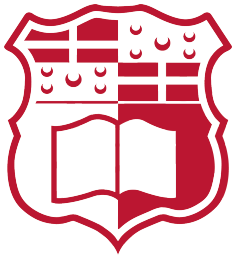

COMET for Low-Resource Machine Translation Evaluation

A case study of English-Maltese and Spanish-Basque

Júlia Falcão

Supervised by Dr Claudia Borg
Co-supervised by Dr Nora Aranberri
Advised by Kurt Abela, M.Sc.

Department of Artificial Intelligence
Faculty of Information and Communication Technology
University of Malta

October, 2023

A dissertation submitted in partial fulfilment of the requirements for the degree of M.Sc. in Human Language Science and Technology (HLST).

WWW.UM.EDU.MT

First edition, October 13, 2023

Acknowledgements

Aos meus pais, Patrícia e Marcelo, e à minha irmã, Marcela. Só Deus sabe o quanto vocês sofreram com a minha ausência pra eu estar aqui na Europa estudando o que eu queria. Obrigada infinitamente pelo apoio, e por me lembrar sempre que eu podia desistir e voltar pra baixo das asas de vocês se ficasse difícil demais. E, igualmente, a todos da minha família que me apoiaram nessa missão e me encorajaram a continuar seguindo meus sonhos.

À minha melhor amiga nesse mundo, Julianna, de quem eu nunca vou me acostumar a morar longe. Esse mestrado só teria sido melhor se eu pudesse te encontrar nos intervalos, tomar um café e comer uma paçoquinha.

À Tatiane, co-protagonista dos meus momentos mais felizes em plena reta final dessa tese. Que bom que você tava aqui.

A todas as amigas que me motivaram a estudar logo pra gente poder sair depois, e preencheram de risadas o meu tempo livre. Muri, Rebeca, Laís, Julia, Júlia, Carol, e mais tantas brasileiras que eu sou grata demais por ter conhecido.

Aos amigos de longa data que eu agora acompanho de longe, mas contando os dias pra voltar e a gente poder dar rolê de novo. Luíza, Sophia, Wallace, Raphael, Raffael.

To all of the incredible friends I made in this continent, the gringos that I love most dearly. I might have been lonely at times but I was never alone. Nastya, Lea, Rosa, Jelena, Anar, Nils, Manu, Uxue, Tabara, Geneviève, Marta, Martin, Hannah, Annika, Agnieszka, Jasmin, Flávia, Sarah, Natalie, Margherita. Thank you for all the love you gave to me; you don't know how much I needed it.

To my stellar team of supervisors, Claudia, Nora, and Kurt. At first I didn't know what to do and then I didn't know how to do it, and I was often overwhelmed, but by the end of every single one of our meetings I was much calmer than when I joined, more confident, more excited about this project, and ready to keep going.

To every participant who contributed to our translation evaluation campaign, and to those who helped spread the word about it as well. To Nora, Claudia, and Marthese Borg, for translating our instructions into Basque and Maltese for the evaluation campaign website, and to everyone in the UPV/EHU and the UM who gave us feedback on the system. To Gorka Labaka and Kurt Abela, for offering their new MT models for us to evaluate.

Finally, I am endlessly thankful to the EMLCT programme for awarding me a scholarship, and the opportunity of a lifetime to study Computational Linguistics in two different countries. It has been a dream come true, and as a bonus, I get to put two diplomas up on my wall.

Abstract

Translation quality is a largely subjective concept, but in Machine Translation, it needs to be measurable. Human judgements are regarded as the gold standard of evaluation methods, but they are expensive and time-consuming to obtain, so the field has turned to automatic metrics, such as BLEU, which measures the lexical overlap between the translation candidate and one or more reference translations. However, lexical overlap is not all there is to a good translation, and BLEU has repeatedly been shown to correlate poorly with human judgements of quality. A new paradigm has emerged in recent years: trainable metrics, based on neural networks that directly predict quality scores of human judgements, have been topping the ranks in the latest meta-evaluation studies. However, as they need to be trained on annotated parallel data, these metrics have limited language support, and so under-resourced languages are mostly left out.

In this work, we look at the most prominent trainable evaluation system proposed for MT so far, the COMET framework, and take English–Maltese and Spanish–Basque as a case study to investigate the extent of COMET's language support restrictions: how well can it evaluate languages outside of its training data, and languages not supported by its underlying encoder, as is the case of Maltese and Basque? We run a crowd-based evaluation campaign to collect human judgements, and then use this data to analyze the performance of COMET out of the box. We also explore potential avenues of improvement: by fine-tuning existing models, or training new models from scratch. Our results, based on correlations between human evaluations and metric outputs, attest to the potential of fine-tuning to improve existing models, but also indicate that COMET is highly susceptible to the distribution of scores in its training data, which is especially concerning in low-resource scenarios.

This dissertation is a step towards the inclusion of under-resourced languages in the development of better metrics for MT evaluation, and we also release our anonymized campaign results to public for future works.

Contents

List of Figures

List of Tables

List of Abbreviations

NLP Natural Language Processing
MT Machine Translation
NMT Neural Machine Translation
SMT Statistical Machine Translation
LM Language Model
GT Google Translate
NLLB No Language Left Behind
UPV/EHU Universidad del País Vasco/Euskal Herriko Unibertsitatea
UM University of Malta
DA Direct Assessment
DQF Dynamic Quality Framework
MQM Multidimensional Quality Metrics
WER Word Error Rate
TER Translation Edit Rate
HTER Human-targeted Translation Edit Rate
BLEU Bilingual Evaluation Understudy
QE Quality Estimation
WMT Conference on Machine Translation

1

Introduction

In the book "Nineteen Ways of Looking at Wang Wei", American translator Eliot Weinberger takes a close look at 19 different translations of a thousand year-old, four-line Chinese poem. Each translation is infused not only with its translator's personal interpretation of the work, but also a multitude of linguistic decisions to transpose that into the target language—some of which the author criticizes thoroughly. While literary translation—especially that of poetry—is quite an extreme case, even the most literal sentence in an instruction manual may be translated a number of different ways. The mapping between two languages is never 1:1, and so the "correctness" of a translation is not binary.

The definition of "translation quality" is a thorny issue for scholars of Translation Studies (Snell-Hornby, 1992); however, as part of a largely theoretical field, translation criticism does not necessarily aim to produce fully objective assessments. In Machine Translation (MT), on the other hand, "translation quality" needs to be *measurable*. Researchers and developers working on MT systems need reliable ways of quantifying how well their systems are performing, and to compare them to other systems. However, the "correctness" of MT outputs cannot be measured in terms of "accuracy", because there is no ground truth; the subjective nature of "correctness" in this context makes evaluation an ongoing challenge.

Therefore, lacking actual ground truth results, the field of MT relies on the expertise of human professionals: human translations are taken as "reference" translations against which MT outputs can be compared, and human judgement is widely regarded as the gold standard of evaluation methods (Bojar et al., 2016a). This is, of course, also flawed; as put by Zehnalová (2013), "*why should we expect then that a necessarily subjective evaluation of something subjective by its very nature will be perfectly objective?*"

Evaluation was performed only by humans in the early decades of MT development, but since the early 2000s, and notoriously since the creation of BLEU (Papineni et al., 2002), automatic evaluation has taken over. BLEU and most other automatic methods are *reference-based*: they rate an MT output by comparing it to one or more reference translations. This comparison usually involves calculating the overlap between the segments through performance metrics such as precision, recall and the F-score, but adapted to handle specificities of textual content.

This process evokes a number of issues that have been investigated in MT meta-evaluation studies for decades now, which we will discuss in depth in the next chapter. First, references are simply human translations that could themselves be subjected to quality assessments. Secondly, most automatic evaluation methods simply measure the lexical overlap between a translation and the reference, without access to any computational resources about the languages, their vocabulary and grammatical structure. Therefore, these lexical metrics can only enforce that the translations should be as close as possible to the references, punishing variability in word choices. Finally, but most importantly, lexical metrics have repeatedly been shown to correlate poorly with various kinds of human judgements of translation quality; they often underestimate systems which humans find better and vice versa, and also often fail to accurately discriminate between high-performing systems.

As research in MT evolves, now mostly dominated by Neural Machine Translation (NMT) systems, the demand for reliable evaluation only grows, and lexical metrics have fallen behind. A new family of methods has emerged in recent years, known as *trainable metrics*, which are based on neural networks trained directly to predict human judgements of quality. Their power lies in that they can go beyond the lexical level by generating contextualized, meaning representations of the input segments. The most prominent trainable evaluation system so far has been COMET, which has topped the rankings in recent meta-evaluations (Freitag et al., 2021b, 2022; Kocmi et al., 2021; Mathur et al., 2020b); these studies analyze the performance of automatic metrics by measuring their correlation with human judgements obtained in manual evaluation campaigns.

However, trainable methods like COMET have a limitation with regards to the languages that they can evaluate, as they have to encode the input segments into embeddings, and have to be trained on previous human assessments to learn to predict them. This limitation has not yet been the subject of thorough analysis, and trainable methods are most often evaluated on the same languages they were trained on. This dissertation aims to contribute to the current research in MT meta-evaluation by looking at this gap. If trainable metrics are to be widely adopted for MT evaluation, in place of simpler lexical metrics, how do we extend them to evaluate new, under-resourced languages?

1.1 | Motivation

This dissertation is a meta-evaluation study of COMET and its applicability for under-resourced languages, and to that end, we conducted a case study centered around two language pairs[1]: English–Maltese and Spanish–Basque. The motivations for this choice are both linguistic and sociolinguistic.

Maltese is a Semitic language spoken primarily in Malta, where it is an official language alongside English. The language mixes elements of Arabic, Italian, and English, due to the long history of invasions and colonization in the Maltese islands (Brincat, 2005). It is the only Semitic language written in the Latin script, and also the only Semitic language with official status in the European Union (Rosner and Joachimsen, 2012).

Basque is the native language of the Basque people, spoken mainly in Euskal Herria, the cultural territory which comprises the Basque Country autonomous community and parts of Navarre in Spain, where it holds official status alongside Spanish, and also the French Basque Country, where it is not recognized (Eberhard et al., 2023).

Maltese and Basque are in relatively similar positions: they are both widely used in their native regions, but nonetheless threatened by the overwhelming and ever growing presence of major co-official languages, namely English and Spanish. The scarcity of digital resources for Maltese and Basque—resources which are readily available for English and Spanish—leaves the minoritary languages in an even more disadvantageous position.

The development of NLP technologies—such as spell checkers, automatic translators and voice assistants, to name a few—is essential to ensure the survival of minoritary languages in the digital age, and contributes, alongside language preservation politics on the part of the local governments, to keep the languages alive and in vigorous use. Machine translation is often highlighted as a key application in the effort to promote linguistic diversity through NLP, because high-quality automatic translation allows already existing online content to be made available in minoritary languages as well (Bartolo, 2022; Wetzel, 2018). However, MT is also particularly demanding due to its data-hungry nature: NMT systems usually need to be trained on millions of parallel segments to be able to perform competitively (Koehn and Knowles, 2017), and this scale of data simply is not available for most languages of the world.

Joshi et al. (2020)[2] conducted a large-scale study on language resource disparity in NLP, and classified Maltese and Basque as languages for which there is "light at the end of the tunnel": dedicated NLP efforts have collected significant datasets which grant these

[1] In accordance with MT literature, we use "language pairs" in the sense of "translation direction": X–Y is the same as $X \rightarrow Y$ (translation from X into Y), and $X \leftrightarrow Y$ denotes both directions.

[2] `https://microsoft.github.io/linguisticdiversity`

languages the potential to maintain an online presence, provided that there is ongoing research on new language technologies. Basque, particularly, was reported as possessing more unlabeled than labeled data; this gap has already been exploited in approaches such as Unsupervised Machine Translation (Artetxe et al., 2017, 2018).

Basque and Maltese are available on a handful of commercial MT systems, and there is active research on improving MT technologies for these languages, particularly centered around Maltese to/from English, and Basque to/from Spanish, English, and sometimes French, Catalan and Galician.[3] The availability of parallel datasets for Maltese has increased significantly since it became an official EU language when Malta joined the EU in 2004; nowadays, most of the available corpora for Maltese comprise official legal-administrative EU documents. In turn, the lack of corpora in other domains, such as works of literary fiction or casual speech, may hinder the performance of NLP applications trained mostly on legal-administrative corpora (Rosner and Borg, 2022).

In the case of English–Maltese and Spanish–Basque, automatic translation is particularly complex due to the fact that the languages are unrelated: while English is a West-Germanic language, Maltese is in the Semitic family, descending from Siculo-Arabic (Agius, 1981); Basque is remarkably the only surviving pre-Indo-European language in Europe, and is considered a language isolate, while Spanish is a Romance language (Eberhard et al., 2023; Trask, 2013). It is easier to produce high-quality MT between languages that are similar, but with large morphological and grammatical differences, and especially in scenarios with relatively limited amounts of data, MT systems struggle to align the elements properly and any errors become more salient.[4] At the same time, translation between these pairs is all the more important since the languages have virtually no mutual intelligibility (Kolovratník et al., 2009; Popović et al., 2016).

At the same time, these language pairs potentially suffer the most from the overuse of BLEU as the sole evaluation metric. Maltese and Basque are morphologically rich languages, and BLEU strictly counts exact n-gram matches as correct; moreover, they are both languages with fairly flexible word order, and BLEU fails to catch long-range word order errors, as it matches n-grams of up to $n = 4$ only (Bisazza et al., 2021; Callison-Burch et al., 2006; Chatterjee et al., 2007; Tatman, 2019).

These factors increase even more the need for human evaluations, which is doubtlessly the most reliable way to verify translation quality, but they are too expensive to obtain with the frequency they would be needed; therefore, MT research for these language pairs

[3] See "API Support" lists at `https://machinetranslate.org/maltese` and `https://machinetranslate.org/basque`, last accessed 03/10/2023.

[4] `https://cordis.europa.eu/article/id/131535-in-the-quest-for-excellence-in-machine-translation`, last accessed 10/10/2023.

stands to benefit greatly from better evaluation metrics.

1.2 | Aims and Objectives

This dissertation adds to the current research in MT meta-evaluation, focusing on an analysis of a relatively new, trainable metric, COMET. The main research question that we focus on is the extension of COMET and its applicability for the evaluation of under-resourced language pairs. To this extent, we focus on Maltese and Basque, both of which, to the best of our knowledge, do not have a neural approach to evaluating machine translation output and thus to date always rely on metrics such as BLEU.

In order to achieve this, we look at the limitations in COMET's language support, investigate its generalization capabilities, and also explore the possibility of training new COMET metrics from scratch to obtain better evaluations for unsupported languages.

