wmt22-cometkiwi-da` is $\tau = 0.311$, which is comparable to $\tau = 0.356$ achieved by `wmt22-comet-da` and far better than the results of strong reference-based baselines such as PRISM and BERTSCORE-F1 with $\tau = 0.25$ and $\tau = 0.216$, respectively (see Table 3.1).

Large-scale reference-free models can outperform SOTA metrics such as `wmt22-comet-da`. When scaling the size of the encoder model from Large to XXL, we were able to further improve the results of our reference-free model. In Table 3.6, we can see that in many cases, the XXL model is competitive with `wmt22-comet-da` on the English→Russian language pair. If we recall Table 3.3, we see that our XXL model is able to match the performance of UNITE on REF and SRC+REF inputs. Surprisingly, on low/mid resource language pairs, the XXL model is able to outperform BLEURT and achieve a performance close to those of UNITE SRC+REF, which is the best-performing metric for the DARR data.

We note that the metrics against which we compare our XXL QE model could also be built using larger encoders such as XLM-R XXL. Our point is not that reference-free evaluation is superior to reference-based evaluation, but rather that we can successfully build evaluation models that rely solely on source information and far exceed the performance of popular metrics such as BLEU or BERTSCORE-F1. With sufficient parameters, these models can even perform as well as well-known SOTA metrics such as COMET and BLEURT, demonstrating the potential of reference-free evaluation in achieving state-of-the-art results.

3.9 The Importance of Source Information

As we have observed, source-based evaluation can be competitive with reference-based evaluation. For instance, in Tables 3.3 and 3.4, we can observe that combining reference information with source information increases correlations for the UNITE metric. Additionally, while reference translations can be difficult to obtain, they can also be of low quality and introduce undesirable

biases that decrease the precision of metrics (Freitag et al., 2020).

In this section, we will highlight the importance of using the source in MT metrics. We will begin by demonstrating the impact that a poor reference can have on the overall performance of a metric. Subsequently, we will explore how source information can be leveraged to address the challenge of ambiguous translations.

3.9.1 Impact of a Low-Quality Reference

In the WMT 2021 Metrics Shared Task, an additional domain (other than News) was introduced, namely TED talks. The transcripts of these talks were extracted from OPUS[11] and released by Reimers and Gurevych (2020). While the talks were originally in English, they were translated into multiple languages by volunteers, resulting in Chinese→English translations that were originally English→Chinese. To ensure natural-sounding translations, the organizers asked a Chinese speaker to select the talks where the source was considered "natural sounding" (Freitag et al., 2021b). However, no quality control was applied to the English translations, and after collecting MQM annotations, it was discovered that the English reference (Ref.A) contained numerous errors. To address this issue, professional translators were enlisted to provide high-quality translations, resulting in a second reference (Ref.B). This unfortunate incident highlights the challenges of finding reliable references and the potential pitfalls of blindly using references from publicly available parallel corpora.

We argue that one of the major advantages of incorporating the source in the evaluation process is the increased robustness against noisy references. To demonstrate this, we compare the performance of UNITE[12] and COMET with and without source information when evaluating against these two references. While COMET typically expects a source input, we created a modified version that does not use the source[13]. The results are presented in Figure 3.6.

Reference-free evaluation outperforms evaluating with a poor reference. Examining the results in Figure 3.6a, we observe that the correlation for UNITE SRC is higher compared to UNITE REF when using Ref.A. Although the use of SRC+REF appears to improve the results, the improvement is not statistically significant compared to using the source alone.

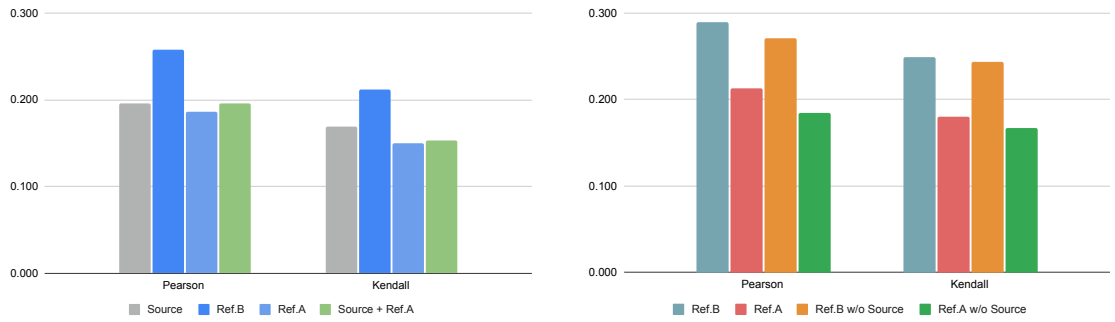
Using a poor reference has a greater impact on models that rely solely on the reference. We can observe that using Ref.A significantly affects both UNITE and COMET. When using Ref.A instead of Ref.B, UNITE REF experiences a drop in Pearson correlation from $\tau = 0.258$ to $\tau = 0.186$, while COMET (without source) drops from $\tau = 0.270$ to $\tau = 0.185$. However, even when using a poor reference, UNITE SRC+REF achieves a correlation of $\tau = 0.196$. Similarly, COMET (`wmt22-comet-da`) exhibits the same trend, with a correlation of $\tau = 0.213$.

Performance of COMET using the source surpasses COMET without the source. The `wmt22-comet-da` model achieves higher Pearson ($\rho = 0.289$) and Kendall ($\tau = 0.249$) correlations

[11] `https://opus.nlpl.eu/TED2020.php`

[12] UniTE-MUP checkpoint: https://huggingface.co/Unbabel/unite-mup

[13] Our modified estimator closely follows the architecture of RUSE, utilizing XLM-R Large as the encoder and trained with the same data and hyperparameters as `wmt22-comet-da`



(a) MQM correlations when using UNITE SRC, REF, and SRC+REFwhen evaluating against Ref.B and Ref.A

(b) MQM correlations when using COMET with and without source input when evaluating against Ref.B and Ref.A

Figure 3.6: Impact of low-quality references on neural fine-tuned metrics with and without source input. Correlations are measured using WMT 21 TED Talk MQM annotations with reference B (Ref.B) and reference A (Ref.A). While Ref.B has an MQM score of 0.42 (less than a minor error per sentence on average), Ref.A has an MQM score of 5.52 (on average, a major error per sentence).

compared to the reference-only implementation of COMET ($\rho = 0.270$ and $\tau = 0.244$)[14]. These findings align with the results for UNITE presented in Tables 3.3 and 3.4 and the results reported in (Rei et al., 2020a) regarding the Translation Ranking Model.

3.9.2 Dealing with Ambiguous Translation

Word Sense Disambiguation	**Discourse Connectives**	**Names Gender**
Src. (de): Was heisst "**Brühe**"? Ref. (en): What does "**stock**" mean? ✓: What does "**vegetable stock**" mean? ✗: What does "**penny stock**" mean?	Src (fr): Aucun test de qualité de l'air n'ait été réalisé dans ce bâtiment **depuis** notre élection. Ref (en): No air quality test has been done on this particular building **since** we were elected. ✓: No air quality test has been done on this particular building **from the time** we were elected. ✗:No air quality test has been done on this particular building **because** we were elected.	Src (de): Der Manager feuerte **die** Bäcker**in**. Ref (en): The manager fired the baker. ✓: The manager fired the **female** baker. ✗: The manager fired the **male** baker.

Figure 3.7: Examples of different types of ambiguous translations from the ACES challenge set (Amrhein et al., 2022).

Another compelling example that highlights the importance of using the source in MT evaluation is the case of ambiguous translations, where the reference translation may not provide all the necessary information to determine the quality of the translation. A notable study conducted by Amrhein et al. (2022) focused on creating a comprehensive challenge set to assess the robustness of MT metrics when evaluating translations with specific phenomena. One such phenomenon they investigated was ambiguous translations.

Ambiguous translations occur when the source text can be translated in multiple ways, resulting in different possible interpretations. The ACES challenge set examined various types of

[14]Results are statistically significant according to the Perm-Both hypothesis test using 500 re-sampling runs.

ambiguity, including gender, word sense, and discourse connectives. Figure 3.7 showcases examples of different types of ambiguous translations from the ACES paper.

To investigate the performance of MT metrics in handling ambiguous translations, we examined the results of COMET, UNITE, and COMETKIWI using the WMT Kendall Tau $\hat{\tau}$ metric (Eq. 7). The results are presented in Table 3.8.

Table 3.8: Comparison between neural fine-tuned metrics with and without source input on Ambiguous Translations from the ACES challenge set. Results are measured in terms of WMT Kendall Tau ($\hat{\tau}$) (Eq. 7).

	Ambiguous Translation			
	Word Sense Disambiguation	Discourse Connectives	Gender	AVG.
COMETKIWI	**0.524**	**0.611**	0.581	0.572
COMET	0.066	0.194	0.77	0.112
UNITE SRC	0.454	0.505	**0.716**	**0.559**
UNITE REF	0.000	0.187	0.000	0.062
UNITE SRC+REF	0.030	0.230	0.142	0.134
UNITE Avg.	0.163	0.279	0.539	0.327

The results shown in Table 3.8 further support our findings. When reference information is poor, such as in the case of ambiguous translations, reference-free evaluation performs better. Surprisingly, even though both COMET and UNITE SRC+REF use source information, they perform much worse than reference-free models like COMETKIWI. This indicates that when a reference is used, these models place more weight on the reference information. It aligns with the behavior of COMET, which intentionally incorporates more reference information into the feed-forward estimator. In the case of UNITE, this behavior is learned during training.

3.10 Conclusion

In this chapter, we have demonstrated the effectiveness of the COMET framework in constructing neural fine-tuned metrics that achieve high correlations with human judgments across various languages and domains. We have introduced `wmt22-comet-da`, a robust metric built using 1 million direct assessments across 36 language pairs. Through comparisons with other metrics, we have once again established the superiority of neural fine-tuned metrics over previous approaches.

Furthermore, we have presented several reference-free metrics ranging from 560 million to 10.7 billion parameters, showcasing the potential for state-of-the-art evaluation without the need for reference translations, given sufficient computational resources.

Throughout this chapter, we have also emphasized the importance of incorporating source information in metrics. By highlighting the value of source information, we have demonstrated its contribution to more accurate and robust evaluation.

Overall, the findings in this chapter solidify the value of neural fine-tuned metrics, reinforce the significance of leveraging source information, and pave the way for advanced evaluation methods that align with human judgments.

Chapter 4

Towards Interpretable MT Evaluation Neural Metrics

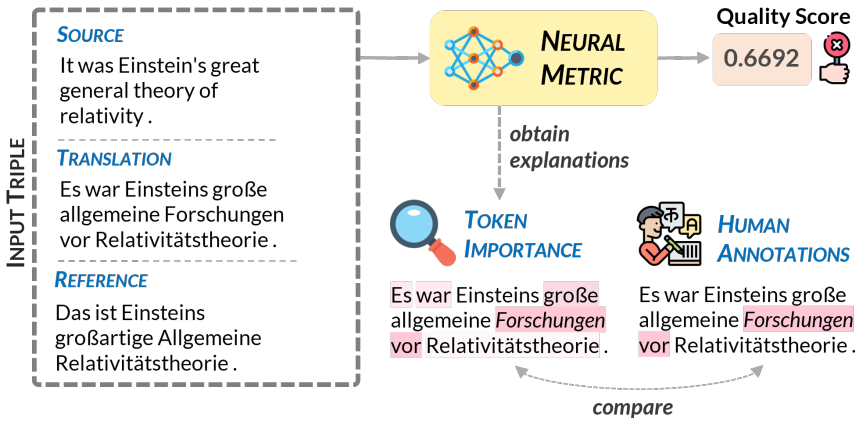


Figure 4.1: Illustration of our approach. In this example, the metric assigns the translation a low score. We aim to better understand this sentence-level assessment by examining the correspondence between our token-level explanations and human annotated error spans.

In Chapter 3, we explored the effectiveness of neural metrics for evaluating MT. These metrics have demonstrated significant improvements in correlating with human judgments compared to traditional metrics like BLEU, which rely on lexical overlap. However, a drawback of neural metrics is their inherent opacity, as they operate as "black boxes", providing a single sentence-level score without revealing the underlying decision-making process.

To address this limitation, this chapter proposes a framework for investigating how these metrics utilize token-level information. Our aim is to shed light on the inner workings of neural metrics by comparing their results with MQM annotations and synthetically-generated critical translation errors. By analyzing the token-level neural saliency maps, we can attribute the information leveraged by these metrics directly to translation errors. Figure 4.1 illustrates our framework.

The primary objective of our framework is to provide explanations for sentence-level quality assessments in reference-based metrics by producing token-level explanations that align with translation errors. However, since recent metrics also incorporate source information, we intend

to further investigate the significance of various input types, particularly the utilization of source information, to gain a better understanding of their impact on the final metric.

The work presented in this chapter was showcased at the 61st Annual Meeting of the Association for Computational Linguistics (ACL 2023) and builds upon the findings from our participation in the shared task for explainable Quality Estimation (QE) (Fomicheva et al., 2021; Treviso et al., 2021). We will commence by providing a brief background and overview of our submission to the shared task, followed by the presentation of our ACL 2023 work titled "The Inside Story" (Rei et al., 2023).

4.1 Background: Explainable QE Shared Task

Before delving into our submission to this shared task, let us provide a brief overview of the setup as described in Section 2.2.5. In the first edition of the Explainable QE shared task (Fomicheva et al., 2021), the organizers adapted the *Appraise* platform to incorporate human annotators' rationales for their score decisions, in addition to the sentence-level scores. Annotators were specifically asked to highlight the words in translated sentence corresponding to translation errors that justified the assigned sentence score. The test data included four language pairs: Estonian→English, Romanian→English, Russian→German, and German→Chinese. For training, participants were provided with the MLQE-PE corpus (Fomicheva et al., 2022), restricting models in the *constrained* task to only use this corpus.

In our participation to the shared task we extensively explored various explainability methods to determine which ones show promise in extracting explanations from QE systems like `wmt22-cometkiwi-da`. Specifically, we investigated rationalizers, attention mechanisms, gradient-based approaches, and perturbation-based methods:

Attention-based methods. Since the backbone of our models consists of pre-trained multilingual transformers, we studied their main component, the multi-head attention mechanism, expecting to find interpretability patterns that assign higher scores to words associated with translation errors. We extracted the following explanations from the multi-head attention mechanism:

- **Attention weights:** For a source sentence with N tokens and a translation with M tokens, we compute the average of the attention matrix $\mathbf{A}$ row-wise for all heads in all layers. This results in a total of 384 explanation vectors $\boldsymbol{a} \in \mathbb{R}^{N+M}$ for XLM-R-based models and 576 explanation vectors for RemBERT-based models ($32 \times 18 = 576$).
- **Cross-attention weights:** by manual inspection of attention weights, we noticed that some attention heads learn plausible connections from source-to-hypothesis and hypothesis-to-source. Therefore, instead of computing a row-wise average of the entire attention matrix, we average only cross-alignment rows.[1]
- **Attention $\times$ Norm:** following the findings of (Kobayashi et al., 2020), we scale attention weights by the norm of value vectors $\left\|\boldsymbol{V}\boldsymbol{W}_h^V\right\|_2$, where $\boldsymbol{V}$ is the transformer attention value matrix and $\boldsymbol{W}_h^V$ is the corresponding learned linear transformation.

[1]Note that we can get cross-attentions from XLM-R and RemBERT by selecting only the words of the source that attend to the hypothesis and vice-versa.