Therefore, we outlined the following objectives:

1. Familiarizing ourselves with the COMET framework, how to use it to evaluate MT outputs, and how to train new COMET models.
2. Defining the correlation metrics that would be used for meta-evaluation.
3. Gathering parallel, annotated data for preliminary tests and analysis.
4. Developing a system to host a crowd-based human evaluation campaign.
5. Gathering parallel datasets for English–Maltese and Spanish–Basque, and manually selecting segments that would be appropriate for crowd-based human evaluation.
6. Selecting MT systems to evaluate for each language pair, and implementing translation through them.
7. Selecting and testing encoders that support our languages, to plug into new COMET models.
8. Defining the quality control procedures and creating quality control items.
9. Launching the campaign and reaching out to bilingual speakers to participate and evaluate translations.
10. Pre-processing and filtering the raw results of the campaign.
11. Fine-tuning COMET on new datasets, and training new COMET models from scratch.

12. Evaluating the performance of new models in comparison with existing pre-trained models.

1.3 | Contributions

We first take a general look at the gap between supported and unsupported languages in the COMET framework. In many ways, this is akin to the gap between high- and low-resource languages in the context of MT. Parallel data is already scarce, and parallel data annotated with human judgements of quality is even harder to find. With the exception of a few low-resource languages included in the datasets, thanks to efforts to encourage more research towards low-resource languages in WMT competitions, most languages supported by COMET models can be considered at least medium- to high-resource.

We explore this restriction on language support by testing pre-trained and custom COMET models on new datasets: we used WMT data that had not been used to train COMET, and also obtained our own quality scores by conducting a manual evaluation campaign aimed at English/Maltese and Spanish/Basque bilingual speakers. We also attempt to improve the performance of the models by fine-tuning it on this new data, and by training entire new COMET models from scratch.

We evaluate the performance of all of these COMET models by using them to evaluate a test set of translations from different systems, and then measuring its correlation with human judgements, using Kendall's Tau and Spearman correlation coefficients. We also compare COMET to other lexical metrics in terms of correlation with human judgements.

The main contributions of this work are as follows:

- We report an analysis of how well COMET models perform when evaluating languages it has not previously seen in its training.
- We provide a preliminary estimation of how much data is needed for fine-tuning to improve a pre-trained COMET model in terms of correlation with human judgements.
- We carry out a human evaluation campaign for English–Maltese and Spanish–Basque, and make our results publicly available for future research.[5]
- We fine-tune existing COMET models, and train custom ones from scratch on top of publicly available cross-lingual encoders that specifically support Maltese and Basque, in order to evaluate their performance in comparison with human scores.

[5] `https://github.com/juliafalcao/direct-assessments`

1.4 | Document Structure

We begin in Chapter 2 with an overview of the theoretical background of this work, which covers the history of MT and MT evaluation, the main types of manual evaluation methods and automatic metrics, and then also delve deeper on the structure of the COMET framework.

Chapter 3 details our preliminary experiments with COMET models, which yielded valuable insights that helped guide our decisions for the next steps in the case study of our two language pairs. Chapter 4 describes the design of our manual evaluation campaign, where we collected translation quality scores in the form of direct assessments for English–Maltese and Spanish–Basque from bilingual speakers. In Chapter 5, we present the experiments we conducted with this assessment data, and discuss the implications of our results. Finally, in Chapter 6, we recapitulate the conclusions of our work and what we achieved, and briefly discuss its limitations as well as some directions for future research.

2

Background & Literature Overview

This chapter aims to provide all of the necessary theoretical background for the present dissertation, a meta-evaluation study on the applicability of COMET in low-resource MT evaluation scenarios.

We begin with an overview of the history of MT up to the present day. This overview is interlaced with a brief history of MT evaluation methods and metrics, highlighting how the demand for quick and simple evaluation methods only grew as MT evolved from specialized rule-based systems to large neural networks. Then, we delve into a literature review on MT evaluation, presenting a classification of methods for manual evaluation (Sec. 2.2) and also the main types of automatic methods (Sec. 2.3), and discussing the main advantages and shortcomings of each.

Once we have established the most important aspects of human and automatic evaluation, we clarify our motivation for focusing on COMET, a trainable, automatic metric that optimizes for correlation with human judgements. We provide greater detail on the structure of the framework, so that later we can look at its limitations.

2.1 | Machine Translation and its Evaluation

Although ideas about how to automate translation have been elicited since the 17th century, the true dawn of Machine Translation happened around the 1950s, driven by the needs of the Cold War (Hutchins, 2006b). The first proposals were based on successful advances in code breaking from World War II (Poibeau, 2017), and in 1954, IBM and Georgetown University demonstrated a system which automatically translated around 60 sentences from Russian to English (Hutchins, 2004).

It was a rule-based system, consisting only of 6 grammar rules and 260 lexical items in its vocabulary. The quality of the translations was good enough for a demonstration

which caused an unprecedentent media attention on the "electronic brains" that could "turn Russian into English".[1] However, as we now know better, rule-based systems are irrevocably restricted in that the mapping of one language to another through explicit grammar rules is immensely complex, if not impossible, because grammatical structures do not always align neatly across languages. The Georgetown–IBM demonstration, for example, avoided issues of article insertion when mapping Russian into English by including only a few English articles to fit the words present in the corpus (Hutchins, 2006a).

Unfortunately, this demonstration gave the public the wrong impression that fully automated systems for high-quality translation would be feasible much sooner than was actually realistic (Hutchins, 2006a). This led to a decade of more liberal funding for MT in the United States, which produced largely unsatisfying results, as researchers were faced with the inflexibility of rule-based systems in handling syntactic and semantic ambiguity, for example (Hutchins, 1986).

Funding was brought to an end in 1966 by the publication of an infamous report from the Automatic Language Processing Advisory Committee (ALPAC), which evaluated the state of MT—and computational linguistics as a whole at the time—, and concluded that there was "*no immediate or predictable prospect of useful machine translation*" (Pierce and Carroll, 1966, p. 32).

Among other shortcomings, the ALPAC report noted that there were no reliable measures of translation quality, and thus included a study that aimed to "*lay the foundations for a standard procedure for measuring the quality of scientific translations, whether human or mechanical*" (Carroll, 1966). This evaluation study was the main basis for ALPAC's conclusion that MT served no useful purpose without post-editing, and furthermore, that post-editing MT might be more costly than simply obtaining human translations (Hutchins, 1996).

Research in MT continued in a much smaller scale in the aftermath of the ALPAC report, but in the few remaining research groups, the focus shifted to more realistic goals, aiming for readability and fidelity to the original text, rather than for perfect translations. Besides, they were working on translating between other language pairs. The new systems were still rule-based, and there were more ambitious approaches, based on linguistic theories in syntax and semantics; nonetheless, there was a gradual realization that these approaches were not bringing solutions, and the limitations of rule-based MT seemed insurmountable (Hutchins, 1986).

At the time, the evaluation of these new systems was largely ad hoc, suffering from the lack of objective evaluation methods for human translations in the first place. MT

[1] `https://aclanthology.org/www.mt-archive.info/50/Georgetown-IBM-1954-reports.htm`, last accessed 18/09/2023.

outputs most often needed to be post-edited—revised by professional translators—to be useful, and the productivity of post-editing was measured in terms of time spent and also on the amount and severity of errors corrected (Slype, 1979). Naturally, post-editing productivity depends on the complexity of the original text, and also on the attitude of the translators—whether they saw MT as an aid or as an inhibitor of their creative expression. Some evaluations also compared raw MT outputs against post-edited versions and against human outputs, by rating them on "comprehensibility" and "clarity", amongst other criteria varying from project to project (Hutchins, 1986).

Eventually, given the unsatisfying results of rule-based MT, this classical approach gave way to new, corpus-based methods. In 1981, Makoto Nagao argued that the rule-based approach is inadequate for very different languages like English and Japanese, where there is little structural correspondence between a sentence and its translation (Nagao, 1984). His approach, inspired by the analogy principle in human thinking, later became known as example-based machine translation (EBMT); it used bilingual corpora to match fragments from the source sentence with real examples, and then the corresponding translated fragments were recombined to generate the target sentence (Somers, 1999). EBMT was a solid first step towards machine translation based on bilingual corpora, leaving behind carefully crafted grammatical rules and lexical resources.

Other sub-fields of NLP were following in a similar direction, reconciling that language is too complex to be encoded as a set of rules; instead, machines could be made to "learn the rules" automatically by looking at large corpora (Koehn, 2009). The revolution in MT came around the late 1980s with the invention of Statistical Machine Translation (SMT), a strategy for finding the most probable translation for a given input by analyzing the statistical relations in parallel data—matching pairs of source texts and human translations (Brown et al., 1988). This strategy leveraged the growing amounts of text available on the Internet, as well as the development of increasingly powerful computers, to do what computers do much better than humans: processing huge amounts of data (Poibeau, 2017).

After several years being dismissed as an "expensive failure" (Garvin, 1980), machine translation was back in the forefront of computational linguistics research. Much bigger systems were being developed, based on gradually more sophisticated statistical approaches, and so the demand for standardized evaluation methods only grew. Evaluation was mostly done manually, with the help of human translators, who were asked to judge MT outputs on attributes such as "comprehensibility", "adequacy" and "fluency". This scenario only started to change significantly with the widespread adoption of BLEU (Papineni et al., 2002), an automatic metric based on the precision score. It not long before publications started reporting only BLEU scores, rather than performing regular

human evaluations (Marie et al., 2021), even though the limitations of BLEU have been discussed since its early years (Melamed et al., 2003; Koehn and Monz, 2006). Critically, metrics based on lexical matching only compare the surface form of an MT output against one or more reference translations provided by humans, thus enforcing the specific wording of the reference and failing to capture any nuances which may lead to better or worse quality regardless of similarity to the reference. Such metrics have therefore been proven to correlate poorly with human judgements, an issue we will discuss further in Section 2.3.1.

The introduction of deep learning methods for translation brought about another revolution, and NMT has been the dominating approach in the field for the past decade, with the invention of increasingly powerful model architectures (Bahdanau et al., 2014; Kalchbrenner and Blunsom, 2013; Sutskever et al., 2014). Early NMT models consisted of Recurrent Neural Networks (RNNs) (Kalchbrenner and Blunsom, 2013), later evolved to sequence-to-sequence, encoder-decoder architectures (Sutskever et al., 2014), where the encoder is responsible for encoding the source sequence into a shared feature space, and the decoder, for predicting each next word in the target sequence. The attention mechanism was created to address the issue of maintaining context or handling long sequences, allowing the decoder to focus on more important parts of the sequence during decoding (Bahdanau et al., 2014). Finally, Vaswani et al. (2017) introduced the Transformer architecture, based on self-attention, which allows the model to capture long-range dependencies; transformers are also designed to be highly parallelizable, exploiting the processing power of GPUs.

The scale of machine translation models nowadays is beyond anything imaginable in the days of rule-based systems: huge neural networks with millions of parameters can be trained on terabytes of parallel texts. Recent advances in NMT research have made it potentially possible to translate between severely under-resourced languages, even without parallel data, by leveraging monolingual data to perform *unsupervised* or *zero-shot* machine translation (Artetxe et al., 2017; Lample et al., 2017, inter alia). In 2022, Google Translate extended its support to 24 under-resourced languages through a novel self-supervised method (Bapna et al., 2022; Caswell and Bapna, 2022).

The sheer size of NMT models now and the speed at which new architectures and techniques are being developed has substantially increased the demand for quick and efficient evaluation. Neural models need to be evaluated constantly and consistently during development, in order to validate the quality of the outputs after each training epoch, to guide hyperparameter selection, and to compare variants of the same model architecture. This only exarcebates the issue of MT researchers over-relying exclusively on BLEU, as well as similar automatic metrics with well-documented shortcomings, for

measuring system performance and reporting results.

The field of MT meta-evaluation is responsible for the study of existing evaluation metrics, usually by measuring their correlation with human judgements, and thus it also analyzes methods for collection of human judgements. Researchers in this field also work on the development of new metrics, in an attempt to provide better, more reliable tools for the evaluation of MT systems. The Conference on Machine Translation (WMT) has held annual shared tasks encouraging the development of new metrics for MT since 2008[2]; the performance of the metrics is evaluated with the human evaluation scores obtained in the main translation task. So while the main translation task reports on the current state of MT systems, the metrics task is the main reference in MT meta-evaluation.

2.2 | Manual Evaluation

Before any automatic metrics were conceived, MT outputs were evaluated only manually, with the help of human assessors. Although human judgements are widely regarded as the gold standard of evaluation methods (Koehn and Monz, 2006), they do not come without their own complications. While one assessor may look at two translations and prefer the one that sounds more natural, another assessor may find that the more literal but less natural-sounding translation is more faithful to the original text.

Human judgements of translation quality are inherently subjective, but when provided by trustworthy evaluators and constrained by a well-defined method, they yield valuable information that can greatly aid in the development of better MT systems. The most widely used methods nowadays for human evaluation can be divided into three groups, to be detailed in upcoming subsections: comparison tasks, attribute evaluation, and error analysis.

When it comes to the choice of task for a given manual evaluation campaign, there is an interesting tradeoff to be considered: the more information you wish to obtain from an evaluation, the harder the task will be for the assessors. Alternatively, simpler tasks yield less fine-grained information, but might make it possible to reach a larger amount of participants and thus obtain more evaluations. Therefore, it is always necessary to carefully consider what is reasonable to expect from the assessors, depending on who they will be and how the campaign will be organized.

Evaluation can be conducted at different levels of granularity, the most common being segment-level, where a "segment" is a piece of text containing one or more sentences—or sometimes not even a full sentence, which occurs in web crawled text with headers and

[2] `https://www.statmt.org/wmt08/shared-evaluation-task.html`, last accessed 23/09/2023.

bullet points, for example. Segments are most often evaluated out of context, though sometimes, preceding and subsequent segments are provided for context, if available. There is also document-level evaluation, where each item is a full text, usually a few paragraphs long. Word-level evaluation can also be conducted in the form of error analysis, where the translation is marked for errors in specific words or spans. The granularity level is another decision that will depend on the goals of an evaluation campaign and the availability of data and assessors.

Choosing the assessors also involves a number of decisions: for example, whether they need to be bilingual speakers, and whether they will be language professionals or crowd-sourced participants. All of these decisions are inevitably subjected to the availability of participants and the resources that the organizers have at their disposal; we will discuss several works in MT meta-evaluation that have analyzed the impact of some of these choices. The subject of manual evaluation is highly important because we will proceed with our study by performing our own evaluation campaign, and then analyzing COMET framework, which is based on predicting human judgements.

2.2.1 | Comparison Tasks

Comparison tasks consist of obtaining relative rankings for MT outputs, by showing N outputs to assessors and asking them to choose the best one, or to rank them by quality.

In the pairwise comparison method, evaluators are presented with a source sentence and two different MT outputs, and asked to choose the best one; they might also have the option to vote that the hypotheses are of equal quality. Ranking is also a comparison task, but with more MT outputs being assessed at once: evaluators either have to pick the best out of 3 or more translations, or to rank all of them from best to worst (Chatzikoumi, 2019).

Both of these tasks yield relative ranks: the results might attest that system A is better than system B; however, there is no measure of absolute quality, or the degree to which A is better. Their main advantage is that they are cognitively easier than the methods we will see in upcoming sections. The instructions are easy and straightforward, as judges must only evaluate translations against each other; they do not need to provide absolute scores of the quality of any translation, or identify any potential errors in them.