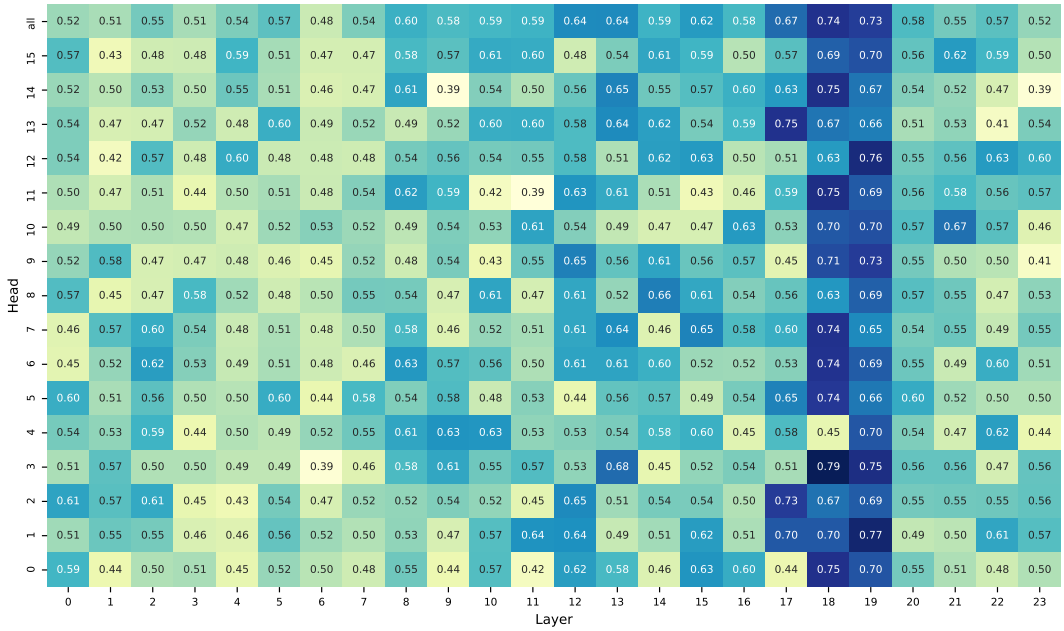


Figure 4.2: Target AUC of different attention heads at each layer of our XLM-R model for Romanian→English. The last tick on the y-axis represents the average of all attention heads.

Gradient-based methods. Explanations extracted by storing gradients computed during the backward propagation is a standard tool used to interpret NLP models. For this shared task, we investigate the following gradient-based methods:[2]

- **Gradient × Hidden States:** we compute gradients w.r.t. the hidden states of each layer, and multiply the resultant vectors by the hidden state vectors themselves: $\nabla_{\boldsymbol{H}_i} \times \boldsymbol{H}_i \in \mathbb{R}^{N+M}$, for $0 \leq i \leq L+1$.
- **Gradient × Attention:** the same as before, but we use the output of the multi-head attention module instead of the hidden states.
- **Integrated Gradients**: we extract integrated gradient explanations w.r.t. the hidden states of each layer. We use a zero-vector as the baseline. We map gradients to explainability scores by normalizing them by their L2 norm and summing the hidden dimensions: $\mathbf{1}^\top \nabla_{\boldsymbol{H}_i} / \left\| \nabla_{\boldsymbol{H}_i} \right\|_2$.

Perturbation-based methods. As baselines, we also extracted explanations using **LIME** (Ribeiro et al., 2016) and a **leave-one-out** strategy, where we replace the `erased` token by the `<mask>` token, which is used for the masked-language model training of XLM-R and RemBERT.

Rationalizers. We append a differentiable binary mask layer (Bastings et al., 2019) on top of the XLM-R model in order to select which tokens are passed on for an estimator for the prediction of a sentence-level score. For each instance, we take the model representations and pass it to an encoder module, in which we sample a binary mask $\boldsymbol{z} \in [0,1]^{N+M}$ from a relaxed Bernoulli distribution

[2]Our implementation is based on Captum: `https://captum.ai/`

(Maddison et al., 2017; Jang et al., 2017), and pass $\boldsymbol{z} \odot [\boldsymbol{s}; \boldsymbol{t}]$ to an estimator module, which re-embeds the masked input and pass it to a linear output layer. Therefore, good explanations $\boldsymbol{z}$ will aid the estimator in producing good sentence-level scores. In training time, the parameters of the encoder and the estimator are jointly trained. In test time, we do not sample the binary masks. Instead, we use the relaxed Bernoulli distribution probabilities as explanations.

Regarding the models used in all these experiments, we trained two models: one based on RemBERT and another based on XLM-R. It is important to note that, as mentioned earlier, for this shared task, we were limited to using the MLQE-PE corpus (Fomicheva et al., 2022). Therefore, our models cannot be directly compared with `wmt22-cometkiwi-da`.

Evaluation. The explanations are evaluated by comparing them to the ground-truth word-level labels using Area Under the Curve (AUC) and Recall at Top-K (Recall@K)[3].

4.1.1 Findings

#	ENCODER	EXPLAINER	AUC	R@K
1	XLM-R	Attention - Layer 18 - Head 3	0.7894	0.6054
2	XLM-R	Attention - Layer 18 - Head 0	0.7462	0.5197
3	XLM-R	Cross-attention - Layer 18 - Head 3	0.8066	0.6293
4	XLM-R	Cross-attention - Layer 18 - Head 0	0.7374	0.4883
5	XLM-R	Attention × Norm - Layer 18 - Head 3	**0.8136**	**0.6342**
6	XLM-R	Attention × Norm - Layer 19 - Head 2	0.8099	0.6153
7	XLM-R	Gradient × Hidden States - Layer 15	0.6780	0.4044
8	XLM-R	Gradient × Attention - Layer 17	0.7618	0.5628
9	XLM-R	Integrated Gradients - Layer 15	0.6560	0.3853
10	XLM-R	LIME	0.5892	0.3300
11	XLM-R	Leave-one-out	0.5921	0.3567
12	XLM-R	Relaxed-Bernoulli Rationalizer	0.5434	0.2914
15	RemBERT	Attention × Norm - Layer 23	0.7904	0.5723
16	RemBERT	Attention × Norm - Layer 22 - Head 5	0.7167	0.4278
1	Ensemble	(5) + (6) + (15)	**0.8398**	**0.6606**

Table 4.1: Area Under Curve (AUC) and Recall@K on the validation set of Romanian→English.

Attention heads are better alone. We found that some attention heads (mostly at upper layers) learned to focus on words associated with BAD tags, achieving great performance in terms of AUC on the validation set. We show in Figure 4.2 the target AUC of different attention heads per layer as a heatmap for Romanian→English, with darker colors indicating higher results.[4] We can see that attention heads in layers 18 and 19 perform better than other layers in general, and that some attention heads solely outperform the average of all attention heads for all respective layers. For example, the attention head 3 at layer 18 achieves an AUC score of 0.79, while the average of all

[3]Recall@K is calculated only for the subset of translations that contain errors. It represents the proportion of words with the highest attribution that correspond to translation errors, relative to the total number of errors in the annotated error span.

[4]We got similar findings for Estonian→English.

attention heads from layer 18 gets an AUC score of 0.74 (5 points difference). This finding also seem to align with Figure 5.3b where we show that layer 15 to 19 are the most important ones for COMET models.

Attention $\times$ Norm outperforms other explainers. By scaling attention probabilities by the L2 norm of value vectors, we improved the performance further. All of our best results consist of attention-based explainers, with the majority being the explanations that consider the norm of value vectors. We show the results for all our explainers on the validation set of Romanian→English in Table 4.1.[5] When using XLM-R or RemBERT as encoder the results are similar, except that the best explainer comes from different attention heads at different upper layers.

Overall, we observed that attention methods outperform gradient and perturbation methods by a considerable margin, and gradients w.r.t. attention outputs yield better results than gradients w.r.t. hidden states, indicating that the information stored in attention heads is valuable.

4.2 Explanations via Attribution Methods

Building upon the previous findings for QE systems, we extend them to the reference-based scenario. Specifically, we employ the following techniques to extract explanations:

- **embed–align:** the maximum cosine similarity between each translation token embedding and the reference and/or source token embeddings (Tao et al., 2022). It was not previously explored in our research, but it has been shown to be a top-performing method used in explainable QE shared tasks proposed by other authors;
- **grad** ℓ_2: the ℓ_2-norm of gradients with respect to the word embeddings of the translation tokens (Arras et al., 2019);
- **attention**: the attention weights of the translation tokens for each attention head of the encoder (Treviso et al., 2021);
- **attn** $\times$ **grad**: the attention weights of each head scaled by the ℓ_2-norm of the gradients of the value vectors of that head (Rei et al., 2022c).

4.3 Experimental Setting

In our experimental setting, we aim to analyze the effectiveness of the explainability methods outlined in Section 4.2 for identifying translation errors and understanding the performance of COMET and UNITE. To achieve this, we will use two datasets: one with MQM annotations performed by experts, and another one with synthetically-generated critical errors.

MQM annotations. We use MQM annotations from the WMT 2021 Metrics shared task (Freitag et al., 2021b),[6] covering three language pairs — English-German (en→de), English-Russian

[5] Results for other language pairs such as Estonian→English follow the same trend.

[6] https://github.com/google/wmt-mqm-human-evaluation

(en→ru), and Chinese-English (zh→en) —in two different domains: News and TED Talks. For each incorrect translation, human experts marked the corresponding error spans. In our framework, these error spans should align with the words that the attribution methods assign higher importance to.

SMAUG perturbations. Publicly available MQM data consists primarily of high quality translations, with the majority of annotated errors being non-critical. However, it is important to assess whether our explanations can be accurately attributed to critical errors, as this may reveal potential metric shortcomings. To this end, we employ SMAUG (Alves et al., 2022)[7], a tool designed to generate synthetic data for stress-testing metrics, to create corrupted translations that contain critical errors. Concretely, we generate translations with the following pathologies: negation errors, hallucinations via insertions, named entity errors, and errors in numbers.[8]

Models. For COMET, we utilize the latest publicly available model: `wmt22-comet-da` (Section 3.5). To ensure a comparable setup, we train our own UNITE model using the same data as COMET. The resulting UNITE model is on par with COMET and the original version, which we examined in Section 2.2.1.3. The full list of hyperparameters is provided in AppendixA.

4.4 Results

4.4.1 High-level analysis

Explanations are tightly related to the underlying metric architecture. The results in Table 4.2 show that the predictive power of the attribution methods differ between UNITE and COMET: attn $\times$ grad is the best method for UNITE-based models, while embed–align works best for COMET. This is expected as UNITE constructs a joint representation for the input sentences, thus allowing attention to flow across them; COMET, in contrast, encodes the sentences separately, so it relies heavily on the separate contextualized embeddings that are subsequently combined via element-wise operations such as multiplication and absolute difference. Interestingly, embed–align and attn $\times$ grad were the winning explainability approaches of the WMT 2022 Shared-Task on Quality Estimation (Zerva et al., 2022a). This suggests that explainability methods developed for QE systems can translate well to reference-based metrics.

Reference information boosts explainability power. Table 4.2 also shows that, across all metrics, using reference information brings substantial improvements over using only the source information. Moreover, while reference-based attributions significantly outperform source-based attributions, combining the source and reference information to obtain token-level attributions does not consistently yield superior results over using the reference alone. Notably, the best attribution method for COMET does not require any source information. This is interesting: in some cases, reference-based metrics may largely ignore source information, relying heavily on the reference instead.

[7] `https://github.com/Unbabel/smaug`

[8] We corrupt fully correct translations that are not an exact copy of the reference translation. Moreover, as the full suit of SMAUG transformations can only be applied to English data, we focus solely on zh→en translations. Overall, the synthetic dataset consists of 2610 translations.

METRIC	EXPLAINABILITY METHOD	en→de AUC	en→de R@K	zh→en AUC	zh→en R@K	en→ru AUC	en→ru R@K	Avg. AUC	Avg. R@K
				src-only evaluation*					
UNITE SRC	embed–align	0.587	0.339	**0.644**	0.281	0.583	0.167	0.604	0.262
	grad ℓ_2	0.572	0.293	0.535	0.200	0.620	0.169	0.576	0.221
	attention	0.636	0.322	0.612	0.253	0.612	0.189	0.620	0.254
	attn × grad	**0.707**	**0.376**	0.639	**0.294**	**0.633**	**0.211**	**0.660**	**0.294**
				ref-only evaluation					
UNITE REF	embed–align[mt, ref]	0.658	0.396	0.667	0.328	0.635	0.218	0.653	0.314
	grad ℓ_2	0.596	0.319	0.571	0.260	**0.661**	0.202	0.609	0.261
	attention	0.637	0.344	**0.670**	0.335	0.652	0.224	0.653	0.301
	attn × grad	**0.725**	**0.425**	0.667	**0.380**	0.660	**0.248**	**0.684**	**0.351**
				src, ref joint evaluation					
UNITE SRC+REF	embed–align[mt, src; ref]	0.650	0.383	0.670	0.330	0.618	0.213	0.646	0.309
	grad ℓ_2	0.595	0.325	0.579	0.257	0.643	0.191	0.606	0.257
	attention	0.657	**0.421**	0.670	**0.383**	0.649	0.223	0.659	0.342
	attn × grad	**0.736**	**0.421**	**0.674**	**0.383**	**0.671**	**0.248**	**0.693**	**0.351**
COMET	embed–align[mt, src]	0.590	0.371	0.674	0.314	0.577	0.220	0.614	0.301
	embed–align[mt, ref]	0.694	**0.425**	0.696	0.355	0.647	0.275	**0.679**	**0.352**
	embed–align[mt, src; ref]	0.688	0.416	**0.697**	**0.357**	0.622	**0.279**	0.669	0.350
	grad ℓ_2	0.603	0.312	0.540	0.252	0.604	0.185	0.582	0.250
	attention	0.604	0.351	0.592	0.259	0.633	0.209	0.608	0.268
	attn × grad	**0.710**	0.365	0.633	0.278	**0.662**	0.244	0.669	0.295

Table 4.2: AUC and Recall@K of explanations obtained via different attribution methods for COMET and UNITE models on the MQM data. *Although UNITE SRC is a *src-only evaluation* metric, it was trained with reference information (Wan et al., 2022).

Explanations identify critical errors more easily than non-critical errors. Figure 4.3 shows that explanations are more effective in identifying critical errors compared to other non-critical errors (see Table 4.2). Specifically, we find significant performance improvements up to nearly 30% in Recall@K for certain critical errors. Overall, hallucinations are the easiest errors to identify across all neural metrics. This suggests that neural metrics appropriately identify and penalize hallucinated translations, which aligns with the findings of (Guerreiro et al., 2023). Moreover, explanations for both UNITE models behave similarly for all errors except numbers, where the source information plays a key role in improving the explanations. Notably, contrary to what we observed for data with non-critical errors, COMET explanations are less effective than those of UNITE REF and UNITE SRC+REF for identifying critical errors.

Explanations can reveal potential metric weaknesses. Figure 4.3 suggests that COMET explanations struggle to identify localized errors (negation errors, named entity errors and discrepancies in numbers). We hypothesize that this behavior is related to the underlying architecture. Unlike UNITE-based metrics, COMET does not rely on soft alignments via attention between the sentences in the encoding process. This process may be key to identify local misalignments during the encoding process. In fact, the attention-based attributions for UNITE metrics can more easily identify these errors. COMET, however, encodes the sentences separately, which may result in grammatical features (e.g. numbers) being encoded similarly across sentences (Chi et al., 2020; Chang et al., 2022). As such, explanations obtained via embedding alignments will not properly identify these misalignments on similar features. Importantly, these findings align with

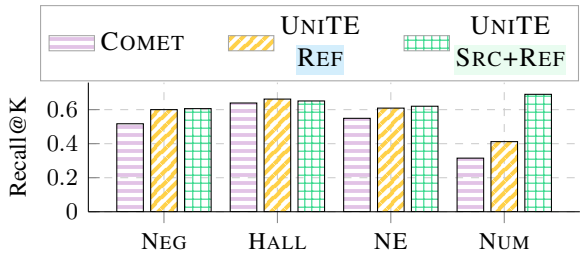


Figure 4.3: Performance of the best attribution methods for COMET, UNITE REF and UNITE SRC+REF in terms of Recall@K on translations with critical errors: negations (NEG), hallucinations (HALL), named entity errors (NE), and errors in numbers (NUM).

observations made in (Amrhein and Sennrich, 2022; Raunak et al., 2022). This showcases how explanations can be used to diagnose and reveal shortcomings of neural-based metrics.

4.5 Comparison between COMET and XLM-R Alignments

We have observed from previous tables that COMET achieves the best explanations when using alignments. However, since COMET is fine-tuned, a lingering question remains: How do these alignments compare to the alignments obtained from the underlying encoder of COMET before the fine-tuning process? To address this question, we conducted an experiment comparing the results obtained directly from XLM-R.