The ranking task allows for direct comparison of more systems at once, and thus requires less evaluators; one evaluation of 5 distinct MT outputs yields 10 pairwise comparisons. However, in terms of cognitive effort, it is considerably harder than pairwise comparison. Görög (2014) argues that the maximum amount of outputs that a judge can reliably rank is three, while WMT has used up to five outputs at a time—generating

a total of 10 pairwise comparisons—which they claim "seems to be a good compromise between efficiency and reliability" (Bojar et al., 2016a, p. 29).

2.2.2 | Attribute Evaluation

Attribute evaluation, unlike comparison tasks, aims to obtain an absolute score of the quality of an MT output, according to certain attributes. That is, each output will have its own score on a given scale, not in relation to other outputs, and these scores allow for conclusions about which system is better and how much better it is.

The term "attribute" refers to the different aspects in which one translation might be deemed better than another, since "quality" of translation cannot be defined concretely by itself. The ALPAC report (1966) defined that the main characteristics a translation should be judged on are "intelligibility" (or "comprehensibility") and "fidelity" (or "accuracy"), and claimed that these aspects are conceptually independent (Carroll, 1966). In the ARPA MT Initiative, the translation quality score was based on "adequacy", "fluency" and "comprehension" (White et al., 1994). "Adequacy" and "fluency" were the quality score attributes under the criteria developed by the Linguistics Data Consortium for the annual NIST Machine Translation Evaluation Workshop (LDC, 2005), where adequacy refers to how well the translation expresses the meaning of the source text, and fluency refers to the correctness of the translation in terms of grammaticality and proper use of the target language (White and O'Connell, 1994). Adequacy and fluency scores can be seen as representing the semantic and the syntactic appropriateness of a translation hypothesis, respectively (Banchs et al., 2015).

Attributes can be evaluated on a full continuous scale of 0–100, generally presented on a visual analogue scale. In order to try and make the task cognitively easier for the assessors, it can also be collected using interval-level scales, usually with five or seven points, presented as radio buttons with labeled categories; as an example, in Callison-Burch et al. (2007), assessors could evaluate a hypothesis on fluency by selecting "flawless", "good", "non-native", "disfluent" or "incomprehensible". However, this strategy might also prove counter-productive, forcing humans to choose between categories that might not accurately represent their judgements, as quality estimates are inherently continuous in nature (Graham et al., 2013). Direct estimation of adequacy and fluency separately on 5-point scales was the main evaluation method used during the first two editions of the WMT shared translation task (Callison-Burch et al., 2007; Koehn and Monz, 2006). Upon analysis, however, they found this method to be inconsistent, unpopular with the annotators, and hard to normalize: scores by individual annotators were distributed very differently and skewed in different directions, and the organizers found no clear way of

combining judgements from multiple annotators (Bojar et al., 2016a).

Moreover, researchers have found issues with the separation of quality judgements into attributes like adequacy and fluency. Denkowski and Lavie (2010) hypothesized that high correlation between these two attributes, as reported by Callison-Burch et al. (2007) on the 2017 WMT results, for example, may actually happen due to the attributes influencing each other. A translation must be considerably fluent in order to express the meaning of a reference, and a highly disfluent translation is difficult to understand and may therefore lead to low adequacy scores as well. There is also the matter of combining these two scores in a way that it can be used for other purposes, such as tuning automated metrics.

Nowadays, absolute evaluations of quality are collected mostly on a single, continuous 0-100 scale, in which the evaluator is asked to rate the overall quality of the translation, rather than rating separate attributes. This form of evaluation is known as Direct Assessment (DA) (Graham et al., 2013), and can be further divided into source-based and reference-based DA, depending on what sentence is shown to the evaluator as a base for judging the hypothesis.

Source-based DA, also known as bilingual DA, consists of asking how well the hypothesis expresses the meaning of the source sentence, and therefore the assessors should be bilingual speakers. Reference-based DA, on the other hand, is referred to as monolingual DA because it can employ monolingual speakers, by asking for judgements of the hypotheses based on how well they express the same meaning as a reference translation. However, using the reference may introduce reference bias into the evaluation, and therefore source-based DA is more recommended by the literature (Läubli et al., 2020) and tends to produce more reliable results (Bentivogli et al., 2018).

DA yields absolute scores in a continuous scale, which can then be standardized in order to smooth over differences in each assessor's scoring strategy; this cannot be done if interval-level scales are used. Additionally, statistical tests can be applied to continuous scores, to filter out contributions from participants deemed unreliable based on quality control methods (more on this in Section 4.6).

DA scores can also be converted to relative rankings (DARR), like those obtained by pairwise comparison and ranking methods, if there are scores for at least two translation hypotheses and the difference between these scores is significant enough to assume one hypothesis is of superior quality (Bojar et al., 2017). Naturally, the granularity of DA scores is lost in this conversion process, but the resulting pairwise rankings serve other purposes.

2.2.3 | Error Analysis

Error analysis is a family of evaluation methods designed to gather information about specific translation errors in each hypothesis. Errors can be annotated on different dimensions; they are usually categorized according to a hierarchical error typology, and also rated on their severity.

This method allows for quite a fine-grained analysis, though at the cost of considerable cognitive effort and time investment. Moreover, it should ideally be performed by professional assessors, having enough knowledge of the languages to be able to distinguishing certain linguistic phenomena in order to classify the errors. Assessors have to be provided with guidelines on what and how to annotate, usually based on a standardized framework.

In 2011, TAUS proposed the Dynamic Quality Framework (DQF), built on the assumption that translation quality depends on the type of content, the purpose of the content, and on its audience (A. Görög, 2014). It was created with the goal of standardizing MT error analysis, offering a way to categorize and count translation errors based on criteria commonly used in the industry. In 2014 it was unified[3] with another popular framework, Multidimensional Quality Metrics (MQM), and now DQF is a subset of MQM 2.0.[4]

MQM was originally developed at the DFKI for the EU-funded QTLaunchPad project (Burchardt, 2013), and it lists over 100 types of issues in a catalog[5] from which evaluation organizers can select the relevant issues to annotate for their tasks. The issues are divided in three main types: "style", "fluency", and "accuracy", and then weighted depending on their severity to calculate the final score, which ranges from $-\infty$ to 100. The score is calculated according to equation 2.1, where I_{minor}, I_{major} and $I_{crit.}$ are the numbers of errors classified as "minor", "major" and "critical".

$$MQM = 100 - \frac{I_{minor} + (5 * I_{major}) + (10 * I_{crit.})}{sentence\ length * 100} \tag{2.1}$$

Other evaluation campaigns might come up with their own frameworks for error analysis, or set out to investigate translation errors from different perspectives. An interesting example is the work of Popović (2021), which analyzed the nature and the causes of translation errors in order to identify what phenomena are dependant on specific domains or language pairs, and what can be seen as "universally challenging" for MT systems.

[3] `https://o.taus.net/academy/news/press-release/dqf-and-mqm-harmonized-to-create-an-industry-wide-quality-standard`, last accessed 05/08/2023.

[4] `https://themqm.org/faq`, last accessed 06/08/2023.

[5] `https://themqm.org/error-types-2/typology`, last accessed 06/08/2023.

MQM has been used as the standard evaluation method in WMT campaigns since 2021 (Freitag et al., 2021b, 2022). In addition to the final quality scores, it also provides word-level tags which may be used for various purposes, such as word-level QE (Bojar et al., 2015), which aims to attribute a binary OK or BAD label to each token in the hypothesis.

2.3 | Automatic Evaluation

The previous section covered the main families of methods for manual evaluation of MT quality; it should be clear by now that human judgements, although valuable, are considerably expensive and time-consuming to obtain. As MT research continued to evolve and more systems and techniques were being developed, evaluations were needed more and more often. Naturally, there was a lot of interest in creating methods that could measure translation quality automatically.

Banerjee and Lavie (2005) named a few basic criteria an automatic metric must satisfy in order to be useful and effective for MT evaluation: it must correlate well with human judgements of MT quality; it should be sensitive to differences in quality between systems; it should be consistent, producing similar scores for similar texts from the same system, and reliable, so that similarly scored systems can be considered of similar quality; and it should be general, usable for different tasks across a variety of domains.

It is immediately clear that these are not simple criteria to satisfy all at once by any single metric, given the inherent complexity of evaluating translation quality in the first place. In the following sections, we will describe three families of automatic evaluation methods, which have surfaced in a chronological manner as research evolved and more resources became available. We also emphasize their limitations and shortcomings, so that we may explore these limitations in upcoming chapters.

2.3.1 | Lexical Metrics

This category comprises the most widely used automatic methods for MT evaluation, also known as "string-based", "n-gram matching" or "lexical overlap" metrics. They require two elements: a reference translation, and a formula by which the overlap between the reference and the hypothesis will be calculated. The metric operates at the lexical level, measuring the overlap in terms of n-grams, words or sub-word units. It is usually a simple formula, easy to compute, thus making the metric fast and lightweight. Additionally, these metrics are relatively language-independent, requiring only that the segments can be tokenized into n-grams for the calculation.

These would be the main reasons why lexical metrics are so widely used, fulfilling the demand for constant evaluation in MT development.

One well-established metric is Word Error Rate (WER), which computes the edit distance (Eq. 2.2): the amount of edits (substitutions (S), insertions (I) and deletions (D)) necessary for a hypothesis translation to match a reference of length N. WER is still used nowadays for evaluating speech recognition systems, and in MT, it is also known as Translation Edit Rate (TER), with an adjustment where N can be the average length of the references in multi-reference scenarios (Snover et al., 2006b).

Snover et al. (2006a) proposed a variation called Human-targeted Translation Edit Rate (HTER), aiming for a human-in-the-loop approach by calculating the edit distance between an MT output and a post-edited version of it. HTER requires more effort, in that translators need to be hired and instructed to post-edit each hypothesis until it is fluent.

$$WER = \frac{S + I + D}{N} \tag{2.2}$$

The overlap between hypothesis and reference can also be measured based on two performance scores well known in Machine Learning: precision and recall. These scores consider the amount of overlapping tokens as the "true positives", which is then divided by the hypothesis' own length ($|h|$) in the case of precision (Eq. 2.3), to contabilize how many of the generated tokens are relevant (i.e. present in the reference); for recall, it's divided by the length of the reference ($|r|$), to measure how many of the reference's tokens were correctly generated (Eq. 2.4).

$$P = \frac{|overlapping\ tokens|}{|h|} \tag{2.3}$$

$$R = \frac{|overlapping\ tokens|}{|r|} \tag{2.4}$$

Precision and recall are widely used in Machine Learning with numerical data, but for NLP, they are rather limited: these formulas alone cannot account for word order in sentences, or repetition, for example. They have been modified in a variety of ways for measuring MT quality.

The current most used metric in the field, BLEU ("Bilingual Evaluation Understudy", Papineni et al. (2002)), is based on clipped (or modified) precision, by only counting an n-gram as correct as many times as it occurs in the references. The clipped precision score is calculated for n-grams of $N = \{1, 2, 3, 4\}$; the idea behind this is that the uni-gram matching should capture the adequacy of the translation, while the longer n-grams should

account for fluency. The final score will be their weighted geometric average (Eq. 2.5); this average is then multiplied by a Brevity Penalty (BP, Eq. 2.6), which serves to penalize translations that are too short compared to the closest reference length.

$$BLEU = BP * \sum_{n=1}^{N} w_n \ \log p_n \tag{2.5}$$

$$BP = \begin{cases} 1 & \text{if } c > r \\ \exp\left(1 - r/c\right) & \text{if } c \leq r \end{cases} \tag{2.6}$$

The acronym "BLEU" stands for **Bi**lingual **E**valuation **U**nderstudy, because it was originally proposed as "*an automated understudy to skilled human judges which substitutes for them when there is need for quick or frequent evaluations*" (Papineni et al., 2002, p. 1); nowadays, however, it is used extensively on its own and not accompanied by regular human evaluations. Furthermore, its design supports multiple references, to allow for variation in word choices, but in practice, due to scarcity of reference translations, it is most often used with only one.

The limitations of BLEU have been well known and discussed for years (e.g. Callison-Burch et al., 2006; Kocmi et al., 2021; Mathur et al., 2020a; Reiter, 2018). In the latest WMT metrics task, it was ranked **last** amongst the 13 reference-based metrics that were evaluated, based on low correlations with human judgements of translation quality (Freitag et al., 2022).

One major limitation is inherent to all lexical metrics, as they are based entirely on lexical matching between the hypotheses and references: the metric will enforce the wording of the references, failing to reward translations which might be of similar quality but worded differently (Popović, 2019). This is acknowledged in the BLEU paper, where they consistently recommend the use of multiple references from different translators, and if that is not possible, single references can be used for a test corpus if they were not all written by the same translator (Papineni et al., 2002).

Lexical metrics also fail to discriminate the severity of translation errors (Federico et al., 2014; Freitag et al., 2021a), since minor errors are treated the same as critical ones which can highly alter the meaning or the polarity of a sentence (Saadany and Orasan, 2021).

Most importantly, lexical metrics have been repeatedly shown to correlate poorly with human judgements (e.g. Edunov et al., 2020; Freitag et al., 2019; Mathur et al., 2020b). One of the first publications to point this out, Melamed et al. (2003), found that BLEU, in a single-reference evaluation of English–Arabic translations, had low correlations with human judgements of adequacy and fluency.

The shortcomings of BLEU have been a re-ocurring topic in all the latest WMT metrics tasks. The results in Barrault et al. (2019) indicated that BLEU failed to correlate with human DA scores at segment-level, and also failed to distinguish the highest performing MT systems. Freitag et al. (2020) pointed out a discrepancy between the systems that were deemed the best by automatic metrics and by humans in the WMT English–German tasks of 2018 and 2019 (Barrault et al., 2019; Bojar et al., 2018). In Freitag et al. (2022), BLEU was ranked last in both MQM and DA+SQM[6] human evaluation campaigns.

Mathur et al. (2020a) found cases in which systems with insignificant BLEU deltas were judged by humans to differ significantly in quality, and other cases in which system pairs that humans thought to be of similar quality had BLEU deltas as large as 3–5 points. Additionally, Callison-Burch et al. (2006) reported cases in which hypotheses assigned the same BLEU score were judged very differently in human evaluation; in practice, improvement in BLEU scores is insufficient to claim better translation quality. These analyses are particularly concerning because an increase higher than 1 or 2 BLEU points in comparison to a previous state-of-the-art system is widely regarded as a reliable evidence of improvement (Kocmi et al., 2021), and a sort of minimum improvement threshold for paper acceptance in the field (Mathur et al., 2020a). As Kocmi et al. (2021) argues, the use of BLEU might have even hampered progress in MT, as the metric can underestimate systems that showed improvements according to human judgements.

Other n-gram matching metrics have been proposed to try and mitigate some of the issues of BLEU, and although they are similarly limited, some of them have significantly outperformed BLEU in terms of correlations.