Table 4.2 clearly demonstrates that the alignments between the reference and/or source and the translation provide effective explanations for COMET. This leads us to examine how these alignments compare to the alignments obtained from XLM-R without any fine-tuning, as shown in Table 4.3.

Overall, the explanations derived from the alignments of COMET are found to be more indicative of error spans compared to those obtained from XLM-R alignments. This suggests that during the fine-tuning phase, COMET modifies the underlying XLM-R representations to achieve improved alignment with translation errors.

METRIC	EXPLAINABILITY METHOD	en→de		zh→en		en→ru		Avg.	
		AUC	R@K	AUC	R@K	AUC	R@K	AUC	R@K
XLM-R	embed–align$^{[mt, src]}$	0.587	0.359	0.668	0.311	0.576	0.199	0.610	0.289
	embed–align$^{[mt, ref]}$	0.671	0.405	0.689	0.345	0.634	0.244	0.664	0.331
	embed–align$^{[mt, src; ref]}$	0.666	0.395	0.690	0.347	0.616	0.242	0.657	0.328
COMET	embed–align$^{[mt, src]}$	0.590	0.371	0.674	0.314	0.577	0.220	0.614	0.301
	embed–align$^{[mt, ref]}$	**0.694**	**0.425**	0.696	0.355	**0.647**	0.275	**0.679**	**0.352**
	embed–align$^{[mt, src; ref]}$	0.688	0.416	**0.697**	**0.357**	0.622	**0.279**	0.669	0.350

Table 4.3: AUC and Recall@K of explanations obtained via alignments for COMET and XLM-R without any further fine-tuning on human annotations.

4.6 COMET Explanation Examples

In Figures 4.4 and 4.5, we show examples of COMET explanations for Chinese→English and English→German language pairs, respectively. We highlight in gray the corresponding MQM annotation performed by an expert linguist and we sort the examples from highest to lowest COMET scores. From these examples we can observe the following:

- Highlights provided by COMET explanations have a high recall with human annotations. In all examples, subword tokens corresponding to translation errors are highlighted in red but we often see that not everything is incorrect.
- Explanations are consistent with scores. For example, in the third example from Figure 4.4, the red highlights do not correspond to errors and in fact the translation only has a major error griffen. Nonetheless, the score assigned by COMET is a low score of 0.68 which is faithful to the explanations that was given even if the assessment does not agree with human experts.

4.7 Conclusion

In this chapter, our investigation focused on the application of explainability methods to gain a deeper understanding of neural fine-tuned metrics, specifically COMET and UNITE. Our goal was to explore how these metrics utilize token-level information to derive a sentence-level score, with a particular emphasis on their alignment with human annotations such as MQM.

Through our analysis, we discovered the impact of reference information on the quality of explanations and how these explanations can help identify the limitations of these metrics. We found a strong correlation between the architecture of the underlying metric and the quality of its explanations. Notably, our findings support the notion that COMET relies heavily on reference information rather than source information, which aligns with the results obtained in Section 3.9.2.

Furthermore, our investigation unveiled how explanations can shed light on the weaknesses of reference-based metrics, particularly in their failure to appropriately identify some categories of errors such as numbers, where we observed a lower Recall@K for COMET. This insight offers valuable guidance for further refining these metrics.

Overall, our exploration of explainability methods in this chapter deepens our understanding of the inner workings of neural fine-tuned metrics and reveals crucial aspects for enhancing their performance and effectiveness.

Source:
And yet, the universe is not a silent movie because the universe isn't silent.

Translation:
Und dennoch ist das Universum kein Stummfilm, weil das Universum nicht still ist.

COMET score: 0.8595
COMET explanation:
_Und _dennoch _ist _das _Univers um _kein _Stu mm film , _weil _das _Univers um _nicht _still _ist .

Source:
And yet black holes may be heard even if they're not seen, and that's because they bang on space-time like a drum.

Translation:
Und dennoch werden Schwarze Löcher vielleicht gehört, auch wenn sie nicht gesehen werden, und das liegt daran, dass sie wie eine Trommel auf die Raumzeit schlagen.

COMET score: 0.7150
COMET explanation:
_Und _dennoch _werden _Schwarz e _Lö cher _vielleicht _gehört , _auch _wenn _sie _nicht _gesehen _werden , _und _das _liegt _daran , _dass _sie _wie _eine _Tro mmel _auf _die _Raum zeit schlagen .

Source:
Ash O'Brien and husband Jarett Kelley say they were grabbing a bite to eat at Dusty Rhodes dog park in San Diego on Thursday, with their three-month-old pug in tow.

Translation:
Ash O'Brien und Ehemann Jarett Kelley sagen, dass sie am Donnerstag im Hundepark Dusty Rhodes in San Diego einen Happen zu essen griffen, mit ihrem drei Monate alten Mops im Schlepptau.

COMET score: 0.6835
COMET explanation:
_Ash _O ' Bri en _und _Ehe mann _Ja rett _Kel ley _sagen , _dass _sie _am _Donnerstag _im _Hunde park _Du sty _Rhod es _in _San _Diego _einen _Happ en _zu _essen _ griff en _ , _mit _ihrem _drei _Monate _alten _M ops _im _Schle ppt au .

Source:
It was Einstein's great general theory of relativity.

Translation:
Es war Einsteins große allgemeine Forschungen vor Relativitätstheorie.

COMET score: 0.6692
COMET explanation:
_Es _war _Einstein s _große _allgemein e _Forschung en _vor _Relativ ität s the ori e .

Source:
There's mask-shaming and then there's full on assault.

Translation:
Es gibt Maskenschämen und dann ist es voll bei Angriff.

COMET score: 0.2318
COMET explanation:
_Es _gibt _Mask en schä men _und _dann _ist _es _voll _bei _Angriff _ .

Figure 4.4: Saliency map for COMET explanation scores for a set of English→German examples. Comparing the token-level explanations with the MQM annotation (highlighted in gray) reveals the source of correspondence between specific token-level translation errors and the resulting scores.

Source:
我想告诉大家 宇宙有着自己的配乐， 而宇宙自身正在不停地播放着。 因为太空可以想鼓一样振动。
Translation:
I want to tell you that the universe has its own iconic soundtrack and the universe itself is constantly playing non-stop because space can vibrate like a drum.
COMET score: 0.8634
COMET explanation:
_I _want _to _tell _you _that _the _univers e _has _its _own _icon ic _soundtrack _and _the _univers e _itself _is _constantly _playing _non - stop _because _space _can _vibra te _like _a _drum .

Source:
另外,吉克隽逸和刘烨作为运动助理,也围绕运动少年制造了不少爆笑话题。
Translation:
In addition, as sports assistants, Ji Kejunyi and Liu Ye have also created a lot of hilarious topics around sports teenagers.
COMET score: 0.8214
COMET explanation:
_In _addition , _as _sports _assistant s , _Ji _Ke ju nyi _and _Li u _Ye _have _also _created _a _lot _of _ hila rious _topic s _around _sports _teenager s .

Source:
一番言论让场上的少年和运动领队们都倒吸一口凉气。
Translation:
The remarks made the teenagers and the sports leaders on the field gasp a sigh of relief.
COMET score: 0.7793
COMET explanation:
_The _re marks _made _the _teenager s _and _the _sports _leaders _on _the _field _gas p _a _sig h _of _relief _ .

Source:
强烈的阳光是如此地刺眼，
Translation:
The intense sunlight is so harsh;
COMET score: 0.7561
COMET explanation:
_The _intense _sun light _is _so _har sh ;

Source:
如今， 我们希望能够 给这部关于宇宙的 宏伟的视觉作品 配上声音。
Translation:
Today, we hope to be able to give this magnificent visual work of the universe a sound.
COMET score: 0.7073
COMET explanation:
_Today , _we _hope _to _be _able _to _give _this _magnific ent _visual _work _of _the _univers e _a _sound .

Figure 4.5: Saliency map for COMET explanation scores for a set of Chinese→English examples. Comparing the token-level explanations with the MQM annotation (highlighted in gray) reveals the source of correspondence between specific token-level translation errors and the resulting scores.

Chapter 5

Searching for COMETINHO: The Little Metric That Could

In this chapter, we describe several experiments that attempt to reduce COMET computational cost and model size to make it more efficient at inference. Our techniques are particularly useful in settings where we have multiple translations from different systems on the same source sentences. Since the models are based on triplet encoders, we will first analyse the impact of *embedding caching* and *length batching*. Then, we will try to further reduce the computational cost by using *weight pruning* and *knowledge distillation*. Our results show that embedding caching and length batching alone can boost COMET performance 39.19% when scoring one system and 65.44% when scoring 8 systems over the same test set. Furthermore, with knowledge distillation we are able to create a model that is 80% smaller and 2.128x faster with a performance close to the original model and above strong baselines such as BERTSCORE and PRISM. Figure 5.1 shows time differences for all proposed methods when evaluating a varying number of systems.

It is worth noting that the experiments in this section were conducted before the development of the `wmt22-comet-da` metric, as presented in Chapter 3. Therefore, all the experiments described here utilized a previous version of the COMET model. The main difference lies in the training data, which only ranged from WMT 17 to WMT 19, and the utilization of different hyperparameters.

This work was published at the 23rd Annual Conference of the European Association for Machine Translation (EAMT 2022) and received the Best Paper Award.

5.1 Length Sorting and Caching

Before exploring approaches that reduce the number of model parameters, we experiment with techniques to optimize the inference time computational load. One which is commonly used is to sort the batches according to sentence length to reduce tensor padding (Pu et al., 2021). Since COMET estimators receive three input texts (source, hypothesis and reference), for simplicity, we do length sorting according to the source length. Figure 5.2a shows the speed difference between an unsorted test set with varying size and length-based sorting.

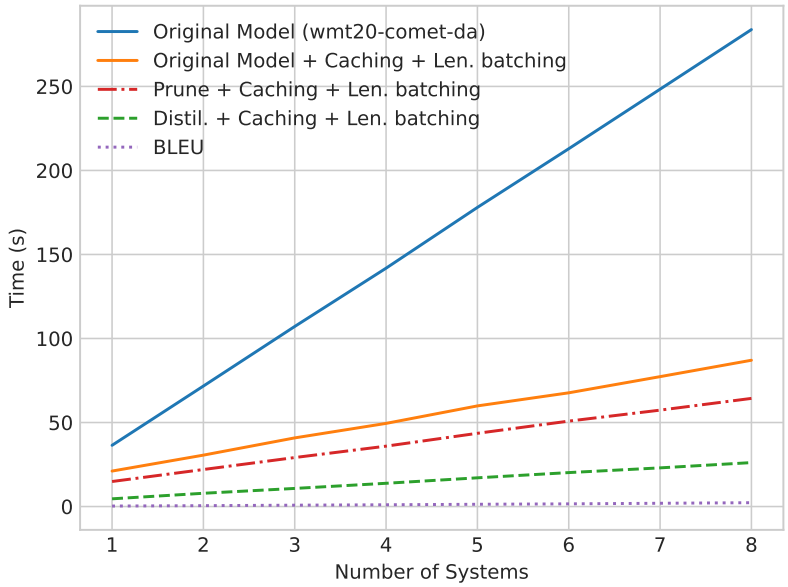


Figure 5.1: Comparison between a COMET estimator with XLM-R Large, that same model with caching and length batching, PRUNE-COMET and DISTIL-COMET. We report the average of 5 runs for each model/metric for a varying number of systems. All experiments were performed using the German→English WMT20 Newstest, with a `NVIDIA GeForce GTX 1080 TI` GPU and a constant batch size of 16. For comparison we also plot the runtime of BLEU in a `Intel(R) Core(TM) i7-6850K CPU @ 3.60GHz`.

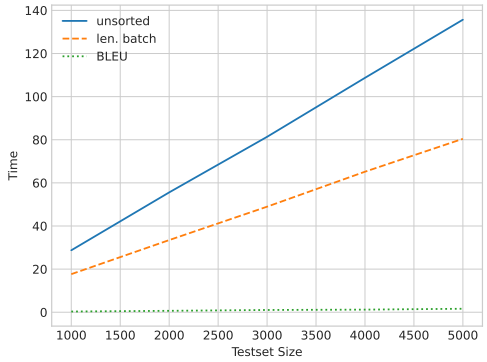
As previously pointed out, estimator metrics are based on triplet encoders[1] which means that the source and reference encoding does not depend on the provided MT hypothesis as opposed to other recent metrics such as BLEURT which have to repetitively encode the reference for every hypotheses.

With that said, using a COMET estimator we only need to encode each unique sentence (source, hypothesis translation or reference translation) once. This means that we can cache previously encoded batches and reuse their representations. In Figure 5.2b, we show the speed gains, in seconds, when scoring multiple systems over the same test set. This reflects the typical MT development use case in which we want to select the best among several MT systems.

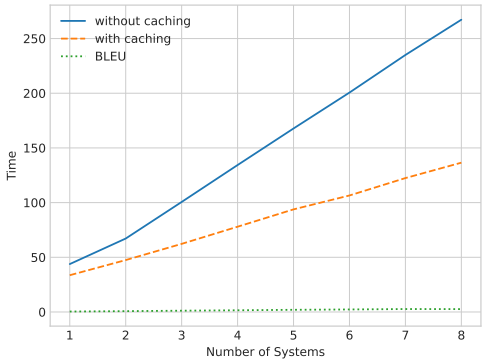
These two optimizations altogether are responsible for reducing model inference time from 34.7 seconds to 21.1 seconds while scoring 1 system (39.19% faster) and from 265.9 seconds to 91.9 seconds when scoring 8 systems (65.44% faster).

For all experiments performed along the rest of the chapter we always use both optimization on all COMET estimators being compared.

[1] A triplet encoder, is a model architecture where three sentences are encoded independently and in parallel. Architectures such as this have been extensively explored for sentence retrieval applications due to its efficiency (e.g. Sentence-BERT (Reimers and Gurevych, 2019))



(a) Runtime (in seconds) varying the size of the test-set in nº of sentences.



(b) Runtime (in seconds) varying number of systems for the `de-en` WMT20 Newstest.

Figure 5.2: Both experiments were performed with an `NVIDIA GeForce GTX 1080 TI` GPU, a constant batch size of 16. The time reported is the average of 5 runs using the COMET estimator architecture. For comparison we also plot the runtime of BLEU in a `Intel(R) Core(TM) i7-6850K CPU @ 3.60GHz`.

5.2 Model Pruning

Model pruning has been widely used in natural language processing to remove non-informative connections and thus reducing model size (Zhu and Gupta, 2018). Since most COMET parameters come from the XLM-R model, we attempt to reduce its size. We start with layer pruning by removing the top layers of XLM-R. Then we experiment with making its encoder blocks smaller either by reducing the size of the feed-forward hidden layers or by removing attention heads. The main advantage of these approaches is their simplicity: within minutes we are able to obtain a new model with reduced size and memory footprint with minimal performance impact.

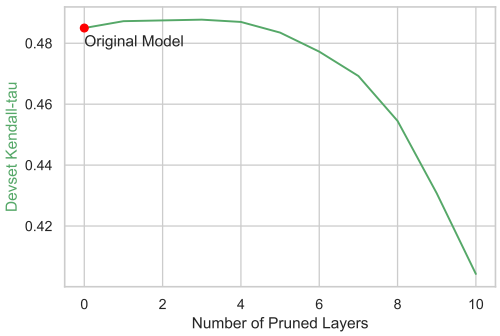
For all the experiments in this section, we used the development set from the Metrics shared task of WMT 2020. This set contains DA for English→German, English→Czech, English→Polish and English→Russian. We use these language pairs because they were annotated by experts exploring *document context* and in a *bilingual setup* (without access to a reference translation)[2]. Nonetheless, in Section 5.4 we show the resulting model performance on all language pairs from WMT 2021 for both DA and MQM.

5.2.1 Transformer Block Pruning

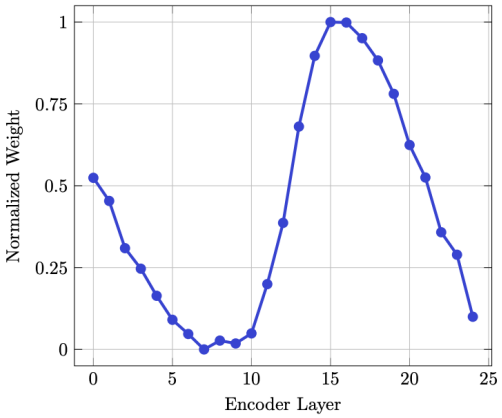
The Transformer architecture is composed of several encoder blocks (layers) stacked on top of the other. In the previous section, we reduce model size by removing the topmost blocks (depth pruning). In this section we reduce the size of each block instead (width pruning).

Each transformer block is made of two components: a *self-attention* (composed of several attention heads) and a *feed-forward neural network*. In XLM-R-large, each block is made of 16

[2]In the WMT 2020 findings paper Mathur et al. (2020b), most metrics showed suspiciously low correlations with human judgements based on crowd-sourcing platforms such as Mechanical Turk. Thus, we decided to focus just on 4 language pairs in which annotations are deemed as trustworthy.



(a) In this figure We can observe that removing up to 5 layers does not affect model performance but provides a 10% reduction in model size.



(b) In this figure we can observe layers between 15-19 are the most relevant ones with a normalized weight between 0.75 and 1. The representations learnt by layers 15-19 depend on previous layers but we can prune the top layers (20-25) without impacting the layers that the model deemed more relevant.

Figure 5.3: impacts of Layer Pruning in terms of performance using the WMT 2020 development set (Figure (a)) and the normalizes weights assigned to each layer when computing the final sentence level representation (Figure (b)).

self-attention heads followed by a feed-forward of a single hidden layer with 4092 parameters.

Using the `TextPruner` toolkit[3], we can easily prune both the attention heads and the feed-forward hidden sizes. Figure 5.4a shows the impact of pruning the hidden sizes from 4096→{512, 1024, 2048, 3072} while Figure 5.4b shows the impact of reducing the attention heads from 16→{4, 6, 8, 10, 12, 14}.

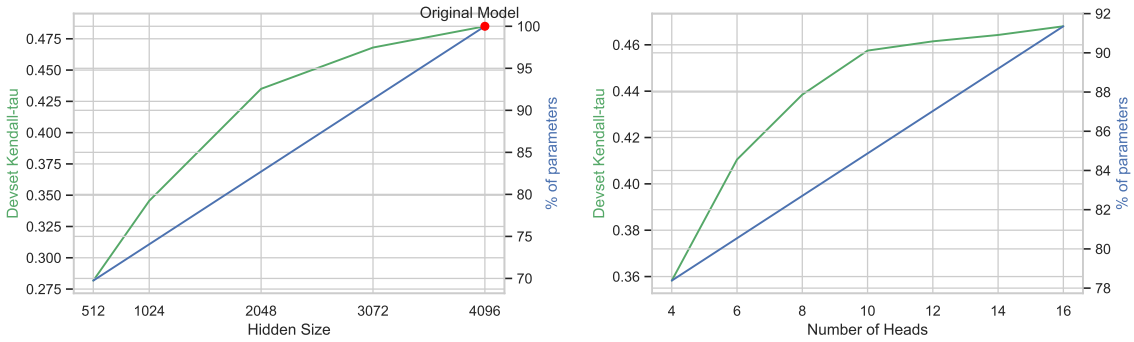
5.2.2 PRUNED-COMET

After experimenting with these three different pruning techniques, we created a pruned version of COMET in which we keep only 19 XLM-R layers, we reduced the feed-forward hidden size by $3/4$ (3072 hidden size) and we removed 2 heads (out of 16). According to our experiments above, the resulting model's performance drop should be almost the same as the original model but the resulting model is 21.1% smaller.

The resulting model is able to score 1000 samples in just 19.74 seconds, while the original model takes around 31.32 seconds. It is important to notice that most of the XLM-R parameters come from its huge embedding layer. Since the embedding size memory does not affect the inference time, the obtained 20% reduction in parameters translates into speed improvements of around 36.97%.[4]

[3] `https://textpruner.readthedocs.io/en/latest/`

[4] Experiments performed in a `NVIDIA GeForce GTX 1080 TI` GPU and a constant batch size of 16. The resulting time is the average of 5 runs.



(a) Impact of Feed-Forward hidden size pruning.

(b) Impact of attention-head pruning.

Figure 5.4: Impact of gradient based pruning techniques on model size (in blue) and performance on the WMT 2020 development set (in green). Note that in Figure (a) we apply pruning just for the feed-forward hidden size. In Figure (b) pruning is applied to several heads while freezing the hidden size to 3072 (3/4 of the original hidden size of XLM-R).

5.3 Distillation

Another commonly used way to compress neural networks is through knowledge distilation (Bucilua et al., 2006; Hinton et al., 2015) in which, for large amounts of unlabeled data, a smaller neural network (the student) is trained to mimic a more complex model (the teacher).

As the teacher network, we used an ensemble of 5 COMET models trained with different seeds. The student network follows the same architecture as the original model and the same hyper-parameters. However, instead of using XLM-R-large, it uses a distilled version with only 12 layers, 12 heads, embeddings of 384 features, and intermediate hidden sizes of 1536. This model has only 117M parameters compared to the 560M parameters from the large model.

Regarding the unlabeled data for distillation, we extracted 25M sentence pairs from OPUS ranging a total of 15 language pairs. To guarantee high quality parallel data we used Bicleaner tool (Ramírez-Sánchez et al., 2020) with a threshold of 0.8. Then, using pre-trained MT models available in Hugging Face Transformers, we created 2 different translations for each source: one using a bilingual model (in theory a high quality translation) and another using pivoting (which can be thought as lower quality). Finally, we scored all the data using our teacher ensemble. The resulting corpus contains 45M tuples with (`source`, `translation`, `reference`, `score`).

The resulting model which name DISTIL-COMET, scores 1000 sentences in 14.72 seconds resulting in a 53% speed improvement over the original model[3].

5.4 Correlation with Human Judgements

In this section, we show results for {PRUNE and DISTIL}-COMET in terms of correlations with MQM annotations from WMT 2021 Metrics task for two different domains: News and TED talks. Since these annotations only cover high-resource language pairs (English→German, English→Russian, Chinese→English), we also evaluate models on mid/low resource language pairs using DARR from WMT 2021, the same data we used in Chapter 3. For a detailed comparison, we also present

results for CHRF and BLEU, two computationally efficient lexical metrics, and other neural metrics such as PRISM[5], BLEURT and BERTSCORE (F1)

From Table 5.1, we can observe that PRUNE-COMET has minimal performance drops compared with COMET with only 80% of its parameters. DISTIL-COMET performance is on average 0.013 Kendall's bellow COMET for high resources languages, which is impressive for a model that only has 20% of COMET's parameters. For low-resource languages, we can observe bigger performance differences between COMET, PRUNE-COMET, and DISTIL-COMET which confirm results by Pu et al. (2021) that shows that smaller MT evaluation models are limited in their ability to generalize to several language pairs. Nonetheless, when comparing with other recently proposed metrics such as PRISM and BERTSCORE, {PRUNE and DISTIL}-COMET have higher correlations with human judgements for both high and low resource language pairs. The only exception is BLEURT which shows stronger correlations than COMET on high-resource language pairs and competitive performance in low-resource ones. We would like to emphasize that the model used in this section is not the most recent version, namely `wmt22-comet-da`, but rather an earlier iteration.

Table 5.1: Kendall's tau correlation on high resource language pairs using the MQM annotations for both News and TED talks domain collected for the WMT 2021 Metrics Task.

		zh-en		**en-de**		**en-ru**		
Metric	**Params**	**News**	**TED**	**News**	**TED**	**News**	**TED**	**avg.**
BLEU	-	0.166	0.056	0.082	0.093	0.115	0.067	0.097
CHRF	-	0.171	0.081	0.101	0.134	0.182	0.255	0.154
BERTSCORE	179M	0.230	0.131	0.154	0.184	0.185	0.275	0.193
PRISM	745M	0.265	0.139	0.182	0.264	0.219	0.292	0.229
BLEURT	579M	**0.345**	**0.166**	**0.253**	**0.332**	**0.296**	**0.347**	**0.290**
COMET	582M	0.336	0.159	0.227	0.290	0.284	0.329	0.271
PRUNE-COMET	460M	0.333	0.157	0.219	0.293	0.274	0.319	0.266
DISTIL-COMET	119M	0.321	0.161	0.202	0.274	0.263	0.326	0.258

Table 5.2: Kendall's tau-like correlations on low resource language pairs using the DARR data from WMT 2021 Metrics task.

Metric	**Params**	**zu-xh**	**xh-zu**	**bn-hi**	**hi-bn**	**en-ja**	**en-ha**	**en-is**	**avg.**
BLEU	-	0.381	0.1887	0.070	0.246	0.315	0.124	0.278	0.229
CHRF	-	0.530	0.301	0.071	0.327	0.371	0.186	0.373	0.308
BERTSCORE	179M	0.488	0.267	0.074	0.365	0.413	0.161	0.354	0.303
BLEURT	579M	**0.563**	**0.362**	**0.179**	0.498	0.483	0.186	0.469	0.391
COMET	582M	0.550	0.285	0.156	**0.526**	**0.521**	**0.234**	**0.474**	**0.392**
PRUNE-COMET	460M	0.541	0.264	0.163	0.519	0.513	0.197	0.439	0.377
DISTIL-COMET	119M	0.488	0.254	0.135	0.498	0.471	0.145	0.419	0.344

[5]PRISM does not support the low-resource language pairs used in our experiments, thus we only report PRISM correlations with MQM data

5.5 Conclusion

In this chapter we presented two simple optimizations that lead to significant performance gains on neural metrics such as COMET and two approaches to reduce its number of parameters. Together these techniques achieve impressive gains in performance (both speed and memory) at a very small cost in performance.

To showcase the effectiveness of our methods, we presented DISTIL-COMET and PRUNE-COMET. These models were obtained using an earlier iteration of `wmt22-comet-da` knowledge distillation and pruning respectively. To test the proposed models, we used the data from the WMT 2021 Metrics task which covers low resource languages as well as high resource languages. Overall the results of PRUNE-COMET are stable across the board with only a small degradation compared to the original metric. Knowledge distillation leads to much higher compression rates but seems to confirm previous findings (Pu et al., 2021) which suggest the lack of model capacity when it comes to the multilingual generalization for low resource languages.

Chapter 6

MT-TELESCOPE: An interactive platform for contrastive evaluation of MT systems

In this chapter we will present MT-TELESCOPE. The fundamental goal of MT-TELESCOPE is to widen access to state-of-the-art, robust MT comparison, to the benefit of the MT community at large. To do so MT-TELESCOPE explores features such as named entities and glossary handling which play a fundamental role in determining the suitability of an MT system for a production environment. Furthermore, the platform applies a bootstrapped t-test for statistical significance (Koehn, 2004) as a means of exposing the experimental rigor of system comparisons. These features are not widely available in other tools and provide a uniquely tailored solution to MT comparison that is highly informative and easy to use.

MT-TELESCOPE is open source, written in Python and uses a dynamic web interface implemented in streamlit[1]. In this manner, MT-TELESCOPE provides a uniquely accessible framework that requires little technical skill to operate and exposes information about the critical differences between MT outputs that is interactive, informative and highly customizable.

6.1 MT-TELESCOPE: Features

In this section, we describe the main features and visualizations implemented in MT-TELESCOPE and illustrate the user experience with examples:

6.1.1 User input and data

MT-TELESCOPE is opened in a web browser and takes four text (*.txt*) files as input; source and reference segments and one set of MT outputs for each of the compared systems. Users drag and drop these files directly onto the interface to begin evaluation. COMET is provided as a default metric given its proven value in the WMT Metrics Shared Task 2020 (Mathur et al., 2020b). Optionally the user can choose an alternate metric using a selection box. Currently available

[1] `https://streamlit.io/`

metrics include BLEU, METEOR and CHRF, and a selection of more recently proposed metrics such as PRISM, BLEURT, and BERTSCORE.

6.1.2 Visualizations

High-level results of the analysis are output in table format with the corresponding system scores. MT-TELESCOPE then exposes segment-level comparison in three primary visualizations:

First, a bubble plot (Figure 6.1) where the position of bubbles show how scores between the two systems differ for each segment, notable differences being highlighted with variations in bubble size and color. This method of visualization of MT is unique to MT-TELESCOPE in that it is fully interactive; by hovering the cursor over individual data points the user can preview the segments and output as well as relevant scores and the magnitude of the difference between them (as depicted in Figure 6.1). This plot allows for interactive exploration of the data which easily exposes differences in model behaviour at a glance. In particular, the distribution of points along the diagonal of this plot is highly informative; clustering along the diagonal indicates that the systems have minor differences whereas the contrary can indicate more dramatic change in behavior which can be hidden by the corpus-level mean.

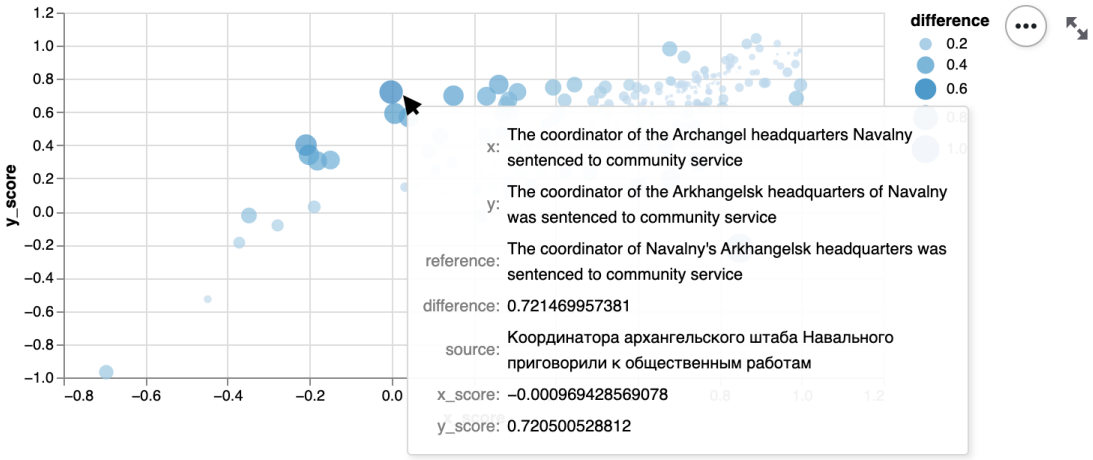


Figure 6.1: Segment comparison bubble plot.

Second, MT-TELESCOPE provides a bucketed error analysis in the form of a stacked bar plot (Figure 6.2). This plot serves to isolate specific bands of translation quality. These bands are highly customizable but can serve as a means of evaluating system utility; the plot can expose the extent to which either model outputs critical error for example. This is particularly useful in a commercial setting where the utility of a production system is inhibited by the presence of particular error types.

Segments are grouped into four buckets: *residual errors*, *minor errors*, *major errors*, and *critical errors*. The thresholds for each bucket can be dynamically adjusted by the user with appropriate sliders and (as with many of the features of MT-TELESCOPE) the plots are updated in real-time to reflect adjustments. Defaults were determined in line with suggestions outlined in the COMET GitHub documentation and with distributions of system-level scores from the WMT News Translation Shared Task 2020.

Residual Errors: The highest tier of quality by default reflects scores greater than 0.70, which generally equates to almost human-like translation with only minor, inconsequential error.

Minor Errors: By default this band reflects scores between 0.30 and 0.70 to reflect the division of quartiles from the distribution of system-level scores from the WMT News Translation Shared Task 2020. In general the band is associated with translation that is adequate but with minor flaws.

Major Errors: Translations scoring between 0.10 and 0.30 by default inhabit this band and are generally inadequate due to more serious error.

Critical Errors: Any translation scoring under 0.10 here is considered to contain critical error.

These bands are intended as a guide and utility of the default thresholds will vary according to use case. Translation quality and the difference between adequate and inadequate translation is highly subjective and language dependant; optimization of these thresholds is a critical direction for future work. Notwithstanding, we find that exposure of the general shift in distribution of inadequate translation in general is potentially informative, particularly given that corpus-level scores do not expose this type of analysis.

Finally, MT-TELESCOPE provides a histogram plot (Figure 6.3) for general evaluation of the distribution of scores between models. We propose that this kind of plot can potentially provide a high-level overview of the shift in performance between models. A corpus-level score (particularly an arithmetic mean) can mask variance between distributions of scores.