In the U.S., the National Institute of Standards and Technology (NIST) adopted a variation of BLEU, simply known as the NIST metric, which differs by weighting the n-grams by informativeness, based on how rare a particular n-gram is, whereas BLEU weighs all n-grams equally (Doddington, 2002). Zhang et al. (2004) found NIST to have more discriminative power than BLEU, although they say both metrics are flawed. NIST scores were used alongside BLEU during the OpenMT Evaluation series, which ran from 2001 to 2015[7], but the NIST metric has mostly fallen out of use since then (Marie et al., 2021).

chrF (Popović, 2015) is a metric that allows for more flexible matching by using character n-grams instead of words; unlike BLEU, it is tokenization-agnostic, and thus appropriate for more target languages. chrF is based on the F-score, which is the har-

[6]DA+SQM is a variation of DA where evaluators rate segments on a 0-100 sliding scale annotated with 7 labeled tick marks, ranging from "Nonsense / No meaning preserved" to "Perfect meaning and grammar" (Kocmi et al., 2022).

[7]`https://catalog.ldc.upenn.edu/LDC2010T10`, last accessed 23/09/2023.

monic mean of precision and recall. For the CHRF score (Eq. 2.7), the parameter β is used to weigh recall β times more important than precision. The standard value was $\beta = 1$ for equal weights in the original publication, but following Popović (2016), both the original[8] and the SacreBLEU implementation[9] use $\beta = 2$. CHRF was recommended by Kocmi et al. (2021) as the best lexical metric to use, if necessary—as other metrics we will discuss in upcoming sections require different resources—, and it continues to rank higher than BLEU in the latest WMT metrics tasks (Freitag et al., 2021b, 2022). There are also variations of CHRF, named CHRF+ and CHRF++, which include word uni-grams and bi-grams in the calculation respectively (Popović, 2017).

$$\text{CHRF} = (1 + \beta^2)\frac{P * R}{\beta^2 P + R} \tag{2.7}$$

The F-score is also the base for METEOR (Banerjee and Lavie, 2005), with recall weighted higher than precision; the authors believe this is important because recall is what measures how thoroughly the source sentence is covered by the translation, and they argue that the brevity penalty of BLEU does not make up for it. METEOR also aims for more generalized word-matching by allowing stems and synonyms besides exact matches. For this, however, it requires language-specific resources: a stemmer and a module for synonymy checks (e.g. WordNet for English.)

Ultimately, although BLEU is still by far the most used metric for evaluating MT, the latest meta-evaluations suggest that the field might be moving away from this type of metric: in last year's WMT metrics task, for the first time since 2008, there were no submissions of new metrics based purely on n-gram matching (Freitag et al., 2022).

2.3.2 | Embedding-Based Metrics

A new family of embedding-based metrics emerged in recent years. They aim to capture semantic similarity between words by using word embeddings, creating soft alignments between hypotheses and references and computing the similarity in embedding space (Bojar et al., 2017). The most prominent of these, scoring consistently higher than n-gram matching metrics in the latest years of WMT metrics tasks, have been YISI-1 (Lo, 2019) and BERTSCORE (Zhang et al., 2019). They both use contextual embeddings from pre-trained Language Models (LMs) such as RoBERTa (Liu et al., 2019) and XLM-RoBERTa (Conneau et al., 2019); YISI-1 measures similarity by aggregating the IDF-weighted distributional lexical semantic similarities between hypothesis and reference,

[8] `https://github.com/m-popovic/chrF`, last accessed 25/09/2023.
[9] `https://github.com/mjpost/sacrebleu#chrf--chrf`, last accessed 25/09/2023.

whereas BERTscore uses cosine similarity to generate an alignment matrix and returns precision, recall and F1 score.

While they may be performing better than lexical metrics, semantic similarity with a reference is hardly all there is to a good translation. Freitag et al. (2021b) found that embedding-based metrics did not *significantly* outperform simpler lexical metrics, and Freitag et al. (2022) reported, in results on the HWTSC Challenge Set (Chen et al., 2022), that these metrics performed poorly on relating synonyms and also seemed to generalize poorly, due to susceptibility to different text styles. We hypothesize that this might be why they have largely failed to find widespread acceptance in MT; for the added trouble of including additional resources such as WordNet or lists of synonyms for each language, it seems that the gains were not enough.

2.3.3 | Trainable Metrics

Lexical metrics are simple and cheap to compute, and for that they have been crucial in the progress of MT to where we are today. However, as succintly put by Koehn and Monz (2006), "*they are only an imperfect substitute for human assessment of translation quality*".

In the search for a better substitute, a new paradigm of methods has come up: trainable metrics, also known as learnable or neural metrics, are systems based on neural network models that explicitly optimize for correlation with human judgements (Mathur et al., 2019). Rather than attempting to compute lexical or semantic overlap between the input segments, these models are trained with data from human evaluations in order to predict different types of human judgements, based on long-established frameworks such as DA and MQM.

While they are still not widely used, trainable metrics have undeniably made an impact in the field of MT evaluation: they have been achieving the best results in terms of correlation with human judgements in large-scale meta-evaluation studies, consistently outperforming lexical and embedding-based metrics (e.g. Freitag et al., 2022; Kocmi et al., 2021).

Ruse, which was one of the first trainable metrics, is a multi-layer perceptron regressor based on sentence embeddings and trained on DA scores from WMT 2015–2017 metrics tasks (Shimanaka et al., 2018). Ruse introduced an approach where the input segments—the translation reference and hypothesis—are encoded separately and then combined through vector operations such as concatenation and element-wise product and difference.

It achieved high correlations in the 2018 task (Bojar et al., 2018), but it was not

submitted to the competition in subsequent years and has not gotten updates since 2019.[10]

ESIM (Chen et al., 2017) is a neural metric developed for Natural Language Inference and adapted for MT evaluation (Mathur et al., 2019); it uses a Bidirectional LSTM (Graves et al., 2013) to encode reference and hypothesis and then passes them through a feed-forward regressor trained on WMT 2016 DA scores. Prism (Thompson and Post, 2020) is a metric that treats MT evaluation as a zero-shot paraphrasing task, and scores outputs with a sequence-to-sequence model conditioned on reference translations. It is trained only on parallel data, not requiring prior human judgements. Similarly to RUSE, both ESIM and Prism performed well in the 2019–2021 WMT metrics tasks, but did not seem to gain much attention elsewhere and their development stalled since then.[11,12]

The first trainable evaluation system to get some traction was COMET (Rei et al., 2020a), created by researchers at Unbabel AI who believed that the existing automatic metrics failed to distinguish accurately between better and worse translation hypotheses, and consequently, between better and worse MT systems (Lavie, 2020). COMET is built on top of a cross-lingual encoder, such as XLM-RoBERTa (Conneau et al., 2019), which it uses to encode the source, reference and hypothesis segments into a shared feature space, and then passes them through a feed-forward regressor to predict human judgements of quality—it can be trained on DA scores, like most other trainable models we mentioned, but also on MQM or HTER. COMET introduced the novel approach of including the source segment as input alongside the reference; previously, this was only done in Quality Estimation, a reference-free evaluation strategy we will describe in the upcoming Section 2.3.4.

COMET models have been at the top of the rankings in the WMT 2020–22 metrics tasks (Freitag et al., 2021b, 2022; Mathur et al., 2020b), and it was the primary automatic metric recommended for use by Kocmi et al. (2021), provided that the languages are supported.

In the 2022 WMT metrics task, COMET-22 was followed by BLEURT and UNITE, two other trainable metrics that have been showing good correlations.[13] BLEURT (Pu et al., 2021; Sellam et al., 2020) encodes each hypothesis jointly with the corresponding reference and is then fine-tuned to produce a DA score. It was originally built on top of English-BERT, and thus had limited applicability, but by now it has been tested on 14 (mostly high-resource) languages and should theoretically work for the 100+ languages

[10] `https://github.com/Shi-ma/RUSE`, last accessed 23/09/2023.

[11] `https://github.com/nitikam/mteval-in-context`, last accessed 23/09/2023.

[12] `https://github.com/thompsonb/prism`, last accessed 23/09/2023.

[13] COMET, BLEURT and UNITE were outperformed only by the proprietary METRICX XXL, a massive multi-task metric fine-tuned on a 30B mT5 large LM (Freitag et al., 2022).

in the training data of its new RemBERT (Chung et al., 2020) model.[14] UniTE (Wan et al., 2022)[15] is a new architecture proposed by Wan et al. (2022) for Comet, aiming to unify the functionalities of three MT evaluation scenarios: src (source-only, as in Quality Estimation), ref (reference-only, as is the case of Ruse, ESIM, Prism and BLEURT), and src-ref (like Comet, except for Comet-QE). UniTE encodes the input sentences in a joint manner to obtain contextualized representations, while Comet encodes the inputs separately and then combines them (Rei et al., 2023b).

Comet is currently used at Unbabel as the primary evaluation method for their MT systems, and it has also started being used to report results in MT publications, often alongside other metrics like BLEU, and sometimes chrF, TER and/or BLEURT (e.g. Agarwal et al., 2023; Macken et al., 2023; Zeng et al., 2023; Zhao et al., 2023). At the moment, it seems that Comet is the neural metric with the highest potential to achieve widespread use; new versions are in development and it is constantly receiving improvements, often based on findings from meta-evaluation studies.

Despite the undeniable superior performance of neural metrics, there are a few reasons why they have yet to achieve widespread use. At the root of most criticisms, there is the inherent complexity of using neural networks, which are typically considered "black box" systems, for evaluation purposes (Rei et al., 2023b). As neural metrics continue to evolve and find wider acceptance MT research, there will be growing demand for work on explainability, to substantiate the reliability of the metric and make it easier for users to interpret the scores that they see (Leiter et al., 2022).

Moreover, the use of neural networks for evaluation naturally raises the question of how much their performance is modulated by the data on which they were trained, and whether they are acquiring biases from this data, which is a complex issue to diagnose. This has also not been the subject of systematic analysis. Kocmi et al. (2021) observed that, despite being fine-tuned on the WMT news domain, the original Comet model did not appear to have overfitted to it, showing equally good results on other domains; similarly, Freitag et al. (2022) reported that Comet, BLEURT and UniTE were robust to different domains, exhibiting superior performance in test sets in the domains of news, social, e-commerce and conversational chat.

Critically, an issue that is at the forefront of this dissertation is language support — unlike lexical metrics, which are mostly language-independent, trainable metrics, as the name implies, have to be trained in order to learn to predict quality scores, and thus they are somehow restricted to a set of languages. Comet and Bleurt, for example, are built by default on top of cross-lingual encoders that support 100 languages—mostly

[14] `https://github.com/google-research/bleurt#language-coverage`, last accessed 03/10/2023.
[15] `https://github.com/NLP2CT/UniTE`, last accessed 30/08/2023.

high-resource languages, but also a few under-resourced ones. Section 3.1 will go into further detail on this issue in the scope of COMET.

2.3.4 | Quality Estimation (QE)

Quality Estimation (QE) is defined as the task of "*predicting the quality of a system's output for a given input, without any information about the expected output*" (Specia et al., 2009, p. 28). It is a concept rather than a type of metric, but it has inspired the development of metrics known as "referenceless" or "reference-free", because they do not look at reference translations to estimate translation quality, instead using only the source segment.[16]

These metrics were not created necessarily to replace reference-based ones, but rather, to fill in the gap where reference translations are unavailable or insufficiently good (Chatzikoumi, 2019). Nevertheless, many of the existing reference-based metrics now have QE alternatives.

The COMET framework has two types of reference-free models, COMET-QE and COMETKIWI (more on this in Section 2.4.2). OPENKIWI (Kepler et al., 2019)[17], incorporated into COMETKIWI, can also be used on its own as an open-source framework for QE which implements four different QE systems, ranked the highest in WMT 2015–18 metrics tasks: QUETCH (Kreutzer et al., 2015), NUQE (Martins et al., 2016, 2017), Stacked Ensemble (Martins et al., 2016, 2017), and Predictor–Estimator (Kim et al., 2017; Wang et al., 2018). Most recently, Unbabel also released COMETKIWI XL and XXL (Rei et al., 2023a), their first QE models based on Large Language Models (LLMs).

UNITE, designed to cover all three evaluation scenarios (source-only, reference-only and source-reference-combined), can naturally function as a reference-free metric (UNITE-SRC). Prism (Thompson and Post, 2020) can be applied as a reference-free metric as well, by scoring MT outputs conditioned only on source sentences instead of reference translations (Thompson and Post, 2020). REUSE is an embedding-based, purely reference-free metric which computes chunk- and sentence-level similarity by leveraging BERT embeddings (Mukherjee and Shrivastava, 2022).

[16] As mentioned in Section 2.3.3 where we talked about UNITE, this can also be referred to as *source-only* evaluation.

[17] `https://github.com/Unbabel/OpenKiwi`, last accessed 30/08/2023.

2.4 | COMET

COMET (**C**rosslingual **O**ptimized **M**etric for **E**valuation of **T**ranslation) was created in 2020 by Unbabel, a multilingual translation services company, to evaluate the quality of their own specialized MT systems.[18] It was released as an open-source framework for MT evaluation models and a suite of pre-trained models that can evaluate translations to and from 100+ languages, by predicting human scores of translation quality.

COMET-based models have lately been considered the most promising neural metric for MT; they have scored impressive correlation scores in WMT metrics tasks since the 2020 edition (Freitag et al., 2021b, 2022; Mathur et al., 2020b), and were recommended by Kocmi et al. (2021) as the primary metric to use when possible. Other specialized meta-evaluations have also found that COMET correlated better with human judgements than other lexical and neural metrics (e.g. Macken et al., 2023; Sai et al., 2023). These findings are the main motivation behind our choice to focus on COMET in this study; we believe that the outstanding results that COMET metrics have shown are slowly changing the status quo in MT evaluation. Furthermore, we exploit the flexibility of the COMET framework in our experiments, as we are focusing on two language pairs that are not supported in any pre-trained models.

First, this section will provide more detail on the COMET framework structure and its capabilities.

2.4.1 | Framework Structure

"COMET" itself is not a single model or metric, but rather a framework for neural models that function as evaluation metrics. The framework is fully open-source and is currently in version 2.0. It includes several pre-trained models ready to be used out-of-the-box for scoring MT hypotheses, as well as the possibility to train custom COMET models. All of the pre-trained models were trained on publicly available data from WMT metrics tasks.

The base structure that is shared by all the models consists of two parts: a cross-lingual encoder and a pooling layer (Rei et al., 2020a).

The cross-lingual encoder is a pre-trained masked language model, which is used to map all of the inputs—the source (s), reference (r) and hypothesis (h) segments—into a shared, multilingual feature space. This encoder generates an embedding $e_j^{(l)}$ for each token x_j and each layer $l \in \{0, 1, ..., k\}$. The default encoder architecture used in pre-trained COMET models is XLM-RoBERTa[19], but COMET configuration also supports

[18] `https://unbabel.com/research/comet`, last accessed 01/10/2023.

[19] COMET-20 uses `xlm-roberta-base`, and COMET-22 uses `xlm-roberta-large`.

any pre-trained encoder from HuggingFace Hub which is compatible with XLM-R, BERT (Devlin et al., 2018), RemBERT (Chung et al., 2020), or MiniLM (Wang et al., 2020).

The pooling layer employs layer-wise attention to pool information from the most important encoder layers into a single embedding. Rather than using only the last encoder layer, which has been shown to potentially generate worse results (Zhang et al., 2019), this pooling method is motivated by findings that intermediate layers capture different linguistic information that might be important for different tasks (Peters et al., 2018; Tenney et al., 2019). The network also employs layer dropout in order to avoid overfitting, and finally, the word embeddings are average-pooled to generate a sentence embedding for each of the segments.

2.4.2 | Architectures

The structure described in the previous section is the common base of all COMET models, but the framework provides a few distinct model architectures, which vary in aspects such as the type of training data and whether they are reference-dependent. Models of all architectures can be used in the same way for inference, and take the same input data— except that referenceless models do not use references—, but they differ in how the training works, as they leverage different types of data for different purposes.

These are the four architectures available currently, also described as diagrams in Figure 2.1:

- **Estimator** (or **Regression**) models are trained on absolute scores of human judgements, namely DA and MQM.[20] The inputs s, h and r are combined into a single vector $x = [\, h; r; h \odot s; h \odot r; |h - s|; |h - r| \,]$. The combinations of embeddings serve to highlight the differences between them in the shared feature space, and furthermore, as this space is multilingual, the source segment is included in combinations so that the model may be able to identify errors that cannot be detected only by comparing hypothesis and reference (Amrhein et al., 2022). Nonetheless, s is not included in raw form (like h and r are), due to concerns that the feature space between different languages might be poorly aligned (Zhao et al., 2020). With the features encoded and combined, the model is then trained to regress directly on the quality scores by minimizing the mean squared error (MSE).

  This is the main architecture in the COMET framework, used by the default pre-trained models, initially COMET-20 and now COMET-22, which are both DA Esti-

[20] The original COMET paper also featured experiments with a COMET-HTER model, which was later discontinued because its correlation with human judgements was inferior to COMET-DA and COMET-MQM (`https://github.com/Unbabel/COMET/issues/28`, last accessed 26/09/2023).

mators. Although MQM has been gaining more traction in recent years of WMT metrics tasks, it requires a lot more effort and so MQM datasets are scarce; using DA allows for larger and more diverse training datasets.

- **Translation Ranking** models are trained on pairwise rankings of hypotheses, where h^+ is a "better" hypothesis and h^- is a "worse" one. These rankings can be obtained in the form of DARR, for example. Inputs are embedded and then combined into a vector $x = [\ s; h^+; h^-; r\ ]$, and the model is trained to minimize the Triplet Margin Loss (Schroff et al., 2015), which encourages it to encode better translations closer to the anchors (source and reference) and to push worse translations away.

- **Referenceless** models are the reference-free versions of COMET, based purely on Quality Estimation (QE). They are trained to predict absolute scores (DA or MQM) by looking only at the source and hypothesis. There are currently two sub-types of referenceless models:

  - COMET-QE (Rei et al., 2021a) was the first QE architecture, which works similarly to Estimator models, except that the combination of inputs is different, as references are not used: $x = [\ h; s; h - s; h * s\ ]$. Note that, unlike in the Estimator architecture, here the source is also passed raw alongside the combined features. COMET-QE (`wmt20-comet-qe-da`) was the default referenceless model in COMET 1.0.
  - COMETKIWI (Rei et al., 2022c) is a newer QE model based on OpenKiwi (Kepler et al., 2019), an open-source framework for QE. This model combines the predictor-estimator architecture of OPENKIWI with COMET's training style, and leverages word-level OK/BAD annotations. It is currently the default referenceless model (`wmt22-cometkiwi-da`), as of the COMET 2.0 release.

- The **Unified** architecture, also known as the UNITE model, was proposed by Wan et al. (2022) and incorporated into COMET 2.0. It employs novel strategies to combine into a single model the functionalities of source-only, reference-only and source-reference-combined MT evaluation. Although there are no pre-trained COMET models with this architecture as of yet, UNITE is available on its own GitHub repository[21] and the Unified architecture is available in the COMET framework for custom models.

[21] `https://github.com/NLP2CT/UniTE`

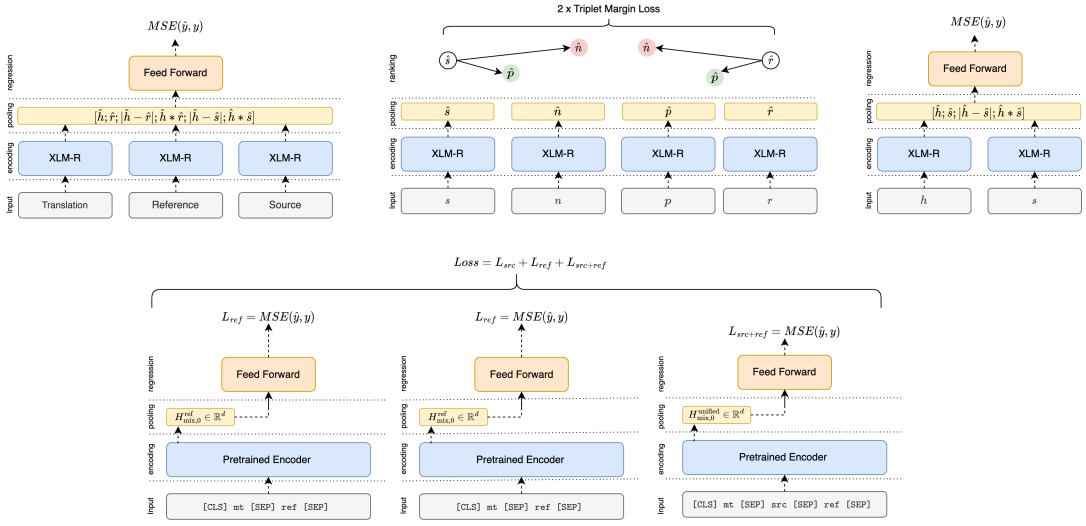

Figure 2.1: Diagrams of the four currently available COMET architectures: Estimator (top left), Translation Ranking (top middle), Referenceless (top right), and Unified Metric (bottom). (`https://unbabel.github.io/COMET/html/models.html`, last accessed 10/10/2023.)

In addition to these, the framework also features COMETINHO models, which are lightweight versions of COMET: they have the same architecture as Estimator models, available for both DA and MQM, but built on top of a smaller cross-lingual encoder, MiniLMv2 (Wang et al., 2020). COMETINHO was created in an attempt to bridge the gap between COMET, a neural metric, and string-based metrics which are very light-weight. Rei et al. (2021a) reported COMETINHO to be 19x faster than the original COMET model, whilst maintaining the level of correlation with human judgements. In another, more recent study, the authors employed knowledge distillation and pruning techniques to generate new lightweight versions of COMET (Rei et al., 2022b).

It is worth noting that, as of the current stable version of the framework (2.0), COMET has no support for multiple references. Rei et al. (2020c) introduced an experimental approach for handling one alternative reference at inference time, but the results showed little to no improvement with the inclusion of alternative references, so the authors speculated that quality might be more important than quantity for COMET. Unlike the case of lexical metrics like BLEU, where using multiple references is recommended, as it creates a larger pool of wording variations, approaches like COMET are based on meaning-representation of the segments, and thus it might be better to have only one high-quality reference. Creators have suggested that, in case the user wishes to leverage multiple available references, scoring the hypotheses separately with each set of references and then

averaging them out would be best.[22]

All of the pre-trained COMET models are available for use with the `unbabel-comet` pip package[23], directly through the command line interface or in Python.

2.4.3 | Score Interpretation

For a given test set of N samples, each sample consisting of source, hypothesis, and reference (unless using a referenceless model), any COMET model will return N segment-level scores, and their average (as a "system score").

When using COMET 1.0 models, each score is an unbound rational number, meaning it can be positive or negative; the higher, the better. These scores are not very interpretable on their own, but can be used to compare translations and MT systems.

As of COMET 2.0, quality scores are in the range of [0,1], to increase their interpretability: the translations with scores closest to 1.0 should be the "best" translations, and translations scored close to 0.0 should be considered no better than random chance.[24] In order to obtain scores in this range, the model has to be trained with data in the same range, which is done by applying feature scaling to the standardized z-scores, a procedure we will describe in more detail in Section 3.3.

In any case, the range of quality scores that COMET models produce is highly variable across language pairs, because there are a lot of factors which influence the human judgements used as training and test data: the complexity of translating between these languages, the data domain(s), the assessors who provided the assessments, among other factors. Therefore, COMET scores are more applicable for comparing systems or segment translations within the same language pair.[25]

2.5 | Summary

This chapter has been a thorough overview of the theoretical background of this dissertation, which is centered mainly on MT meta-evaluation. We have started by describing the main existing types of methods for human evaluation of MT quality, and afterwards, we have taken a look at three categories of automatic methods: lexical or n-gram matching metrics, embedding-based metrics, and trainable metrics. As our focus will be henceforth

[22] `https://github.com/Unbabel/COMET/issues/20`, last accessed 26/09/2023.

[23] `https://pypi.org/project/unbabel-comet`

[24] `https://unbabel.github.io/COMET/html/faqs.html#interpreting-scores`, last accessed 07/09/2023.

[25] `https://github.com/Unbabel/COMET/issues/110`, last accessed 07/09/2023.

placed on the COMET framework, we have also gone into deeper detail on its overall structure and available architectures for the design of evaluation models.

All of these concepts will be essential for understanding our motivations and the decisions we have taken while designing and executing our own experiments, the first of which will be described in the next chapter.

3

Experiments with COMET

As we discussed in the previous chapter, strictly string-based metrics are so widely used in MT evaluation because they are easy and cheap to compute, as they only compare the surface forms of the segments. This is also why their results might correlate poorly with human judgements; naturally, *some* knowledge of the source and target languages is necessary for any method to be able to discriminate translation quality on a deeper level, and as a consequence, more computational power is needed to access this knowledge.

Embedding-based metrics (see Section 2.3.2) make use of linguistic resources, such as lexical databases or language models, in order to in order to measure semantic similarity—they create a soft alignment between reference and hypothesis, in an attempt to capture distinct wording with equal meaning, which should make them more flexible than strictly lexical metrics. Consequently, they can only be used to evaluate languages for which such resources are available, and still, they remain fully based on comparison with the reference segment, albeit a fuzzy comparison.

Trainable (or neural) metrics, which aim to go even further and learn to predict quality scores directly based on human judgements, require cross-lingual encoders and annotated parallel data, which leaves them with a critical limitation: they can only provide reliable results for languages they have been trained on.

Neural metrics have been topping the rankings in the latest WMT metrics tasks, the competition that is the biggest reference in the field of MT meta-evaluation. Moreover, according to the findings of these competitions and from other meta-evaluation studies cited in the previous chapter, many researchers in the field believe that COMET has a lot of potential as a neural metric that has been proven to be more reliable than regular string-based metrics, such as BLEU, in multiple scenarios of MT evaluation. Therefore, it is in our best interest to explore these language-related limitations.

We will begin in Section 3.1 by discussing the ways in which COMET's language

support is limited, based on its architecture, and what that might mean for the evaluation of languages that are not covered. In Section 3.3, we present the preliminary experiments we carried out with COMET, exploring its potential to evaluate language pairs outside of its training data. We did this first because there are publicly available datasets of human quality scores for language pairs that fit this scenario, which is not the case for the under-resourced pairs that are our main focus, English–Maltese and Spanish–Basque. Our results and conclusions are discussed in Section 3.5, as well as the ways in which these conclusions will guide our decisions in designing our upcoming experiments.

3.1 | Language Support in COMET

COMET is a neural evaluation framework with a base structure composed of two parts: a pre-trained cross-lingual encoder, and a feed-forward regression model (see Section 2.4.1). This architecture leads to two ways in which language support is somehow limited, also shown in the diagram in Figure 3.1:

1. Cross-lingual encoders are trained with monolingual data in N languages, and therefore can only reliably encode segments in these languages. Its training is outside of COMET's scope; the encoder is simply plugged in from HuggingFace Hub. By default, most COMET models have been using XLM-RoBERTa, which supports $N = 100$ languages (Conneau et al., 2019), but it can be replaced by any encoder of compatible architecture when training a custom COMET model. We will refer to these N languages as the languages that a COMET model *supports*, and any languages outside this set are therefore *unsupported languages*.

2. The feed-forward regressor is trained on parallel data for M language pairs, annotated with quality scores. For example, COMET-20[1] was trained on WMT 2017–2019 DA scores covering 24 language pairs (Rei et al., 2020c), and COMET-22 additionally included the WMT 2020 data, totaling 32 language pairs. The encoder will first map this data into a shared feature space, and thus the languages in these M pairs should ideally be—and historically have been—a subset of the N languages supported by the encoder. They will be henceforth referred to as the language pairs that a COMET model has *seen*, and thus any pairs outside of this set are *unseen language pairs*.

The first restriction, referring to the encoder's language support, appears to be much stronger than the second one. Encoding the inputs is the first step, so if a language

[1] `https://huggingface.co/Unbabel/wmt20-comet-da`

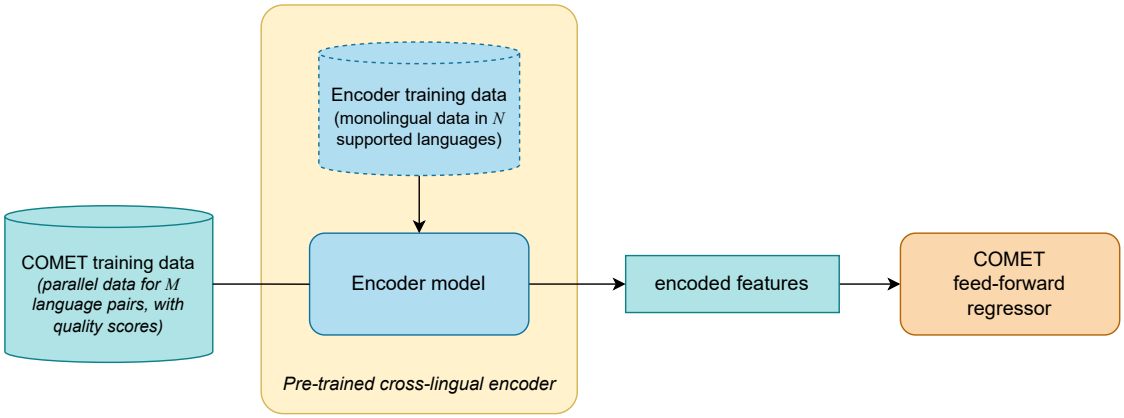

Figure 3.1: Diagram of the base structure of COMET models, showcasing the two sources of language restrictions: the monolingual training data of its cross-lingual encoder, and the training data of COMET itself.

is unsupported, and thus the quality of the embeddings cannot be trusted, then the performance of the COMET model will also be questionable.