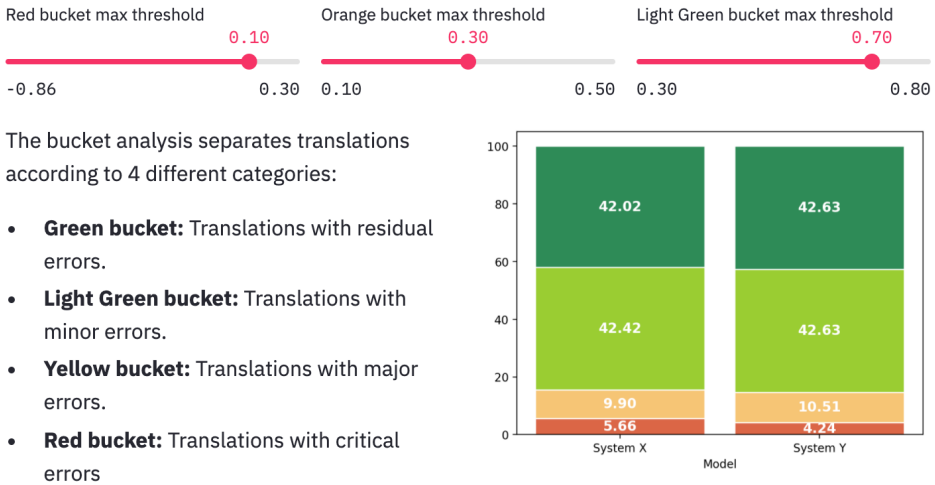


Figure 6.2: Segment-level error bucket analysis plot. In this plot, we can compare the two systems side by side according to the percentage of segments falling into 4 different category buckets: *residual errors*, *minor errors*, *major errors*, *critical errors*. The thresholds for defining these buckets can be dynamically adjusted using the sliders displayed above the plot.

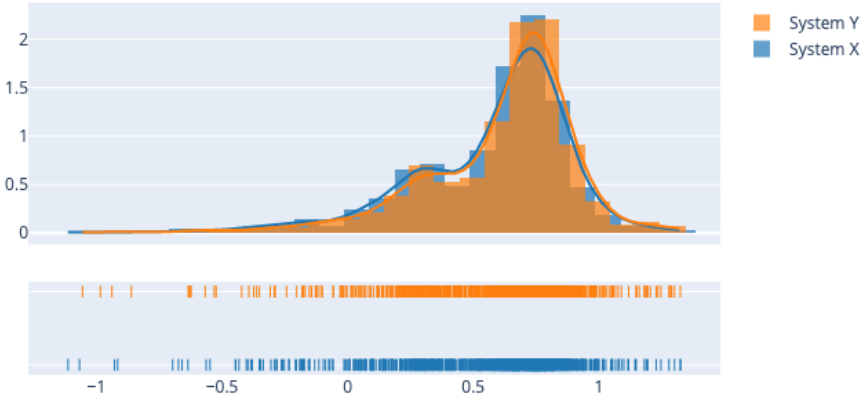


Figure 6.3: Segment-level histogram comparison.

6.1.3 Example evaluation

To demonstrate the utility of the MT-TELESCOPE evaluation we expose analyses for the *Online-G* and the *PROMT* (Molchanov, 2020) systems from the WMT News Translation Shared Task 2020 (Barrault et al., 2020) for Russian-English:

The *Online-G* system (System Y) achieves a COMET score of 0.6081, outperforming the *PROMT* system (System X) which only achieves 0.5972. We have isolated this example in particular as it represents a common occurrence of two systems achieving fairly comparable scores.

Figures 6.1, 6.2 and 6.3 above show the output of MT-TELESCOPE analysis on two sampled systems:

- Figures 6.2 and 6.3 illustrate that the second system (System Y) in general exceeds performance of the first (System X). We can conclude from these plots that the systems perform comparably with System Y producing a higher percentage of adequate translations. In particular we note that System Y outputs fewer critical errors, consistent with its general performance gain.
- Figure 6.1 illustrates isolation of an example where System Y makes substantial gain over System X. Here we note that both systems struggle to render the named entity and the corresponding possessive, but that System Y successfully produces the named entity as reflected in the reference and adds a pronoun to at least give possessive flavor.

6.2 MT-TELESCOPE: Dynamic Corpus Filtering

Given a test corpus, MT-TELESCOPE provides functionality to dynamically evaluate sub-samples of the system outputs as a means of focused analysis tailored to particular phenomena relevant to

MT. On selection of any of the available filtering criteria, the MT-TELESCOPE Dynamic Corpus Filtering feature (DCF) updates the output evaluation in real-time to allow the user to "zoom in" on relevant data points.

Currently, MT-TELESCOPE supports filtering by named entity, glossary and source segment length, as well as an option to remove duplicates. Whenever any of these options is selected, the interface will output the size of the sub-sample as a percentage of the original test corpus.

6.2.1 DCF: Named Entities

Successful rendering of named entities is a known challenge for even modern MT systems and can lead to distortion of locations, organization and other names (Koehn and Knowles, 2017; Modrzejewski et al., 2020). Recently, several methods have been proposed to improve the translation of named entities in Neural Machine Translation (NMT) (Sennrich and Haddow, 2016; Ugawa et al., 2018; Modrzejewski et al., 2020), but precise measurement of translation quality improvements for these techniques is inhibited by the fact that not all sentences in traditional benchmark test sets (e.g. WMT test sets) contain named entities and that scores produced by automated evaluation metrics are not sufficiently fine-grained to reflect this type of variation. MT-TELESCOPE offers a potential solution to this by applying the following filter:

We initially run the Stanza Named Entity Recognition (NER) model (Stanza, Qi et al. 2020)[2] over the source test corpus to isolate segments that contain named entities. If the source language (as specified by the user) is not supported by Stanza, we run NER on the reference. MT-TELESCOPE will then update the output analysis allowing focused evaluation of the handling of segments containing named entities by either MT system.

Table 6.1: Example of named entity errors produced *Online-G* system in comparison to the *PROMT* system from the WMT20 shared task.

		COMET
Source	Маругов врезался на мотоцикле в такси, которым управлял Акбаров.	
Online-G	Murugov crashed into a motorcycle taxi, which was ruled by Akbar.	-0.1799
PROMT	Marugov crashed into a taxi driven by Akbarov on a motorcycle.	0.5154
Reference	Marugov crashed on a motorcyle into the taxi Akbarov was driving.	

To illustrate the utility of DCF analysis on named entities we again compare the outputs of the *Online-G* and the *PROMT* (Molchanov, 2020) systems from the Metrics Shared Task 2020 (Barrault et al., 2020) as above:

Applying DCF for named entities, the *Online-G* system COMET score drops to 0.5851 (previously 0.6081), while the *PROMT* system only drops to 0.5888 (previously 0.5972). We also observe that the percentage of critical segments from the *Online-G* system in our bucketed analysis jumps from 6.26% to 7.0%, while the corresponding percentage output by the *PROMT* system drops from 6.66% to 6.29%.

On the basis of the DCF analysis for named entities we can conclude that whilst in general the *Online-G* exhibits superior quality, it may be under-performing with regard to named entities. Interestingly, the system description paper for the *PROMT* system (Molchanov, 2020) specifically details a targeted approach to handling translation of named entities, which may explain its stronger performance on the isolated sub-sample.

[2] `https://stanfordnlp.github.io/stanza/ner.html`

In Table 6.1 we illustrate an example of a translation in which the *Online-G* system produces critical errors as a consequence of translating named entities incorrectly, specifically isolated by the DCF feature.

6.2.2 DCF: Terminology

Similarly to named entities, enforcing that MT systems use specific terminology during translation is a challenging task with particular relevance in commercial use cases. Measuring terminology adherence typically involves relying on automated metrics for MT as well as measuring the accuracy of terminology output (Dinu et al., 2019; Exel et al., 2020).

This approach presents two concrete problems: a) applying terminology constraints typically results in only minimal variance between translations, which limits the utility of using automated metrics at the corpus level; and b) measuring accuracy in terminology usage typically relies on exact string matching between a translation hypothesis and its respective reference, which implies that properly inflected translated terms often do not receive proper credit.

MT-TELESCOPE offers a DCF Terminology feature which allows a user to optionally upload a glossary by which to isolate a corresponding sub-sample of the test corpus. We apply string matching on the source and filter to only those segments which contain a corresponding glossary match.

6.2.3 DCF: Segment Length

Another common weakness of some MT systems is their inability to accurately translate long segments (Koehn and Knowles, 2017). In general, corpus level evaluation on a distribution that includes very short segments can artificially inflate performance, with substantial drops in scores being observed when these segments are specifically excluded (Koehn and Knowles, 2017). In the same manner, quality-based decisions regarding two systems can change when we consider segments of different lengths.

Using our example systems outlined above in Section 6.2.1, when comparing the *Online-G* and the *PROMT* systems using only the top 50% longest segments, the *PROMT* system outperforms the *Online-G* system according to COMET and CHRF scores, changing the fundamental perception of which system is 'better'. With the above in mind, MT-TELESCOPE also offers an option to filter by segment length. This filter is adaptive to the distribution of segment lengths in the test corpus. We first build the distribution of the source segment lengths (measured in terms of characters) for the entire test set. Then, the user can select which part of the distribution to analyse by adjusting the a and b parameters of the density function $P(a \leq X \leq b)$; a and b being the minimum and maximum length allowed, respectively.

6.2.4 DCF: Duplication

The removal of duplicates can be particularly important in situations where the test corpus sample contains repetition. Repeated segments in a test sample can artificially inflate the corpus-level score, particularly where that score results from an average of segment-level scores. Whilst we acknowledge that removal of duplicate segments is fairly common in public data sets such as

that used in the WMT Shared Tasks and consequently our example here, we propose that it is, nevertheless, a useful tool when evaluating on random samples.

6.3 Statistical Significance Testing

By default, MT-TELESCOPE implements the bootstrapped t-test for statistical significance promoted for use in comparison of MT systems by Koehn (2004). Specifically, we iteratively resample a portion of the test set (of size P) N times, compare corpus-level results of each subsample and record the comparative conclusions. The ratio of wins of a single system is a reasonable proxy to the probability that that system is better than the other. In other words, if one system outperforms the other system 95% of the time, we conclude that the former is better with a significance of $p = 0.05$ (Koehn, 2004).

This is particularly useful in cases where the relative difference between systems is minimal and acts as a measure of the robustness of any resulting decision. In our implementation P is an optional parameter which defaults to 0.5 (50%) or 500 segments, whichever is larger, to ensure reasonable stability in the output conclusion[3]. N is also user defined and by default is set at 300 iterations.

6.4 Conclusion

In this chapter we show how MT-TELESCOPE is designed to provide robust and insightful comparative analysis specific to the MT use case with state-of-the-art metrics. Data visualizations are dynamic, interactive and highly customizable. The tools have been built specifically with ease of use in mind, in the hope of expanding access to high quality MT evaluation.

[3]Employing a lower P value entails balancing statistical power and efficiency. A P value of 0.5 strikes a favorable balance between the two.

Chapter 7

Additional Contributions

7.1 Uncertainty-Aware MT Evaluation

While the metrics above — COMET, BLEURT , UNITE — have enjoyed success in correlations with human judgements, their segment-level quality scores are often unreliable. They all share the limitation that their output is a single *point estimate* – they do not provide any uncertainty information, such as confidence intervals, with their quality predictions. This is an important limitation: often, complex or out-of-domain sentences receive quality estimates that are far from their true quality (as illustrated in Figure 7.1). This may lead to translations with critical mistakes being undetected, and hinders worst-case performance analysis of MT systems.

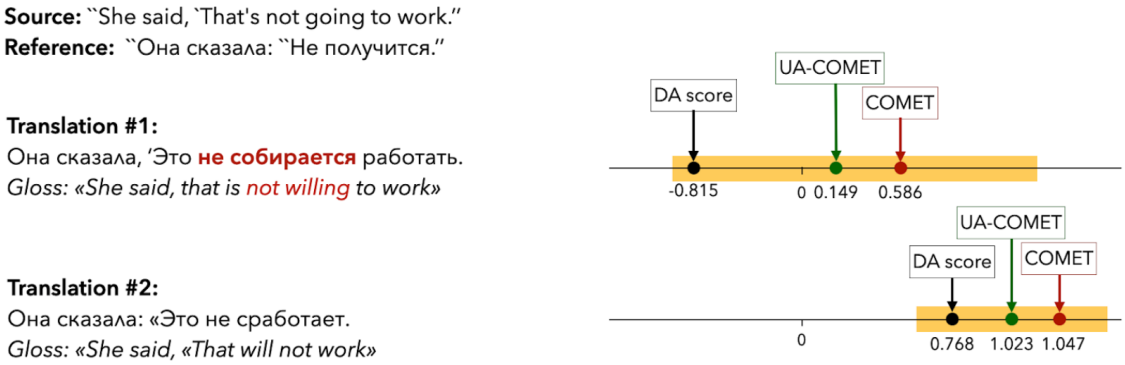


Figure 7.1: Example of uncertainty-aware MT evaluation for a sentence in the WMT20 dataset. Shown are two Russian translations of the same English source "She said, 'That's not going to work." with reference "Она сказала: "Не получится ". For the first sentence, COMET provides a point estimate in red that overestimates quality, as compared to a human direct assessment (DA), while our UA-COMET returns a large 95% confidence interval which contains the DA value. For the second sentence UA-COMET is confident and returns a narrow 95% confidence interval.

In this work, we propose a simple and effective method to obtain **uncertainty-aware** neural fine-tuned metrics, by representing quality as a *distribution*, rather than a single value. To this end, we make use of two well-studied techniques for uncertainty estimation, namely: Monte Carlo (MC) dropout (Gal and Ghahramani, 2016) and deep ensembles (Lakshminarayanan et al., 2017).

		Pearson (τ)	UPS ↑	NLL ↓	ECE ↓	Sha. ↓
EN-DE	MC Dropout	0.452	<u>0.409</u>	<u>1.433</u>	<u>0.024</u>	0.674
	Deep Ensemble	<u>0.459</u>	0.336	1.435	0.035	<u>0.556</u>
	Baseline	0.452	-	1.437	0.094	1.031
ZH-EN	MC Dropout	<u>0.503</u>	<u>0.309</u>	<u>1.402</u>	<u>0.018</u>	0.721
	Deep Ensemble	0.485	0.257	1.415	0.023	<u>0.653</u>
	Baseline	0.503	-	1.398	0.059	0.953

Table 7.1: Results for segment-level MQM prediction. <u>Underlined</u> numbers indicate the best result for each language pair and evaluation metric. Reported are the predictive Pearson score $r(\hat{\mu}, q^*)$ (τ), the uncertainty Pearson score $r(|q^* - \hat{\mu}|, \hat{\sigma})$ (UPS), the negative log-likelihood (NLL), the expected calibration error (ECE), and the sharpness (Sha.) (see Appendix A Section X). Note that the UPS of the baseline is always zero, since it has a fixed variance.

In both cases, our methods are agnostic to the underlying metric, as long as it can be ensembled or perturbed. In our experiments we use the COMET metrics and, to get confidence intervals, we either apply a dropout probability of 0.1 and run $N = 100$ runs of MC dropout, or we get predictions from 5 different models that are replicas of each other trained with 5 different seeds.

The models employed in these experiments are estimator models that were originally submitted to the WMT20 Metrics shared task Rei et al. (2020b). The DA estimator represents an earlier version of our primary metric presented in Section 3.5, trained exclusively on DA data from the years 2017 to 2019. On the other hand, the HTER model resembles the HTER model outlined in Section 3.4, albeit employing an XLM-R Large encoder. For the purpose of deep ensembles, we retrained these models using five different seeds. The ensemble employed as a *teacher* in Chapter 5 is identical to the one developed for this study.

We evaluate our approach using data from the WMT20 metrics task (Mathur et al., 2020b), including its recent extension with Google MQM annotations (Freitag et al., 2021a), and the QT21 dataset (Specia et al., 2017). In Table 7.1 we show results for the MQM corpus were we can observe that the resulting confidence intervals are informative and correlated with the prediction errors, leading to slightly more accurate predictions with informative uncertainty. On top of the benefits of having an uncertainty-aware metric, we were able to improve the performance of the original model (measured by the Pearson correlation between the model's predictions and ground truth). In fact, by using deep ensembles and MC dropout we improve prediction performance over the original model (baseline) across all language pairs. In Appendix A (Section B) we explain with more detail how evaluate uncertainty and we also show the results for the other 2 test sets used.

We next experiment with the WMT20 EN-DE MQM data to get some insights on the impact of using multiple references. This dataset contains 3 human references (Human A, B, and P) for each source sentence generated in different ways: A and B are generated independently by annotators and P is a paraphrased as-much-as-possible version of A. Our goal is to simulate the availability of multiple human references of varying quality levels. As reported in the findings of WMT20 Metrics task (Mathur et al., 2020b), in realistic scenarios the available references have very disparate quality levels, and the quality of human references is not always known. We thus calculate the performance when using each of the Human-A, Human-B and Human-P references individually, and then compare randomly sampling r from $\mathcal{R}$ with averaging predictions over each r in $\mathcal{R}$, hypothesizing that the combination of references will result in reduced model uncertainty.

	#r	Pearson (τ)	UPS ↑	NLL ↓	ECE ↓	Sha. ↓
			$\mathcal{R}$={A,B}			
S-1	1	0.452	0.407	1.403	<u>0.017</u>	0.746
Mul	2	<u>0.471</u>	<u>0.389</u>	<u>1.388</u>	0.020	<u>0.718</u>
			$\mathcal{R}$={B,P}			
S-1	1	0.391	0.327	1.470	0.029	0.837
Mul	2	<u>0.441</u>	<u>0.331</u>	<u>1.429</u>	<u>0.013</u>	<u>0.753</u>
			$\mathcal{R}$={A,P}			
S-1	1	0.406	0.334	1.475	0.026	0.852
Mul	2	<u>0.433</u>	<u>0.339</u>	<u>1.460</u>	<u>0.019</u>	<u>0.719</u>
			$\mathcal{R}$={A,B,P}			
S-1	1	0.402	<u>0.355</u>	1.473	0.026	0.825
S-2	2	0.441	0.348	1.424	0.019	0.756
Mul	3	<u>0.455</u>	<u>0.351</u>	1.417	0.018	<u>0.702</u>

Table 7.2: Performance over multiple references and combination patterns on EN-DE Google MQM annotations. S-N signifies sampling w/o replacement N references from $\mathcal{R}$; Mul signifies combining estimates over multiple references in $\mathcal{R}$. <u>Underlined</u> numbers indicate the best result for each evaluation metric and reference set.

	Pearson (τ)	UPS ↑	NLL ↓	ECE ↓	Sha. ↓
$\mathcal{R}$={A}	<u>0.452</u>	<u>0.409</u>	<u>1.433</u>	0.024	<u>0.674</u>
$\mathcal{R}$={B}	0.442	0.400	1.406	<u>0.015</u>	0.782
$\mathcal{R}$={P}	0.391	0.275	1.511	0.020	0.783

Table 7.3: Performance over singleton reference sets on EN-DE Google MQM annotations. <u>Underlined</u> numbers indicate the best result for each evaluation metric.

We can see in Table 7.2 that when having access to multiple references, combining all available references (Mul) results in narrower confidence intervals compared to sampling single references (S-1) or even pairs of references (S-2) as indicated by the decreasing values in sharpness. Apart from sharpness, the model seems to benefit from the addition of new knowledge, since we see consistent improvement in performance for τ and NLL metrics. Thus, with the incorporation of additional human references we obtain models that are more confident – and rightly so, since they are more predictive too. Combining this information with the performance of singleton reference sets in Table 7.3, we note that even among human references, the estimated reference quality seems to have an impact both on the predictive accuracy (τ) and confidence (UPS, NLL, Sharpness). Both for S-N and Mul approaches, the inclusion of Human-P in the reference set results in performance drop across all metrics. Still, the negative impact of Human-P decreases with the increase of combined references and we can conclude that when there is no information on the estimated quality of references the best approach is to combine them: for $\mathcal{R} = \{A, B, P\}$, Mul results in similar performance to Human-A.

7.2 Quality-Aware Decoding

Despite the progress in machine translation evaluation (both QE and Metrics) in the last years, decoding in neural machine translation (NMT) is mostly oblivious to this and centers around finding the most probable translation according to the model (MAP decoding), approximated with beam search. In this work we bring together these two lines of research and propose *quality-aware decoding* for NMT, by leveraging recent breakthroughs in QE and MT Metrics through various inference methods like N-best reranking and minimum Bayes risk decoding (MBR) where the generation process is decoupled into two steps: *candidate generation* and *candidate selection*. Figure 7.2 illustrates the described decoding framework.

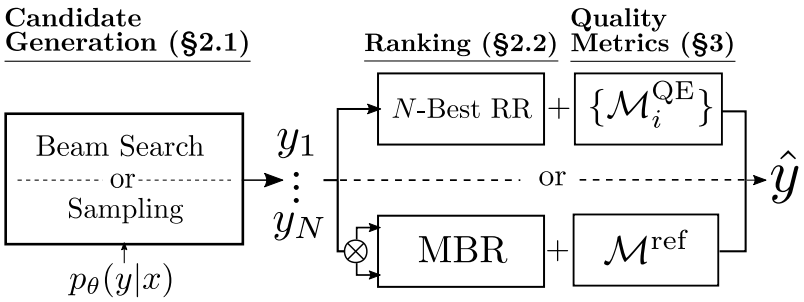


Figure 7.2: Quality-aware decoding framework. First, translation candidates are *generated* according to the model. Then, using reference-free and/or reference-based MT metrics, these candidates are *ranked*, and the highest ranked one is picked as the final translation.

The main question we try to answer in this work is if we can leverage recent advances in MT evaluation to generate better translations and, if so, how can we most effectively do so? To do so we explored the impact of combining 4 different ranking strategies:

1. **Fixed N-best Reranker:** An N-best reranker using a single reference-free metric as a feature.

2. **Tuned N-best Reranker:** An N-best reranker using as features several reference-free metrics, along with the model log-likelihood $\log p_\theta(y|x)$. The weights of each feature/reference-free metric are optimized to maximize a given reference-based metric (e.g COMET) using MERT (Och, 2003), a coordinate-ascent optimization algorithm widely used in previous work.

3. **MBR Decoding:** Choosing as the utility function a reference-based metric, we estimate the utility using a simple Monte Carlo sum where each hypothesis is compared against each other.

4. **N-best Reranker $\rightarrow$ MBR:** Using a large number of samples in MBR decoding is expensive due to its quadratic cost. To circumvent this issue, we explore a *two-stage* ranking approach: we first rank all the candidates using a tuned N-best reranker, followed by MBR decoding using the top M candidates. The first ranking stage *prunes* the candidate list to a smaller, higher quality subset, making possible a more accurate estimation of the utility with less samples, and potentially allowing a better ranker than *plain* MBR for almost the same computational budget.

In this work, for both **ranking and performance evaluation** of MT systems we used BLEU, CHRF, BLEURT and COMET. For N-best reranking we explore four recently proposed reference-free metrics as features at the sentence-level:

- COMET-QE-DA, this was a reference-free model, following a similar architecture as COMET, that was the winning submission for the "QE-as-a-metric" subtask of the WMT20 shared task (Mathur et al., 2020b).
- TransQuest (Ranasinghe et al., 2020), the winning submission for the sentence-level DA prediction subtask of the WMT20 QE shared task (Specia et al., 2020). Similarly to COMET-QE this metric predicts a DA score. Regaring its architecture if is similar to COMETKIWI (Section 3.8).
- MBART-QE (Zerva et al., 2021), based on the mBART Liu et al. (2020) model, trained to predict both the *mean* and the *variance* of DA scores. It was a top performer in the WMT21 QE shared task (Specia et al., 2021).
- OpenKiwi-MQM which is trained to predict MQM and it was ranked second on the "QE-as-a-metric" subtask from the WMT 2021 metrics shared task. This models not only produces a sentence-level score but it also predicts word-level `Ok/Bad` tags.

Regarding our **experimental setup** we study the benefits of quality-aware decoding over MAP-based decoding in two regimes:

- A high-resource, unconstrained, setting with *large* transformer models (6 layers, 16 attention heads, 1024 embedding dimensions, and 8192 hidden dimensions) trained by Ng et al. (2019) for the WMT19 news translation task (Barrault et al., 2019), using English to German (EN $\rightarrow$ DE) and English to Russian (EN $\rightarrow$ RU) language pairs. These models were trained on over 20 million parallel and 100 million back-translated sentences, being the winning submissions of that year's shared task. We consider the non-ensembled version of the model and use *newstest19* for validation and *newstest20* for testing.
- A more constrained scenario with a *small* transformer model (6 layers, 4 attention heads, 512 embedding dimensions, and 1024 hidden dimensions) trained from scratch in *Fairseq* (Ott et al., 2019) on the smaller IWSLT17 datasets (Cettolo et al., 2012) for English to German (EN $\rightarrow$ DE) and English to French (EN $\rightarrow$ FR), each with a little over 200k training examples. We chose these datasets because they have been extensively used in previous work (Bhattacharyya et al., 2021) and smaller model allows us to answer questions about how the training methodology affects ranking performance.

We use beam search with a beam size of 5 as our decoding baseline because we found that it resulted in better or similar translations than larger beam sizes. For tuned *N*-best reranking, we use Travatar's (Neubig, 2013) implementation of MERT (Och, 2003) to optimize the weight of each feature. Finally, we evaluate each system using BLEU, CHRF, BLEURT and COMET.

Our results according to automatic evaluation can be found in Table 7.4. For *fixed* N-best reranker with a single reference-free metric (1^{st} group in Table 7.4), while none of the metrics allows for improving the baseline results in terms of the lexical metrics (BLEU, CHRF), rerankers using COMET-QE-DA or MBART-QE outperform the baseline according to BLEURT and COMET.

	Large (WMT20)				Small (IWSLT)			
	BLEU	chrF	BLEURT	COMET	BLEU	chrF	BLEURT	COMET
Baseline	**36.01**	63.88	0.7376	0.5795	29.12	56.23	0.6635	0.3028
F-RR w/ COMET-QE-DA	29.83	59.91	0.7457	0.6012	27.38	54.89	0.6848	0.4071
F-RR w/ MBART-QE	32.92	62.71	0.7384	0.5831	27.30	55.62	0.6765	0.3533
F-RR w/ OpenKiwi	30.38	59.56	0.7401	0.5623	25.35	51.53	0.6524	0.2200
F-RR w/ Transquest	31.28	60.94	0.7368	0.5739	26.90	54.46	0.6613	0.2999
T-RR w/ BLEU	35.34	63.82	0.7407	0.5891	**30.51**	**57.73**	0.7077	0.4536
T-RR w/ BLEURT	33.39	62.56	0.7552	0.6217	30.16	57.40	0.7127	0.4741
T-RR w/ COMET	34.26	63.31	0.7546	0.6276	30.16	57.32	0.7124	0.4721
MBR w/ BLEU	34.94	63.21	0.7333	0.5680	29.25	56.36	0.6619	0.3017
MBR w/ BLEURT	32.90	62.34	0.7649	0.6047	28.69	56.28	0.7051	0.3799
MBR w/ COMET	33.04	62.65	0.7477	0.6359	29.43	56.74	0.6882	0.4480
T-RR+MBR w/ BLEU	35.84	**63.96**	0.7395	0.5888	30.23	57.34	0.6913	0.3969
T-RR+MBR w/ BLEURT	33.61	62.95	**0.7658**	0.6165	29.28	56.77	**0.7225**	0.4361
T-RR+MBR w/ COMET	34.20	63.35	0.7526	**0.6418**	29.46	57.13	0.7058	**0.5005**

Table 7.4: Evaluation metrics for EN $\rightarrow$ DE for the *large* and *small* model settings, using a *fixed* N-best reranker (F-RR), a *tuned* N-best reranker (T-RR), MBR decoding, and a two-stage approach. Best overall values are **bolded** and best for each specific group are underlined.

If we consider a *tuned* N-best reranker (2$^{\text{nd}}$ group in Table 7.4) using as features *all* the reference-free metrics, and optimized using MERT for a particular metric all the rankers show improved results over the baseline. In particular, optimizing for BLEU, leads to the best results in the lexical metrics, while optimizing for BLEURT leads to the best performance in the others.

Table 7.4 (3$^{\text{rd}}$ group) shows the impact of using MBR with different utility function (BLEU, BLEURT and COMET). For the *small* model, using COMET leads to the best performance according to all the metrics except BLEURT (for which the best result is attained when optimizing itself). For the *large* model, the best result according to a given metric is obtained when using that metric as the utility function.

Finally, looking at Table 7.4 4$^{\text{th}}$ group we see that, for both the *large* and the *small* model, the two-stage ranking approach (N-best Reranker $\rightarrow$ MBR) leads to the best performance according to the fine-tuned metrics. In particular, the best result is obtained when the utility function is the same as the evaluation metric. These results suggest that a promising research direction is to seek more sophisticated pruning strategies for MBR decoding.

To further investigate how *quality-aware* decoding performs when compared to *MAP-based* decoding, we perform another human study, this time based on MQM. According to Table 7.5, despite the remarkable performance of the F-RR with COMET-QE-DA in terms of COMET, the quality of the translations decreases when compared to the baseline, suggesting the possibility of *metric overfitting* when evaluating systems using a single automatic metric that was directly optimized for (or a similar one). However, for both language pairs, the T-RR with COMET and the two stage approach (T-RR$\rightarrow$MBR with COMET) achieve the highest MQM score. In addition, these systems present the smallest number of errors when combining both major and critical errors.

	EN-DE (WMT20)				EN-RU (WMT20)			
	Minor	Major	Critical	MQM	Minor	Major	Critical	MQM
Reference	24	67	0	97.04	5	11	0	99.30
Baseline	8	139	0	95.66	17	239	49	79.78
F-RR w/ COMET-QE	15	204	0	93.47	13	254	80	76.25
T-RR w/ COMET	12	109	0	**96.20**	9	141	45	85.97†
MBR w/ COMET	11	161	0	94.38	8	182	40	83.65
T-RR + MBR w/ COMET	10	138	0	95.44	11	134	45	**86.78**†

Table 7.5: Error severity counts and MQM scores for WMT20 (large models). Best overall values are **bolded**. Methods with † are statistically significantly better than the baseline, with $p < 0.05$.

Chapter 8

Conclusion and Future Work

In this thesis, we have addressed the challenges and limitations of automatic MT evaluation and proposed novel ways to improve the state of the field. We have focused on the following desiderata: **strong correlation with human judgments, robustness to diverse languages and domains, interpretability, and efficiency**. Our contributions can be summarized as follows:

1. We introduced COMET, a PyTorch-based framework for training highly multilingual and adaptable MT evaluation models that can function as metrics. COMET metrics incorporate source-language information and leverage recent advancements in cross-lingual language modeling. These metrics have demonstrated high correlations with human judgments and have been widely adopted by the MT community.

2. We conducted a study titled "The Inside Story," where we investigated the inner workings of COMET metrics. Through the analysis of token-level neural saliency maps and comparisons with human annotations, we revealed that these metrics effectively capture translation errors and provide valuable insights into their decision-making process.

3. To address the computational cost of COMET metrics, we introduced techniques based on pruning and knowledge distillation. These techniques led to the creation of more compact and faster versions of COMET metrics, referred to as COMETINHO's. These improved metrics maintain high correlations with human judgments while enhancing efficiency in scenarios where speed is crucial.

4. We developed MT-TELESCOPE, an analysis tool designed for comparing two MT systems side-by-side under different circumstances. MT-TELESCOPE enables robust MT comparison by incorporating state-of-the-art evaluation metrics, statistical tests, dynamic filters, and a visual interface. This tool facilitates the adoption of best practices in MT evaluation and empowers researchers and industry practitioners.

The contributions presented in this thesis have advanced the field of MT evaluation by offering new evaluation methodologies, interpretability insights, improved efficiency and best practices. We have demonstrated the importance of strong correlation with human judgments and the value of incorporating source-language information into MT evaluation. Our work has shed light on the inner workings of neural fine-tuned metrics and addressed the computational cost of these metrics,

making them more practical in various settings. Additionally, we provided a comprehensive tool for MT system comparison.

Overall, our research has paved the way for more robust, interpretable, and efficient MT evaluation. The adoption of our contributions will lead to better-informed decisions in MT model selection, system development, and deployment.