COMET documentation warns upfront that the results are unreliable for unsupported languages[2], and Ricardo Rei, the main developer of COMET, stated that the models should work well for all languages supported by XLM-R, in theory, but otherwise its results might be unreliable, especially if the *target* language is unsupported.[3,4] There are exceptions, such as the case of English↔Inuktitut in the WMT 2020 metrics task (Mathur et al., 2020b), for which COMET models obtained Pearson correlation scores in the range of 0.6–0.8 despite Inuktitut not being supported by XLM-R nor included in the models' training data. On the other hand, Tom Kocmi from Microsoft reported unstable correlations in a number of language pairs featuring unsupported languages[5]; in the end, the published study (Kocmi et al., 2021) only covered supported languages. Therefore, the extent of this limitation in encoder support has never been systematically analyzed.

As for the second restriction, relating to the language pairs present in COMET's training data, it is an aspect we emphasize on because it is relevant to our study and will be explored further in our experiments, but from what we have seen in published works about COMET and in other meta-evaluations, it has never been pointed out explicitly as an actual "limitation" of the models. One could say that, once the inputs are encoded into

[2] `https://github.com/Unbabel/COMET#languages-covered`, last accessed 22/09/2023.

[3] `https://github.com/Unbabel/COMET/issues/38`, last accessed 11/10/2023.

[4] We understand that this distinction is probably due to the target and reference segment being more important for evaluation, and their embeddings playing a bigger role in the combined vector that is passed to the feed-forward regressor, since the source embedding is not included in raw form (see Section 2.4.2).

[5] `https://github.com/Unbabel/COMET/issues/18`, last accessed 11/10/2023.

a shared, multilingual feature space, the training process uses all embeddings the same way for the model to learn the *task* of predicting human quality scores; however, we know that is not *exactly* the case, as languages are not equally represented in the encoder, and thus embedding quality can vary (Conneau et al., 2019; Wu and Dredze, 2020).

Of course, the training data of any COMET model understandably represents only a small portion of all the possible language pairs that can be made from combining N languages supported by XLM-R. Similarly, published test results tend to cover only a number of language pairs. Nonetheless, as the framework documentation and publications do not make note of this restriction, we assume that the models can be expected to produce reliable scores, and this assumption is what we will investigate with the experiments described in this chapter.

For the overarching purposes of this dissertation, we should note that Maltese and Basque are unsupported by all currently available COMET models, and English–Maltese and Spanish–Basque are unseen language pairs, i.e. not present in any of the training datasets used. Therefore, there is essentially no basis for COMET to perform reliably well for these language pairs, or for any others involving Maltese or Basque. We will come back to how these limitations impact the evaluation of translations involving Maltese and Basque specifically in Chapter 5.

3.2 | Motivation

Before we approach the specific cases of English–Maltese and Spanish–Basque translation, we looked at another scenario that was more readily available for analysis: other language pairs that consist of languages that COMET supports, but that were not in its training data (unseen language pairs). It is important to emphasize that all languages covered in the experiments in this chapter *are* supported by XLM-R.

We decided to perform this analysis because COMET models are most often evaluated on the same language pairs they were trained on, which is the typical scenario of WMT metrics shared tasks. In fact, participants of the shared task have the opportunity to fine-tune their metrics specifically to perform well on the languages that will be evaluated, as was the case of the COMET ensemble model submitted to WMT 2022, which was fine-tuned on MQM data for Chinese–English, English–German and English–Russian and then tested on these language pairs (Rei et al., 2022a).

Differently from this scenario of WMT metrics tasks, a large-scale meta-evaluation study conducted at Microsoft covered 232 language pairs made up only of languages supported by COMET (Kocmi et al., 2021). They obtained 2.3M sentence-level human

judgements for 4,380 MT systems, and recommended COMET[6] as the primary metric to use for supported languages.

Our goal here is to address two main questions:

1. How well does COMET perform for other encoder-supported languages that were not seen in its training data?

2. How much can we improve this performance for each unseen language pair, in terms of correlation with human judgements, by fine-tuning the model on more Direct Assessment (DA) scores?

Recently, Sai et al. (2023) described a similar experiment, focused on 5 Indic languages: Tamil, Gujarati, Hindi, Marathi, and Malayalam. They evaluated translations of FLORES-101[7] sentences from English into these 5 languages, obtained from 7 MT systems, and found that COMET-MQM and COMET-DA[8] obtained the best correlations, in comparison with metrics like BLEU, BLEURT-20 and CHRF, among others. Furthermore, they used the same data to fine-tune COMET-MQM; this fine-tuned model outperformed the COMET baselines, not only when trained on the 5 selected languages at once, but also in a zero-shot setting, using 4 languages in training and evaluating on a separate one. This study further motivates us to explore such possibilities of improving COMET models for unseen language pairs.

3.3 | The Experiment

At the time of submission of this dissertation, the latest stable pre-trained COMET model for general purpose MT evaluation was COMET-22 (Rei et al., 2022a).[9] COMET-22 has been the default model since the release of COMET 2.0, the second open-source version of the framework, after its original release in 2020.[10] COMET-22 is a DA Estimator model

[6]The model evaluated in this study was `wmt20-comet-da` (`https://huggingface.co/Unbabel/wmt20-comet-da`).

[7]`https://github.com/facebookresearch/flores`

[8]The pre-trained COMET models they used were `wmt20-comet-da` and `wmt21-comet-mqm` (`https://unbabel.github.io/COMET/html/models.html`).

[9]To clear up any naming confusions, what we are calling "COMET-22" here is the DA Estimator model described in Rei et al. (2022a), released as `wmt22-comet-da` (`https://huggingface.co/Unbabel/wmt22-comet-da`). In their paper, "COMET-22" refers to an ensemble model composed of this DA Estimator and a new multitask model trained on MQM, but ultimately the ensemble was not released, as it performed poorly on language pairs not included in the competition (`https://github.com/Unbabel/COMET/issues/163`, last accessed 13/09/2023).

[10]`https://unbabel.com/introducing-unbabel-comet-v2-0-improved-models-and-metrics-for-better-machine-translation-evaluation`, last accessed 10/10/2023.

(see Section 2.4.2), built on top of XLM-R and trained on DA scores from 2017–2020 WMT metrics tasks. All the language pairs included in this dataset and the counts of segments for each of them can be seen in Table 3.1.

Language pair	**Size**	**Language pair**	**Size**	**Language pair**	**Size**
Chinese–English	126,947	Estonian–English	20,496	English–Tamil	7,890
English–German	112,420	German–Czech	13,804	Tamil–English	7,577
German–English	99,183	English–Estonian	13,376	English–Gujarati	6,924
English–Chinese	81,805	Polish–English	11,816	Kazakh–English	6,789
Russian–English	70,280	English–Polish	10,572	German–French	6,691
English–Russian	62,749	Lithuanian–English	10,315	English–Latvian	5,810
English–Czech	60,937	English–Japanese	9,578	English–Turkish	5,171
Finnish–English	46,145	Gujarati–English	9,063	Khmer–English	4,722
English–Finnish	34,335	English–Lithuanian	8,959	Pushto–English	4,611
Turkish–English	30,186	Japanese–English	8,939	French–German	3,999
Czech–English	27,847	English–Kazakh	8,219		

Table 3.1: Size of the COMET-22 training set per language pair. (Table generated from the training data files.)

We collected the data released by WMT for the 2021 and 2022 campaigns (Freitag et al., 2021b, 2022)[11], which was *not* used to train COMET-22, and looked at the newly included language pairs, to find those that were not seen by COMET-22 in its training. Out of 9 new language pairs, we selected the ones with at least 10,000 tuples available, which were the top 4: Ukrainian–English (`uk-en`, 16,470 tuples), Hausa–English (`ha-en`, 13,171 tuples), English–Icelandic (`en-is`, 10,838 tuples), and English–Hausa (`en-ha`, 10,812 tuples).

All languages involved (Icelandic, English, Hausa and Ukrainian) are supported by XLM-RoBERTa. Additionally, we verified that the four language pairs we selected are either `X`→`en` or `en`→`X`, where `X` is not seen in the training data of COMET-22 even as part of other language pairs.

The DA datasets provided by WMT consist of two scores for each tuple: the raw score, between 0 and 100, and the z-score, obtained by standardizing the raw scores according to each evaluator's mean score and the standard deviation (Bojar et al., 2016b). In addition to that, as explained in Section 2.4, COMET-22 introduced a new training approach where DA scores are scaled to the range [0,1] before training, so that the model would also mostly produce scores in this range, and anything outside of it is clipped to fit the

[11] `https://unbabel.github.io/COMET/html/faqs.html`, last accessed 08/10/2023.

interval. We rescaled the z-scores in the DA data using feature scaling, similarly to the process described by COMET creators[12], which originally consists of the following 4 steps:

1. Find a reasonable mininum value (r_{min}): the average z-score of all segments with more than 1 annotator where all annotators agreed that the score was 0.
2. Find a reasonable maximum value (r_{max}): the average z-score of all segments with more than 1 annotator where all annotators agreed that the score was 100.
3. Apply a `MinMaxScaler`[13] to the data using r_{min} and r_{max} as the feature range.
4. Clip (truncate) the data to the range of [0,1].

We had to make a change to this process in steps 1 and 2 because, for the 4 language pairs we covered in this experiment, there were no items annotated by more than 1 annotator; therefore, we calculated the minimum and maximum values (r_{min} and r_{max}) out of all segments scored 0 and 100.

Once the data was rescaled, we set aside random sets of 200 samples for validation and 200 for testing, and then defined three training sets: a small one with 400 samples, a medium one with 4,000 samples, and a large one with 10,000 samples. We set up the training configuration and the datasets so that the process could be executed incrementally: the small model was trained on top of COMET-22 with the first 400 samples, the medium model was trained on top of the small model's checkpoint with 3,600 more samples, and finally, the large model was trained on top of the medium checkpoint with the 6,000 remaining tuples. We did it this way to save resources in training, and also to ensure that each subsequent model would only be different to the previous one by its additional data.

The hyperparameters we used are described in Table 3.2; we followed the values recommended in the COMET 2.0 repository, for comparability.

3.4 | Meta-Evaluation Methods

In order to evaluate our custom models in comparison to the base COMET-22 model, we used two correlation coefficients that have been typically used in previous MT meta-evaluation works, namely Kendall's Tau and Spearman rank correlations. They measure the ordinal association between the scores given by the metrics and by humans (Kendall,

[12] `https://github.com/Unbabel/COMET/issues/131`, last accessed 05/10/2023.

[13] `https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MinMaxScaler.html`

Hyperparameter	**Value**
Encoder model	XLM-RoBERTa (large)
Optimizer	AdamW
Num. frozen epochs	0.3
Learning rate	1.5e-5
Batch size	16
Loss function	MSE
FP precision	16
Feed-forward hidden units	(3072, 1024)
Feed-forward activations	Tanh
Feed-forward dropout rate	0.1
Min. epochs	1
Max. epochs	4

Table 3.2: Hyperparameters used to fine-tune COMET-22 on unseen language pairs.

1970), where the human scores are the DA scores, and the metric scores are the outputs from pre-trained and custom COMET models.

Kendall's Tau (Kendall, 1938) is calculated based on the number of "concordant pairs" (n_c), for which the metric points to the same ordering as the humans, and "discordant" pairs (n_c), for which the orderings are different. There may also be tied pairs (t_h and t_m), which are not concordant nor discordant, and are thus handled differently in Kendall's Tau-b (Kendall, 1945). The Tau-b variant (τ_B, Eq. 3.1) was used in the latest WMT edition (Freitag et al., 2022) and is the default variant implemented in `scipy`[14], so it is the one we will be using. Kendall's Tau-b is also the default metric used to monitor the performance of COMET models during training.

$$\tau_B\left(h, m\right) = \frac{n_c - n_d}{\sqrt{(n_c + n_d + t_h)(n_c + n_d + t_m)}} \tag{3.1}$$

The Spearman rank correlation coefficient (Dodge, 2008), denoted ρ, is defined as the Pearson correlation (Freedman et al., 2007) between the rankings ($R(h)$ and $R(m)$) of the raw scores (h_i and m_i), as per Equation 3.2.

$$\rho\left(h, m\right) = \frac{cov(R(h), R(m))}{\sigma_{R(h),R(m)}} \tag{3.2}$$

Spearman correlation is often used alongside Kendall for MT meta-evaluation (e.g. Kocmi et al., 2021). These two correlation coefficients can be considered mostly inter-

[14] `https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.kendalltau.html`

changeable, and tend to produce close values (Crichton, 2001). We decided to calculate and report both, in line with what we have seen in the recent literature.

3.5 | Results & Discussion

We evaluated our 12 fine-tuned models on test sets of 200 samples for each language pair, and compared the results to the base COMET-22 model. Table 3.3 shows the Kendall's Tau and Spearman correlation scores between human judgements and the scores generated by all these models.

	uk-en		en-is		ha-en		en-ha	
Model	τ	ρ	τ	ρ	τ	ρ	τ	ρ
Base (COMET-22)	0.017	0.025	**0.423**	**0.589**	0.110	0.159	0.145	0.190
Small	0.025	0.038	0.426	0.591	0.106	0.152	0.112	0.147
Medium	0.087	0.123	0.476	0.657	0.111	0.156	0.194	0.255
Large	0.099	0.139	**0.488**	**0.673**	0.082	0.114	0.206	0.270
Improvement	**0.082**	**0.114**	0.065	0.084	-0.028	-0.046	0.062	0.081

Table 3.3: Kendall's Tau (τ) and Spearman (ρ) correlation scores for the base and fine-tuned models on our test set. (The scores in red were deemed statistically insignificant, with p-values < 0.05.)

	zh-en		en-de		en-ru	
Model	τ	ρ	τ	ρ	τ	ρ
COMET-20	0.336	0.463	0.206	0.270	0.256	0.330
COMET-21	**0.377**	**0.513**	**0.237**	**0.309**	0.263	0.345
COMET-22	0.362	0.495	0.221	0.289	**0.285**	**0.369**

Table 3.4: Correlation scores for 3 COMET DA Estimator models, tested on the 2021 WMT metrics task MQM annotations. Partially reproduced from Rei et al. (2022a). COMET-{20,21,22} are `{wmt20,wmt21,wmt22}-comet-da`, respectively (Rei et al., 2020b, 2021b, 2022a).

The first thing to notice is that most of the correlation scores for COMET-22 out-of-the-box are quite low, with the notable exception of English–Icelandic, for which the model's results correlate surprisingly well with human judgements, with Kendall's Tau and Spearman correlation values of 0.423 and 0.589, respectively. For comparison, and to exemplify the scale of correlation scores that COMET models can reach, Table 3.4

shows the same correlation measures for 3 DA Estimator models (2020–2022 versions) on 3 high-resource language pairs from the 2022 WMT test set: Chinese–English (`zh-en`), English–German (`en-de`) and English–Russian (`en-ru`).

The impressive performance of COMET-22 on our English–Icelandic test set could be attributed to these two languages sharing a common ancestor, as they are both in the Germanic family; albeit distant relatives, they are closer to each other than English is to Hausa, an Afro–Asiatic language, or to Ukrainian, part of the Indo–European family (Hammarström et al., 2023).