Looking ahead, there are several directions for future work that can further advance MT evaluation. Firstly, there is a need to develop metrics that align better with the MQM framework. These metrics should go beyond providing a sentence-level score and be able to identify error spans with corresponding categories such as minor, major, and critical errors. Although attempts have been made (Rei et al., 2021a, 2022b; Perrella et al., 2022), the results have been limited due to the scarcity of publicly available data to train such models. Addressing this limitation by creating larger and more diverse datasets would be crucial for progress in this area.

Secondly, orthogonal to the first direction, it is important to explore the potential of scaling. Our research has shown that scaling a reference-free model to 10.7 billion parameters leads to improved results. Metrics like COMET, BLEURT, and UNITE, which currently have less than 600 million parameters, represent only the tip of the iceberg. With larger models, these metrics can be expected to continue improving. Moreover, the use of Large Language Models (LLMs) presents an opportunity to frame the evaluation task as a generative task. This approach has the potential to create more interpretable metrics that not only provide a score but also offer explanations in textual form, further enhancing our understanding of the evaluation results.

Finally, in recent months, there has been a lot of interest in Reinforcement Learning (RL) from Human Feedback. In these works, the training is split into 3 phases: 1) Pretraining, 2) Creation of a reward model from human annotations, and 3) fine-tuning with RL. COMET and all these new neural fine-tuned metrics can be seen as strong reward models for MT training using RL. Our quality-aware decoding work can be seen as a first step in that direction, but there is still much to investigate in this area, especially with QE models. This exploration could open the door to unsupervised fine-tuning of MT models.

Bibliography

Duarte Alves, Ricardo Rei, Ana C Farinha, José G. C. de Souza, and André F. T. Martins. 2022. Robust MT evaluation with sentence-level multilingual augmentation. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 469–478, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Chantal Amrhein, Nikita Moghe, and Liane Guillou. 2022. ACES: Translation accuracy challenge sets for evaluating machine translation metrics. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 479–513, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

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

Leila Arras, Ahmed Osman, Klaus-Robert Müller, and Wojciech Samek. 2019. Evaluating recurrent neural network explanations. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 113–126, Florence, Italy. Association for Computational Linguistics.

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

Arsenii Ashukha, Alexander Lyzhov, Dmitry Molchanov, and Dmitry Vetrov. 2020. Pitfalls of in-domain uncertainty estimation and ensembling in deep learning. In *International Conference on Learning Representations*.

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.

Jasmijn Bastings, Wilker Aziz, and Ivan Titov. 2019. Interpretable neural predictions with differentiable binary variables. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2963–2977, Florence, Italy. Association for Computational Linguistics.

Sumanta Bhattacharyya, Amirmohammad Rooshenas, Subhajit Naskar, Simeng Sun, Mohit Iyyer, and Andrew McCallum. 2021. Energy-based reranking: Improving neural machine translation using energy-based models. 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 4528–4537, Online. Association for Computational Linguistics.

Ondřej Bojar, Yvette Graham, and Amir Kamran. 2017. 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. 2016. 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.

Cristian Bucilua, Rich Caruana, and Alexandru Niculescu-Mizil. 2006. Model compression. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '06, page 535–541, New York, NY, USA. Association for Computing Machinery.

Mauro Cettolo, Christian Girardi, and Marcello Federico. 2012. WIT3: Web inventory of transcribed and translated talks. In *Proceedings of the 16th Annual conference of the European Association for Machine Translation*, pages 261–268, Trento, Italy. European Association for Machine Translation.

Tyler Chang, Zhuowen Tu, and Benjamin Bergen. 2022. The geometry of multilingual language model representations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 119–136, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ethan A. Chi, John Hewitt, and Christopher D. Manning. 2020. Finding universal grammatical relations in multilingual BERT. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5564–5577, Online. Association for Computational Linguistics.

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

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

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 7059–7069. Curran Associates, Inc.

Michael Denkowski and Alon Lavie. 2010. METEOR-NEXT and the METEOR paraphrase tables: Improved evaluation support for five target languages. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 339–342, Uppsala, Sweden. Association for Computational Linguistics.

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.

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.

Georgiana Dinu, Prashant Mathur, Marcello Federico, and Yaser Al-Onaizan. 2019. Training neural machine translation to apply terminology constraints. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3063–3068, Florence, Italy. Association for Computational Linguistics.

George Doddington. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the Second International Conference on Human Language Technology Research*, HLT '02, page 138–145, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Miriam Exel, Bianka Buschbeck, Lauritz Brandt, and Simona Doneva. 2020. Terminology-constrained neural machine translation at SAP. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 271–280, Lisboa, Portugal. European Association for Machine Translation.

Christian Federmann. 2010. Appraise: An open-source toolkit for manual phrase-based evaluation of translations. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Christian Federmann. 2018. Appraise evaluation framework for machine translation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 86–88, Santa Fe, New Mexico. Association for Computational Linguistics.

Christiane Fellbaum, editor. 1998. *WordNet: an electronic lexical database*. MIT Press.

Patrick Fernandes, Antonio Farinhas, Ricardo Rei, José G. C. de Souza, Perez Ogayo, Neubig Graham, and André F. T. Martins. 2022. Quality-Aware Decoding for Neural Machine Translation. In *Proceedings at the 2022 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, Seattle, Washington. Association for Computational Linguistics.

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

Marina Fomicheva, 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.

Erick Fonseca, Lisa Yankovskaya, André F. T. Martins, Mark Fishel, and Christian Federmann. 2019. Findings of the WMT 2019 shared tasks on quality estimation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 1–10, Florence, Italy. Association for Computational Linguistics.

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

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André F. T. Martins. 2022. 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*, Online. Association for Computational Linguistics.

Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1050–1059, New York, New York, USA. PMLR.

Yang Gao, Steffen Eger, Wei Zhao, Piyawat Lertvittayakumjorn, and Marina Fomicheva, editors. 2021. *Proceedings of the 2nd Workshop on Evaluation and Comparison of NLP Systems*. Association for Computational Linguistics, Punta Cana, Dominican Republic.

Taisiya Glushkova, Chrysoula Zerva, Ricardo Rei, and André F. T. Martins. 2021. Uncertainty-aware machine translation evaluation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3920–3938, Punta Cana, Dominican Republic. 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, Elena Voita, and André Martins. 2023. Looking for a needle in a haystack: A comprehensive study of hallucinations in neural machine translation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1059–1075, Dubrovnik, Croatia. Association for Computational Linguistics.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330. PMLR.

Geoffrey Hinton, Oriol Vinyals, and Jeffrey Dean. 2015. Distilling the knowledge in a neural network. In *NIPS Deep Learning and Representation Learning Workshop*.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339, Melbourne, Australia. Association for Computational Linguistics.

Julia Ive, Lucia Specia, Sara Szoc, Tom Vanallemeersch, Joachim Van den Bogaert, Eduardo Farah, Christine Maroti, Artur Ventura, and Maxim Khalilov. 2020. A post-editing dataset in the legal domain: Do we underestimate neural machine translation quality? In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3692–3697, Marseille, France. European Language Resources Association.

Eric Jang, Shixiang Gu, and Ben Poole. 2017. Categorical reparameterization with gumbel-softmax. In *International Conference on Learning Representations*.

Xiaoqian Jiang, Melanie Osl, Jihoon Kim, and Lucila Ohno-Machado. 2011. Calibrating predictive model estimates to support personalized medicine. *Journal of the American Medical Informatics Association*, 19(2):263–274.

Alex Kendall and Yarin Gal. 2017. What uncertainties do we need in bayesian deep learning for computer vision? In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Fabio Kepler, Jonay Trénous, Marcos Treviso, Miguel Vera, and André F. T. Martins. 2019. OpenKiwi: An open source framework for quality estimation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 117–122, Florence, Italy. Association for Computational Linguistics.

Ondrej Klejch, Eleftherios Avramidis, Aljoscha Burchardt, and Martin Popel. 2015. MT-ComparEval: Graphical evaluation interface for Machine Translation development. *The Prague Bulletin of Mathematical Linguistics*, 104.

Goro Kobayashi, Tatsuki Kuribayashi, Sho Yokoi, and Kentaro Inui. 2020. Attention is not only a weight: Analyzing transformers with vector norms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7057–7075, Online. Association for Computational Linguistics.

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

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain. Association for Computational Linguistics.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver. Association for Computational Linguistics.

Dan Kondratyuk and Milan Straka. 2019. 75 languages, 1 model: Parsing Universal Dependencies universally. 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 2779–2795, Hong Kong, China. Association for Computational Linguistics.

Volodymyr Kuleshov, Nathan Fenner, and Stefano Ermon. 2018. Accurate uncertainties for deep learning using calibrated regression. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2796–2804. PMLR.

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.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Alon Lavie and Michael Denkowski. 2009. The meteor metric for automatic evaluation of machine translation. *Machine Translation*, 23:105–115.

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

Vladimir Iosifovich Levenshtein. 1966. Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady*, 10(8):707–710.

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

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

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

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

Ilya Loshchilov and Frank Hutter. 2018. Decoupled weight decay regularization. In *Proceedings of the 6th International Conference on Learning Representations (ICLR)*.

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.

Chris J. Maddison, Andriy Mnih, and Yee Whye Teh. 2017. The concrete distribution: A continuous relaxation of discrete random variables. In *International Conference on Learning Representations*.

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.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc.

Maciej Modrzejewski, Miriam Exel, Bianka Buschbeck, Thanh-Le Ha, and Alexander Waibel. 2020. Incorporating external annotation to improve named entity translation in NMT. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 45–51, Lisboa, Portugal. European Association for Machine Translation.

Alexander Molchanov. 2020. PROMT systems for WMT 2020 shared news translation task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 248–253, Online. Association for Computational Linguistics.

Mahdi Pakdaman Naeini, Gregory F. Cooper, and Milos Hauskrecht. 2015. Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, AAAI'15, page 2901–2907. AAAI Press.

Graham Neubig. 2013. Travatar: A forest-to-string machine translation engine based on tree transducers. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 91–96, Sofia, Bulgaria. Association for Computational Linguistics.

Graham Neubig, Zi-Yi Dou, Junjie Hu, Paul Michel, Danish Pruthi, and Xinyi Wang. 2019. compare-mt: A tool for holistic comparison of language generation systems. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 35–41, Minneapolis, Minnesota. Association for Computational Linguistics.

Nathan Ng, Kyra Yee, Alexei Baevski, Myle Ott, Michael Auli, and Sergey Edunov. 2019. Facebook FAIR's WMT19 news translation task submission. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 314–319, Florence, Italy. Association for Computational Linguistics.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 160–167, Sapporo, Japan. Association for Computational Linguistics.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

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.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. 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.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the*

2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. 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.

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.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.

Adrian E. Raftery, Tilmann Gneiting, Fadoua Balabdaoui, and Michael Polakowski. 2005. Using bayesian model averaging to calibrate forecast ensembles. *Monthly Weather Review*, 133(5):1155 – 1174.

Gema Ramírez-Sánchez, Jaume Zaragoza-Bernabeu, Marta Bañón, and Sergio Ortiz-Rojas. 2020. Bifixer and Bicleaner: two open-source tools to clean your parallel data. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 291–298, Lisboa, Portugal. European Association for Machine Translation.

Tharindu Ranasinghe, Constantin Orasan, and Ruslan Mitkov. 2020. TransQuest: Translation quality estimation with cross-lingual transformers. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5070–5081, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Vikas Raunak, Matt Post, and Arul Menezes. 2022. SALTED: A framework for SAlient long-tail translation error detection. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5163–5179, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ricardo Rei, Ana C Farinha, José G. C. de Souza, Pedro G. Ramos, André F. T. Martins, Luisa Coheur, and Alon Lavie. 2022a. Searching for Cometingo: The Little Metric That Could. In *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, Ghent, Belgium. Association for Computational Linguistics.

Ricardo Rei, Ana C Farinha, Chrysoula Zerva, Daan van Stigt, Craig Stewart, Pedro G Ramos, Taisiya Glushkova, André Martins, and Alon Lavie. 2021a. Are References Really Needed? Unbabel-IST 2021 Submission for the Metrics Shared Task. In *Proceedings of the Sixth Conference on Machine Translation*, Online. Association for Computational Linguistics.

Ricardo Rei, Ana C Farinha, Craig Stewart, Luisa Coheur, and Alon Lavie. 2021b. MT-Telescope: An interactive platform for contrastive evaluation of MT systems. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 73–80, Online. Association for Computational Linguistics.

Ricardo Rei, Nuno M. Guerreiro, Marcos Treviso, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2023. The inside story: Towards better understanding of machine translation neural evaluation metrics. In *Proceedings of the 61th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Toronto, Canada. Association for Computational Linguistics.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022b. 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, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020a. 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, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020b. Unbabel's participation in the WMT20 metrics shared task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 911–920, 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. 2022c. 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.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4512–4525, Online. Association for Computational Linguistics.

Marco Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 97–101, San Diego, California. Association for Computational Linguistics.

F. Schroff, D. Kalenichenko, and J. Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 815–823.

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.

Rico Sennrich and Barry Haddow. 2016. Linguistic input features improve neural machine translation. In *Proceedings of the First Conference on Machine Translation: Volume 1, Research Papers*, pages 83–91, Berlin, Germany. Association for Computational Linguistics.

Christophe Servan, Alexandre Bérard, Zied Elloumi, Hervé Blanchon, and Laurent Besacier. 2016. Word2Vec vs DBnary: Augmenting METEOR using vector representations or lexical resources? In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1159–1168, Osaka, Japan. The COLING 2016 Organizing Committee.

Hiroki Shimanaka, Tomoyuki Kajiwara, and Mamoru Komachi. 2018a. Metric for automatic machine translation evaluation based on universal sentence representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 106–111, New Orleans, Louisiana, USA. Association for Computational Linguistics.

Hiroki Shimanaka, Tomoyuki Kajiwara, and Mamoru Komachi. 2018b. RUSE: Regressor using sentence embeddings for automatic machine translation evaluation. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 751–758, Belgium, Brussels. Association for Computational Linguistics.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *In Proceedings of Association for Machine Translation in the Americas*, pages 223–231.

Lucia Specia, Frédéric Blain, Marina Fomicheva, Erick Fonseca, Vishrav Chaudhary, Francisco Guzmán, and André F. T. Martins. 2020. Findings of the WMT 2020 shared task on quality estimation. In *Proceedings of the Fifth Conference on Machine Translation*, pages 743–764, Online. Association for Computational Linguistics.

Lucia Specia, Frédéric Blain, Marina Fomicheva, Chrysoula Zerva, Zhenhao Li, Vishrav Chaudhary, and André F. T. Martins. 2021. Findings of the WMT 2021 shared task on quality estimation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 684–725, Online. Association for Computational Linguistics.

Lucia Specia, Kim Harris, Frédéric Blain, Aljoscha Burchardt, Viviven Macketanz, Inguna Skadina, Matteo Negri, , and Marco Turchi. 2017. Translation quality and productivity: A study on rich morphology languages. In *Machine Translation Summit XVI*, pages 55–71, Nagoya, Japan.

Lucia Specia, Marco Turchi, Nicola Cancedda, Nello Cristianini, and Marc Dymetman. 2009. Estimating the sentence-level quality of machine translation systems. In *Proceedings of the 13th Annual conference of the European Association for Machine Translation*, Barcelona, Spain. European Association for Machine Translation.

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.

Kosuke Takahashi, Katsuhito Sudoh, and Satoshi Nakamura. 2020. Automatic machine translation evaluation using source language inputs and cross-lingual language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3553–3558, Online. Association for Computational Linguistics.

Shimin Tao, Su Chang, Ma Miaomiao, Hao Yang, Xiang Geng, Shujian Huang, Min Zhang, Jiaxin Guo, Minghan Wang, and Yinglu Li. 2022. CrossQE: HW-TSC 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 646–652, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Andre Tättar and Mark Fishel. 2017. bleu2vec: the painfully familiar metric on continuous vector space steroids. In *Proceedings of the Second Conference on Machine Translation*, pages 619–622, Copenhagen, Denmark. Association for Computational Linguistics.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy. 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.