Overall, Ukrainian–English has the worst scores, which might be due to the quality of the data: while the segments for the other three pairs were extracted from news articles[15], the `uk-en` data is made up of segments collected from real use cases of the Charles Translator for Ukraine (Kocmi et al., 2022)[16], and contains segments that are noisier and not as well-formed as the news domain. Additionally, most of the training data of COMET-22 is also in the news domain (928,822 segments); the remanining 9,333 segments are from Wikipedia.

Nonetheless, Ukrainian–English was also the pair which improved the most after fine-tuning; the model fine-tuned on the large training set obtained an improvement of 0.11 in Spearman correlation compared to the base model.

Hausa–English was the only pair to show no improvement; the fine-tuned models have slightly low correlation scores than the base. To investigate this poor performance on the test set, we looked at the training logs, as COMET also calculates Kendall's Tau on the train and validation sets to monitor the performance at each epoch.[17] These scores show (Fig. 3.2) that there was a small but noticeable improvement in correlation on the training set, but the performance on the validation set almost did not change after each epoch. We hypothesize, therefore, that these models might have overfit on the training data and thus performed poorly in testing.

[15] 2021 WMT News Translation Task (`https://www.statmt.org/wmt21`, last accessed 06/09/2023.)

[16] 2022 WMT General MT Task (`https://www.statmt.org/wmt22`, last accessed 06/09/2023.)

[17] We trained every model with a minimum of 1 and a maximum of 3 epochs; some models stopped on the first or second epoch due to early stopping, because the validation performance started decreasing.

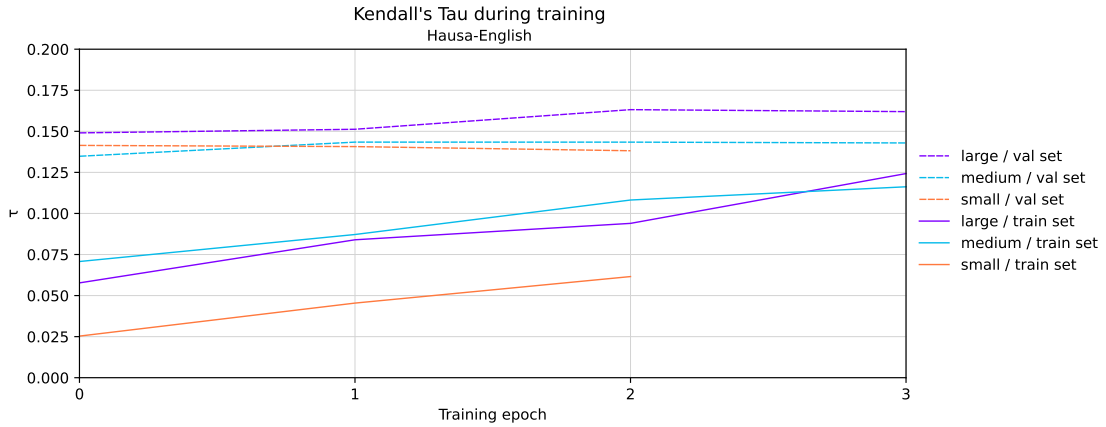

Figure 3.2: Kendall's Tau correlation scores for the three Hausa–English models on their corresponding train and validation sets.

Besides calculating the correlation measures, we also analyzed the segment-level quality scores generated by the COMET models for each segment in the test set. They are shown in Figure 3.3, with the corresponding human judgements for each segment in the y-axis. The diagonal line is where automatic and human scores would be exactly equal, so the closer our data points are to this line, the better.

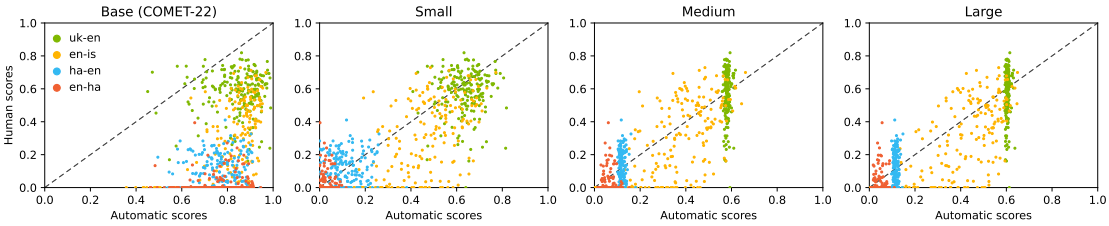

Figure 3.3: Quality scores generated by COMET-22 and by our fine-tuned models (x-axes), and the corresponding human scores (y-axis).

The plots for the medium and large models showcase a curious pattern: for the `ha↔en` and `uk-en` pairs, these fine-tuned models only produced results in the same short range, noticeable as the vertical clusters of markers between 0.0-0.2 for `ha↔en` and around 0.6 for `uk-en`. We hypothesize that this was due to the lack of diversity in human scores present in the data, and consequently in our training sets; we plotted the distributions of these sets to illustrate this issue (Fig. 3.4). Besides the obvious disproportional amount of near-zero scores for `ha↔en`, the medium and large sets are also entirely within the 0.0–0.3 range for these pairs and around 0.6 for `uk-en`, which explains the range of results the models were able to generate, while for `en-is` the data is more evenly distributed, thus,

so are the results. It is only fair to conclude that COMET needs to be trained on balanced distributions of scores to be able to properly discriminate translation quality.

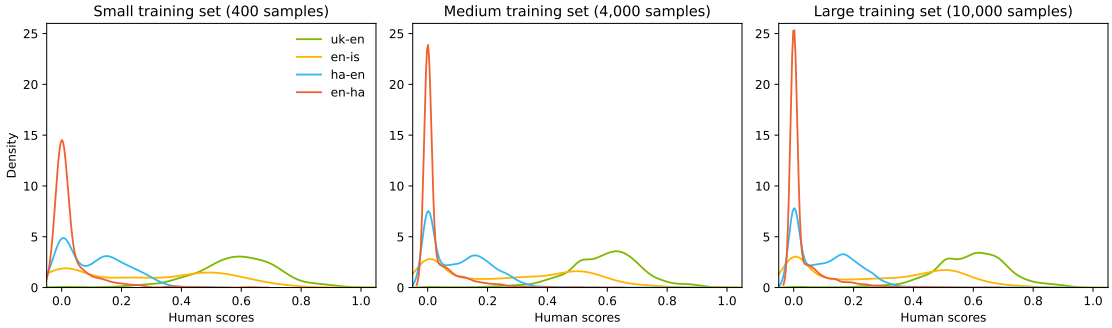

Figure 3.4: Density plots of the human scores in the incremental training sets for our fine-tuned models.

3.6 | Summary

This chapter covered preliminary experiments we did with the latest stable COMET model, in order to investigate the model's ability to generalize and evaluate translations to and from languages it has not seen in training. The languages we selected are supported by the XLM-R encoder, and thus presumed to be supported for evaluation with COMET.

Our results for test sets of 200 tuples showed that, with the exception of English–Icelandic, the correlations between human scores and the scores generated by COMET-22 were relatively low for the other three language pairs, Ukrainian–English and Hausa↔English. This raises the question of how well COMET models—and perhaps trainable metrics in general—are able to perform for language pairs that they were not trained to evaluate. We believe that such models should be evaluated more often on unseen language pairs, and not only on the same pairs as its training data, since there is a large number of language pairs that are presumed to be supported since they are supported by the underlying cross-lingual encoder.

In addition to that, we gathered data for these language pairs that had not been included in training, and used it to fine-tune COMET-22 with gradually large training sets, to analyze its improvement. Our largest training sets contained 10,000 samples, and we did see improvement of up to 0.11 points in Spearman correlation for `uk-en`. There were small improvements for the other language pairs, except for `ha-en`, for which the performance got slightly worse after fine-tuning, which we hypothesize might have happened due to overfitting on the training set.

Our main goal with this analysis was to explore COMET's generalization abilities, as well as play with the idea of fine-tuning as one of the possible strategies to improve models for unseen language pairs. This is highly important for our goals in the scenario of English–Maltese and Spanish–Basque, as Maltese and Basque are currently unsupported languages in the COMET framework, and our conclusions from this chapter have informed some of the choices we will discuss in upcoming chapters.

4

Manual Evaluation Campaign

In order to carry out a meta-evaluation study of MT for English–Maltese and Spanish–Basque, we needed annotated parallel data—parallel datasets of source sentences, reference translations and translation hypotheses, annotated with human quality scores, which we could use to compare against automated metrics and to train new MT evaluation models. To the best of our knowledge, there were no publicly available datasets that satisfied this requirement; therefore, we decided to run our own evaluation campaign to collect human judgements. This project is the first instance of a manual evaluation campaign for English–Maltese translations.

For translations into Basque, manual evaluations of various systems have been reported in previous works. Labaka (2010) proposed an SMT system for Spanish–Basque, and performed an evaluation based on HTER by collecting human post-edits for 100 outputs. Although too expensive to carry out at development time, they chose HTER to try and obtain a more reliable evaluation, as multiple publications had already risen doubts about BLEU (i.e. Melamed et al., 2003, Koehn and Monz, 2006). Later, Labaka et al. (2014) built three MT systems for Spanish–Basque with a hybrid rule-based/statistical architecture, and evaluated their systems in two steps: six native speakers of Basque evaluated 100 outputs through pairwise comparison, and three professional translators post-edited the same 100 outputs for evaluation with HTER. Both human evaluation steps produced results that partially contradicted those obtained from automatic metrics such as BLEU, WER and NIST.

More similarly to our project, Aranberri et al. (2017) carried out a large-scale, crowd-based evaluation campaign of English–Basque translations. As their goal was to investigate whether users found noticeable differences in quality between five MT systems, they conducted their evaluation in the form of pairwise comparisons, where users saw two hypotheses and were asked which one they found better or if both were of equal quality.

When we started planning our own evaluation campaign, the first decision we had to make was regarding who our participants would be. Due to our limited resources and budget, we designed our campaign for crowd-based, bilingual speakers to participate on a voluntary basis. Therefore, as our expected participants would be mostly non-professionals, it was especially important to be careful with the amount of effort required for the task. We wanted volunteers to be able to participate with minimal effort, and to understand the task easily, so that they would be encouraged to contribute as much as possible. These concerns were taken into account in the design of our evaluation tool, the selection of the data, the systems that we evaluated, and most importantly, the format of the evaluation task chosen.

In the following sections we will discuss in greater detail all the decisions involved in the preparation of this campaign.

4.1 | Participants

In an ideal scenario, it is usually recommended that evaluations like this be carried out by experts, such as professional linguists or translators. The main concern is that crowd-sourced judgements are more prone to inconsistencies, as the participants are mostly anonymous and have no verifiable language skills. However, as Graham et al. (2017) argued, it is possible to obtain reliable crowd-sourced judgements with appropriate quality control.

Ultimately, we agree with this view, and we also believe that regular bilingual speakers who actively use both languages in their day-to-day lives can still provide useful insight on what they perceive as good or bad translations (Graham et al., 2013). Both Maltese and Basque are minoritary languages, threatened by the presence of major co-official languages in the region, respectively English in Malta and Spanish and French in the Basque Country, but they have large communities of native speakers who are interested in actively engaging with initiatives to secure the survival of their language in the digital age.

Furthermore, we expect that the actual turnout would feature a mix of experts and language enthusiasts. This dissertation project was carried out in collaboration between the University of Malta (UM) and the University of the Basque Country (UPV/EHU), the two biggest references in terms of linguistics and NLP research for Maltese and Basque respectively. We were in contact with several researchers in both universities, who are bilingual speakers and have been working in Computational Linguistics with these languages for a long time, and we are very thankful that they could help us with the campaign,

participate in the evaluation and provide us with feedback as well.

4.2 | The Task

As we discussed in Section 2.2, there are essentially three major types of manual evaluation tasks: translation comparison or ranking, attribute evaluation (i.e. direct assessment) and error analysis (e.g. MQM). Error analysis would provide the most fine-grained evaluation, but it requires significant cognitive effort and time commitment from assessors, as they would need to identify and classify specific error spans in each text. Pair-wise comparison is cognitively the simplest, and translation ranking is harder but saves time by collecting multiple pairwise comparisons at once; however, these tasks were ruled out because they yield relative rankings, which allows only for a ranking of systems or individual hypotheses, with no measure of absolute quality or magnitude of difference in quality.

We therefore chose to collect direct assessments (DA), as a tradeoff between the cognitive effort for our assessors and the usefulness of the results for our purposes. With DA, a single hypothesis is evaluated at a time, and the assessor gives it an absolute ranking between 0 and 100. Thus, we obtain absolute scores which are more informative and can be used in more fine-grained analyses. Furthermore, as we mentioned in Section 2.4.2, DA is the standard method used for training COMET models, and DA scores are flexible enough that we could use them to train all three architectures: COMET-DA and COMET-QE with the normalized DA scores (with or without references), and COMET-RANK by converting them to relative rankings (DARR).

We structured the task in a format known as source-based DA, where assessors rate a translation hypothesis in the target language based on how well it expresses the meaning of a given source sentence. This is in line with the best practices recommended by most of the literature in MT evaluation (Läubli et al., 2020), in contrast with reference-based DA, where the assessor would instead see a reference translation and rate how well the hypothesis expresses the same meaning. Reference-based DA has the advantage of requiring only speakers of the target language, since it is basically a semantic similarity task, and thus it is commonly used in WMT tasks for language pairs where there might be a scarcity of bilingual assessors (Freitag et al., 2021b). We decided against it mainly because it might be counter-productive to ask for semantic similarity judgements with the goal of evaluating translation, unintentionally leading non-professional assessors into a superficial, lexical comparison of the sentences (Freitag et al., 2021a, 2022). Moreover, the evaluation would have been too tied to the quality of the references, which we preferred

to avoid even though we have only used reliable human-translated references. Thus, for source-based DA, we limited the pool of assessors to bilingual speakers in order to create space for hypotheses to be deemed good regardless of their similarity to a reference, and in addition, this format allows us to use the references as one of the systems for quality control (Bojar et al., 2018), explained further in Section 4.6.

4.3 | The Software

We set up our campaign using Appraise (Federmann, 2018)[1], an open-source, web-based framework for MT evaluation, largely used in the WMT campaigns. Appraise provides a simple and neat user interface with a responsive layout that works well both on computer and mobile screens, an easy registration and authentication process, and the framework for us to set up the source-based DA task with our own data and system outputs.

We did, however, thoroughly customize the system to better fit our scenario, as it was originally designed for a public of pre-selected evaluators. We adapted the interface to be available in both English and Spanish, and the task instructions were also available in Maltese and Basque[2], in order to make sure our users were able to understand them as well as possible. As our participants would be mostly non-professional, bilingual speakers, we aimed to explain the project just enough so that our goals and purposes were clear and that the task seemed simple and interesting, but without overwhelming them with too much technical detail (Fig. 4.1).