Marcos Treviso, Nuno M. Guerreiro, Ricardo Rei, and André F. T. Martins. 2021. IST-unbabel 2021 submission for the explainable quality estimation shared task. In *Proceedings of the 2nd Workshop on Evaluation and Comparison of NLP Systems*, pages 133–145, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Arata Ugawa, Akihiro Tamura, Takashi Ninomiya, Hiroya Takamura, and Manabu Okumura. 2018. Neural machine translation incorporating named entity. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3240–3250, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Yu Wan, Dayiheng Liu, Baosong Yang, Tianchi Bi, Haibo Zhang, Boxing Chen, Weihua Luo, Derek F. Wong, and Lidia S. Chao. 2021. RoBLEURT submission for WMT2021 metrics task. In *Proceedings of the Sixth Conference on Machine Translation*, pages 1053–1058, Online. Association for Computational Linguistics.

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.

Changhan Wang, Anirudh Jain, Danlu Chen, and Jiatao Gu. 2019. VizSeq: a visual analysis toolkit for text generation tasks. 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): System Demonstrations, pages 253–258, Hong Kong, China. Association for Computational Linguistics.

Weiyue Wang, Jan-Thorsten Peter, Hendrik Rosendahl, and Hermann Ney. 2016. CharacTer: Translation edit rate on character level. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 505–510, Berlin, Germany. Association for Computational Linguistics.

Yiming Yan, Tao Wang, Chengqi Zhao, Shujian Huang, Jiajun Chen, and Mingxuan Wang. 2023. Bleurt has universal translations: An analysis of automatic metrics by minimum risk training. In *Proceedings of the 61th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Toronto, Canada. Association for Computational Linguistics.

Chrysoula Zerva, Frédéric Blain, Ricardo Rei, Piyawat Lertvittayakumjorn, José G. C. de Souza, Steffen Eger, Diptesh Kanojia, Duarte Alves, Constantin Orăsan, Marina Fomicheva, André F. T. Martins, and Lucia Specia. 2022a. Findings of the WMT 2022 Shared Task on Quality Estimation. In *Proceedings of the Seventh Conference on Machine Translation*, Abu Dhabi. Association for Computational Linguistics.

Chrysoula Zerva, Taisiya Glushkova, Ricardo Rei, and André F. T. Martins. 2022b. Disentangling uncertainty in machine translation evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8622–8641, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Chrysoula Zerva, Daan van Stigt, Ricardo Rei, Ana C Farinha, José G. C. de Souza, Taisiya Glushkova, Miguel Vera, Fabio Kepler, and André Martins. 2021. IST-Unbabel 2021 Submission for the Quality Estimation Shared Task. In *Proceedings of the Sixth Conference on Machine Translation*, Online. Association for Computational Linguistics.

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

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

Michael Zhu and Suyog Gupta. 2018. To prune, or not to prune: Exploring the efficacy of pruning for model compression. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Workshop Track Proceedings*. OpenReview.net.

Appendix A

COMET Models Hyperparameters

Below you can find the hyperparameters used to train all the models presented in this thesis using the `unbabel-comet>=2.0.0` PyPI version of our framework.

```yaml
regression_metric:
  class_path: comet.models.RegressionMetric
  init_args:
    nr_frozen_epochs: 0.3
    keep_embeddings_frozen: True
    optimizer: AdamW
    encoder_learning_rate: 1.0e-05
    learning_rate: 1.0e-05
    layerwise_decay: 1.0
    encoder_model: XLM-RoBERTa
    pretrained_model: xlm-roberta-base
    pool: avg
    layer: mix
    layer_transformation: softmax
    layer_norm: False
    loss: mse
    dropout: 0.1
    batch_size: 32
    train_data:
      - PATH_TO_TRAIN_DATA.csv
    validation_data:
      - PATH_TO_WMT21_MQM_NEWS_ENDE.csv
      - PATH_TO_WMT21_MQM_NEWS_ENRU.csv
      - PATH_TO_WMT21_MQM_NEWS_ZHEN.csv
    hidden_sizes:
      - 2304
      - 1152
    activations: Tanh
```

Listing A.1: Hyperparameters used for the Estimator models presented in Section 3.4.

```yaml
ranking_metric:
  class_path: comet.models.RankingMetric
  init_args:
    nr_frozen_epochs: 0.0
    keep_embeddings_frozen: True
    optimizer: AdamW
    encoder_learning_rate: 1.0e-05
    learning_rate: 1.0e-05
    layerwise_decay: 1.0
    encoder_model: XLM-RoBERTa
    pretrained_model: xlm-roberta-base
    pool: avg
    layer: mix
    layer_transformation: softmax
    layer_norm: False
    dropout: 0.1
    batch_size: 32
    train_data:
      - PATH_TO_TRAIN_DATA.csv
    validation_data:
      - PATH_TO_WMT21_DARR_NEWS_ENDE.csv
      - PATH_TO_WMT21_DARR_NEWS_ENRU.csv
      - PATH_TO_WMT21_DARR_NEWS_ZHEN.csv
```

Listing A.2: Hyperparameters used for the Translation Ranking model presented in Section 3.4.

```
regression_metric:
  class_path: comet.models.RegressionMetric
  init_args:
    nr_frozen_epochs: 0.3
    keep_embeddings_frozen: True
    optimizer: AdamW
    encoder_learning_rate: 1.0e-06
    learning_rate: 1.5e-05
    layerwise_decay: 0.95
    encoder_model: XLM-RoBERTa
    pretrained_model: xlm-roberta-large
    pool: avg
    layer: mix
    layer_transformation: sparsemax
    layer_norm: False
    loss: mse
    dropout: 0.1
    batch_size: 16
    train_data:
      - PATH_TO_TRAIN_DATA.csv
    validation_data:
      - PATH_TO_WMT21_MQM_NEWS_ENDE.csv
      - PATH_TO_WMT21_MQM_NEWS_ENRU.csv
      - PATH_TO_WMT21_MQM_NEWS_ZHEN.csv
    hidden_sizes:
      - 3072
      - 1024
    activations: Tanh
```

Listing A.3: Hyperparameters used for the `wmt22-comet-da` metric presented in Section 3.5.

```
unified_metric:
  class_path: comet.models.UnifiedMetric
  init_args:
    nr_frozen_epochs: 0.3
    keep_embeddings_frozen: True
    optimizer: AdamW
    encoder_learning_rate: 1.0e-06
    learning_rate: 1.5e-05
    layerwise_decay: 0.95
    encoder_model: XLM-RoBERTa
    pretrained_model: microsoft/infoxlm-large
    sent_layer: mix
    layer_transformation: sparsemax
    word_layer: 24
    loss: mse
    dropout: 0.1
    batch_size: 16
    train_data:
      - PATH_TO_TRAIN_DATA.csv
    validation_data:
      - PATH_TO_WMT21_MQM_NEWS_ENDE.csv
      - PATH_TO_WMT21_MQM_NEWS_ENRU.csv
      - PATH_TO_WMT21_MQM_NEWS_ZHEN.csv
    hidden_sizes:
      - 3072
      - 1024
    activations: Tanh
    input_segments:
      - mt
      - src
    word_level_training: False
```

Listing A.4: Hyperparameters used for the `wmt22-cometkiwi-da` metric presented in Section 3.8.

```
unified_metric:
  class_path: comet.models.UnifiedMetric
  init_args:
    nr_frozen_epochs: 0.3
    keep_embeddings_frozen: True
    optimizer: AdamW
    encoder_learning_rate: 1.0e-06
    learning_rate: 1.5e-05
    layerwise_decay: 0.95
    encoder_model: XLM-RoBERTa
    pretrained_model: xlm-roberta-large
    sent_layer: mix
    layer_transformation: sparsemax
    word_layer: 24
    loss: mse
    dropout: 0.1
    batch_size: 16
    train_data:
      - PATH_TO_TRAIN_DATA.csv
    validation_data:
      - PATH_TO_WMT21_MQM_NEWS_ENDE.csv
      - PATH_TO_WMT21_MQM_NEWS_ENRU.csv
      - PATH_TO_WMT21_MQM_NEWS_ZHEN.csv
    hidden_sizes:
      - 3072
      - 1024
    activations: Tanh
    input_segments:
      - mt
      - src
      - ref
    word_level_training: False
```

Listing A.5: Hyperparameters used for the UNITE metric presented in Chapter 4.

Language Pair	SIZE
zh-en	126947
en-de	121420
de-en	99183
en-zh	90805
ru-en	79280
en-ru	62749
en-cs	60937
fi-en	46145
en-fi	34335
tr-en	30186
et-en	29496
cs-en	27847
en-mr	26000
de-cs	13804
en-et	13376
pl-en	11816
en-pl	10572
lt-en	10315
en-ja	9578
gu-en	9063
si-en	9000
ro-en	9000
ne-en	9000
en-lt	8959
ja-en	8939
en-kk	8219
en-ta	7890
ta-en	7577
en-gu	6924
kk-en	6789
de-fr	6691
en-lv	5810
en-tr	5171
km-en	4722
ps-en	4611
fr-de	3999
Total	1027155

Table A.1: Number of direct assessments per language pair used to train `wmt22-comet-da`, `wmt22-cometkiwi-da` (Chapter 3) and the UNITE model used in Chapter 4

Appendix B

Evaluating Uncertainty

Two crucial aspects to take into account when evaluating uncertainty-aware systems are: (i) the system should not harm the predictive accuracy compared to a system without uncertainty and (ii) the uncertainty estimate should reflect the failure probability of the system well, meaning that the system "knows when it does not know." In what follows, we assume a test or validation set $\mathcal{D} = \{\langle s_j, t_j, \mathcal{R}_j, q_j^* \rangle\}_{j=1}^{|\mathcal{D}|}$with input tuples with a source s, a translation t, a set of reference translations $\mathcal{R} = \{r_1, \ldots, r_{|\mathcal{R}|}\}$ and a ground truth scores q^*.

consisting of examples together with their ground truth quality scores.

Calibration Error One way of understanding if models can be trusted is analyzing whether they are *calibrated* (Raftery et al., 2005; Jiang et al., 2011; Kendall and Gal, 2017), that is, if the confidence estimates of its predictions are aligned with the empirical likelihoods (Guo et al., 2017). In classification tasks, this is assessed by the *expected calibration error* (ECE; Naeini et al. 2015), which has been generalized to regression by (Kuleshov et al., 2018).

It is defined as:

$$\mathrm{ECE} = \frac{1}{M} \sum_{b=1}^{M} |\mathrm{acc}(\gamma_b) - \gamma_b|, \tag{1}$$

where each b is a bin representing a confidence level γ_b, and $\mathrm{acc}(\gamma_b)$ is the fraction of times the ground truth q^* falls inside the confidence interval $I(\gamma_b)$:

$$\mathrm{acc}(\gamma_b) = \frac{1}{|\mathcal{D}|} \sum_{\langle s,t,\mathcal{R},q^* \rangle \in \mathcal{D}} \mathbb{1}(q^* \in I(\gamma_b)). \tag{2}$$

We use this metric with $M = 100$.

Negative log-likelihood To evaluate parametric methods that represent the full distribution $\hat{p}_Q(q)$, we can use a single metric that captures both accuracy and uncertainty, the average negative log-likelihood of the ground truth quality scores according to the model:

$$\mathrm{NLL} = -\frac{1}{|\mathcal{D}|} \sum_{\langle s,t,\mathcal{R},q^* \rangle \in \mathcal{D}} \log \hat{p}(q^* \mid \langle s, t, \mathcal{R} \rangle). \tag{3}$$

This metric penalizes predictions that are accurate but have high uncertainty (since they will become flat distributions with low probability everywhere), and even more severely incorrect predictions with high confidence (as they will be peaked in the wrong location), but is more forgiving to predictions that are inaccurate but have high uncertainty.

Sharpness The metrics above do not sufficiently account for how "tight" the uncertainty interval is around the predicted value, and thus might generally favour predictors that produce wide and uninformative confidence intervals. To guarantee useful uncertainty estimation, confidence intervals should not only be calibrated, but also sharp. We measure sharpness using the predicted variance $\hat{\sigma}^2$, as defined in (Kuleshov et al., 2018):

$$\mathrm{sha}(\hat{p}_Q) = \frac{1}{|\mathcal{D}|} \sum_{\langle s,t,\mathcal{R}\rangle \in \mathcal{D}} \hat{\sigma}^2. \tag{4}$$

Pearson correlations As shown by Ashukha et al. (2020), NLL and ECE alone might not be enough to evaluate uncertainty-aware systems. Therefore, we complement the indicators above with two Pearson correlations involving the system's predictions and the ground truth quality scores coming from human judgements. The first, which we call the **predictive Pearson score** (PPS), is useful to assess the predictive accuracy of the system, regardless of the uncertainty estimate – it is the Pearson correlation $r(q^*, \hat{\mu})$ between the ground truth quality scores q^* and the average system predictions $\hat{\mu}$ in the dataset $\mathcal{D}$ (for the baseline point estimate system, we use $\hat{q}$ instead of $\hat{\mu}$). We expect this score to be similar to the baseline or slightly better due to the ensemble effect. The second is the **uncertainty Pearson score** (UPS) $r(|q^* - \hat{\mu}|, \hat{\sigma})$, which measures the alignment between the prediction errors $|q^* - \hat{\mu}|$ and the uncertainty estimates $\hat{\sigma}$. Note that achieving a high UPS is much more challenging – a model with a very high score would know how to correct its own predictions to obtain perfect accuracy. We confirm this claim later in our experiments.

		PPS ↑	UPS ↑	NLL ↓	ECE ↓	Sha. ↓
EN-DE	MCD	0.576	0.284	1.330	0.014	0.645
	DE	0.581	0.246	1.364	0.023	0.523
	Basel.	0.576	-	1.337	0.079	0.845
EN-ZH	MCD	0.333	0.064	1.779	0.024	0.701
	DE	0.354	0.477	1.435	0.020	0.762
	Basel.	0.329	-	1.570	0.090	1.342
EN-TA	MCD	0.658	0.015	1.226	0.022	0.585
	DE	0.675	0.068	1.200	0.018	0.564
	Basel.	0.655	-	1.237	0.028	0.691
ZH-EN	MCD	0.314	0.109	1.628	0.015	0.971
	DE	0.319	0.174	1.591	0.016	0.928
	Basel.	0.313	-	1.580	0.059	1.374
EN-JA	MCD	0.640	0.165	1.237	0.011	0.591
	DE	0.651	0.093	1.225	0.015	0.556
	Basel.	0.636	-	1.259	0.035	0.725
EN-CS	MCD	0.691	0.207	1.163	0.013	0.548
	DE	0.729	0.163	1.100	0.013	0.455
	Basel.	0.695	-	1.172	0.036	0.608
EN-RU	MCD	0.536	0.142	1.378	0.021	0.767
	DE	0.578	0.139	1.320	0.023	0.670
	Basel.	0.532	-	1.383	0.041	0.925
EN-PL	MCD	0.611	0.199	1.275	0.015	0.650
	DE	0.650	0.176	1.224	0.012	0.581
	Basel.	0.608	-	1.301	0.042	0.783
EN-IU	MCD	0.300	0.223	1.600	0.016	1.016
	DE	0.308	0.319	1.682	0.026	1.052
	Basel.	0.292	-	1.594	0.077	1.410

Table B.1: Results for segment-level DA prediction. Underlined numbers indicate the best result for each language pair and evaluation metric.

		PPS ↑	UPS ↑	NLL ↓	ECE ↓	Sha. ↓
EN-DE	MCD	0.765	0.384	1.054	0.023	0.325
	DE	0.703	0.408	1.110	0.017	0.406
	Basel.	0.761	-	1.052	0.120	0.478
DE-EN	MCD	0.769	0.475	0.964	0.029	0.329
	DE	0.702	0.498	1.100	0.040	0.330
	Basel.	0.767	-	1.046	0.140	0.469
EN-LV	MCD	0.778	0.376	1.209	0.020	0.284
	DE	0.709	0.377	1.064	0.022	0.328
	Basel.	0.772	-	1.017	0.108	0.454
EN-CS	MCD	0.753	0.173	1.097	0.038	0.413
	DE	0.672	0.216	1.222	0.024	0.536
	Basel.	0.752	-	1.076	0.050	0.498

Table B.2: Results for segment-level HTER prediction in QT21. Underlined numbers indicate the best result for each language pair and evaluation metric.