Upon registration, users were only required to create an username and a password (Fig. 4.2); we did not ask for email address or any personally identifiable information in order to maintain privacy and anonymity. The only other information we asked for was their proficiency level in each of the languages in their language pair, out of five options: "beginner", "intermediate", "advanced", "fluent" and "native". The purpose of this was to rule out potential participants who were not proficient enough[3] in both languages to be able to judge translations, as well as later on being able to analyze the distribution of our participants across proficiency levels they identified with.

Once registered and logged in, users could enter the task and rate as many translations as they wanted, over one sitting or multiple separate sessions. Figure 4.3 shows an example screen of the task. They see one source sentence identified as "original text", and one hypothesis labeled as "candidate translation", without any hint of what system it is from.

[1] `https://github.com/AppraiseDev/Appraise`

[2] We thank Marthese Borg (UM) and Nora Aranberri (UPV/EHU) for kindly providing these translations.

[3] We decided to allow participants who marked their proficiency level as "intermediate" or above.

The priming question at the top of the page asks participants to express how much they agree that "the candidate translation adequately expresses the meaning of the original text." This format of priming question for source-based DA could be seen as asking for a judgement of adequacy, which traditionally refers to how well the translation conveys the meaning of the source regardless of its fluency, but we also previously encouraged participants to "deduct points" for any potential disfluencies in the hyptotheses, and we believe such disfluencies would be the most obvious translation errors that people would notice. The user can then select their rating on a simple slider from 0% to 100%, with their selected value being displayed below. Once a rating is submitted, the user cannot go back to see or edit it.

The system was made available on a public website in the UM NLP research group domain.

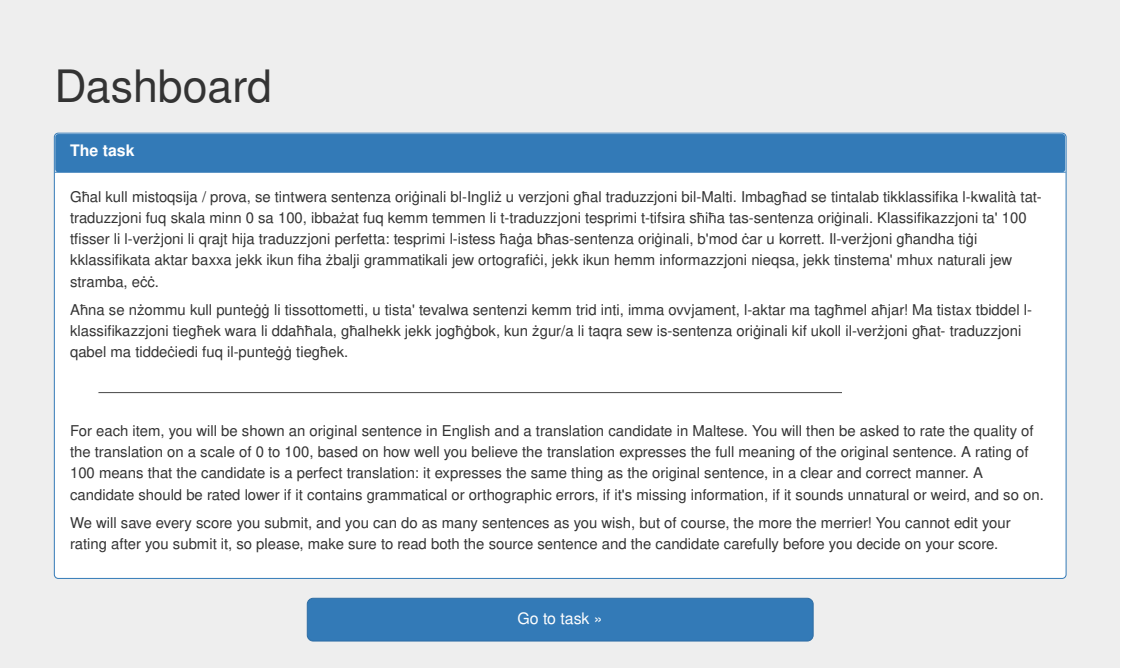

Figure 4.1: The dashboard page on Appraise.

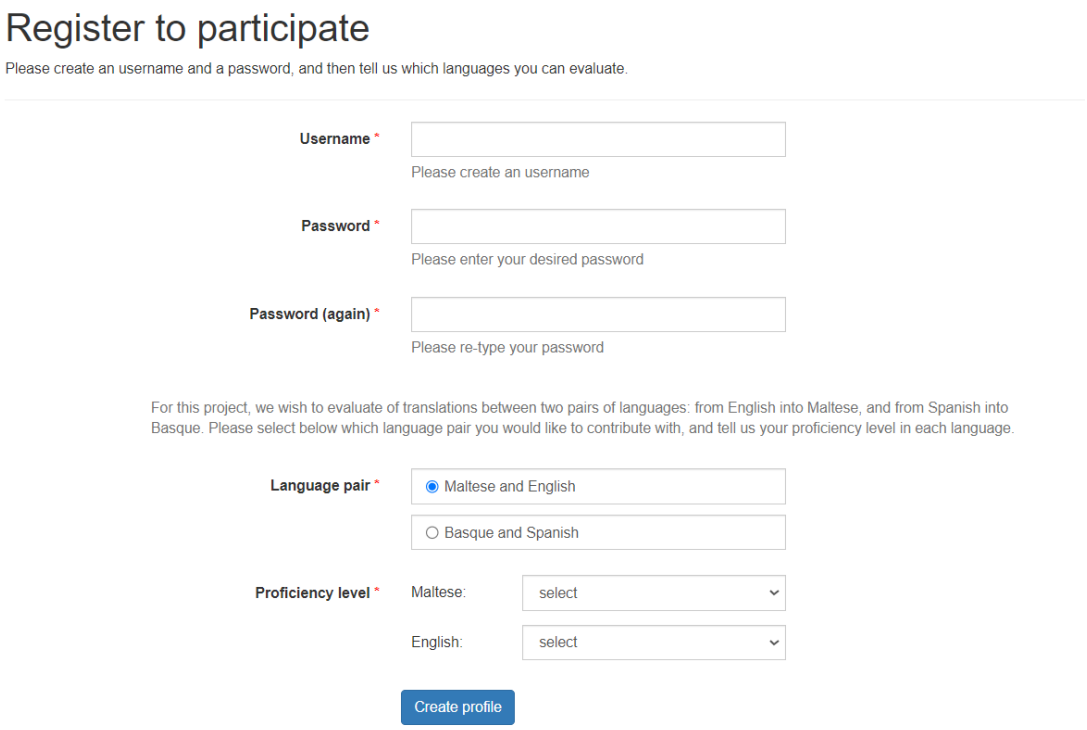

Figure 4.2: The registration page on Appraise.

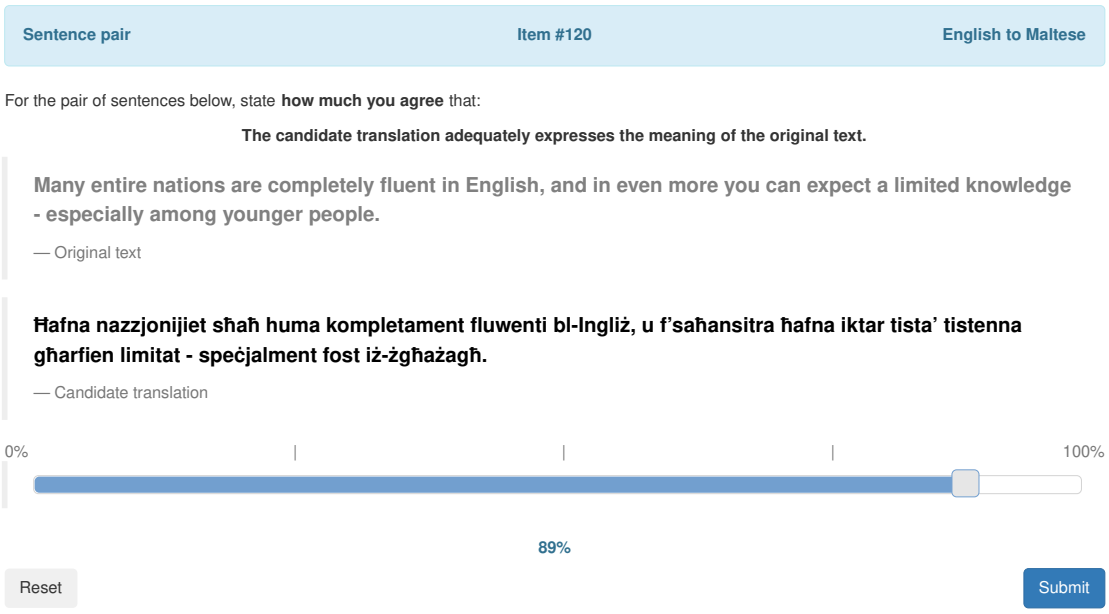

Figure 4.3: The task page on Appraise, showing an English–Maltese hypothesis to be evaluated.

4.4 | Datasets

Our main concern with the data selection for our human evaluation campaign was that the segments had to be relatively simple and easy for all our potential participants to understand. Each segment should be understandable on its own, without the need for additional context, and it shouldn not be too long or feature overly complex vocabulary. Moreover, we wanted the content to be diverse enough to keep the participants engaged.

We ruled out a large portion of the existing parallel datasets for both our language pairs, which consisted of texts in overly specific domains, such as legal documents, medical corpora and parliament publications. Especially in the case of English–Maltese, the overwhelming majority of corpora for this language pair were available thanks to Maltese being an official EU language since Malta joined the EU in 2004, meaning that most of the corpora consist of documents and publications from the EU parliament and other European agencies, such as Covid- and vaccination-related information, and so on.

With the remaining available corpora, we carried out several filtering steps, based on the source sentences. First the segments were tokenized[4], and we kept those between 5 and 50 tokens, although most of them ended up being in the range of 20–30 tokens. We also used regular expressions to check if the segments were well-formed, starting with capital letters and ending with periods or question and exclamation marks, thus removing incomplete segments extracted from the middle of full sentences, bullet points, and so on.

Lastly, from this filtered dataset, we selected the final segments by hand, in order to make sure that they would fit all of our criteria. We also had to check the references manually, as some corpora had misaligned source-reference pairs.

We selected a total of 400 segments for each language pair, an amount we believed would be feasible to obtain evaluations for. These sets are henceforth referred to as our evaluation sets. Table 4.1 shows the amount of sentences per corpus, and below is a brief description of each of these corpora.

- **FLORES-200**: A many-to-many multilingual translation benchmark by Meta AI.[5] It contains segments in 101 languages, including English, Spanish, Maltese and Basque; the texts were extracted from web articles and translated by professional translators (Goyal et al., 2021).
- **CrowS-Pairs**: A challenge dataset created with the goal of measuring the presence of U.S. stereotypical biases in masked language models (Nangia et al., 2020; Névéol

[4]We used the `word_tokenize` function from NLTK (`https://www.nltk.org`), which supports both English and Spanish.

[5]`https://github.com/facebookresearch/flores/blob/main/flores200/README.md`, last accessed 21/09/2023.

Language pair	Corpus	Count
English–Maltese	FLORES-200	281
	CrowS-Pairs	49
	EUbookshop	47
	ELITR-ECA	23
Spanish–Basque	FLORES-200	110
	TED2020	60
	Elhuyar	54
	OpenSubtitles	48
	EhuHac	46
	QED	40
	WikiMatrix	30
	NeuLab-TedTalks	12

Table 4.1: Number of segments per corpus in the evaluation sets.

et al., 2022).[6] Each example consists of a pair of segments, where one is an offensive statement about a minoritary group, and the other is the same statement but edited to be about an advantaged group. The segments were translated from English into Maltese and kindly provided to us via private correspondence; it will be made publicly available later in 2023.

We were careful to select relatively "mild" segments from this dataset, and also included a disclaimer on our system to warn participants that the sentences were not written by us and do not express the views of anyone involved in the project.

- **EUbookshop**: A corpus of documents shared by the Publications Office of the EU[7] and compiled by OPUS (Tiedemann, 2012).
- **ELITR-ECA**: Corpus of documents published by the European Court of Auditors (ECA)[8], crawled by the European Live Translor (ELITR) project (Williams and Haddow, 2021).
- **TED2020**: Parallel corpus of TED and TED-X talk transcripts, translated by volunteers (Reimers and Gurevych, 2020).
- **Elhuyar corpus**: Parallel corpus of textbook translations in Spanish and Basque, compiled by the foundation Elhuyar for the UPV/EHU.[9]

[6] `https://github.com/nyu-mll/crows-pairs`, last accessed 06/09/2023.

[7] `https://op.europa.eu/en/web/general-publications`

[8] `https://www.eca.europa.eu`

[9] `https://www.elhuyar.eus/en/services/language-services-and-basque-plans/translations-and-language-resources/corpus`

- **OpenSubtitles**: A corpus of translated movie and TV show subtitles from the OpenSubtitles website (Lison and Tiedemann, 2016).[10]
- **EhuHac**: Multilingual corpus of books compiled by the UPV/EHU.[11]
- **QCRI Educational Domain Corpus (QED)**: Corpus of subtitles from educational videos, transcribed and translated over the AMARA platform (Abdelali et al., 2014).
- **NeuLab-TedTalks**: Crawl of original subtitles and translations from TED Talks (Qi et al., 2018).
- **WikiMatrix**: Corpus of parallel sentences mined from Wikipedia articles by Facebook Research (Schwenk et al., 2019).[12]

We downloaded most of these corpora from OPUS (Tiedemann, 2012)[13] using the OpusTools package (Aulamo et al., 2020).[14]

4.5 | Systems

Once we had the data, we translated all of it using a variety of MT systems in order to have different translation hypotheses for our participants to evaluate. It is worth noting that this project is a meta-evaluation study, not an evaluation of the state of MT research itself; therefore, our focus is not on the quality of the MT outputs, but on the quality of the methods used to evaluate these outputs.

With this in mind, it was in our best interest to have as much variety as possible amongst the systems that we chose. At the same time, we could only include outputs from a limited amount of systems, considering the amount of evaluations that we could expect to obtain. We picked 3 systems for each language pair: one proprietary MT system, one open-source, publicly available model, and one in-house model, out of the new models currently in development at each of our partner universities, in order to collaborate by providing some human evaluations for these systems.

The one system we used for both language pairs is an open-source suite of many-to-many MT models developed by Meta AI, named No Language Left Behind (NLLB), which translate directly between any pair out of 200 available languages, including Maltese and

[10] `http://www.opensubtitles.org`
[11] `https://www.ehu.eus/ehg/hac`
[12] `https://github.com/facebookresearch/LASER/tree/main/tasks/WikiMatrix`
[13] `https://opus.nlpl.eu`
[14] `https://github.com/Helsinki-NLP/OpusTools`

Basque (Team et al., 2022). We used the 1.3B Dense Transformer variant[15], henceforth referred to simply as "NLLB", which is available through the HuggingFace `transformers` library.

[15] `https://huggingface.co/facebook/nllb-200-1.3B`

These are the systems we picked only for English–Maltese:

- **Google Translate (GT)**: Popular multilingual translation service developed by Google. There are not many details available to the public about the engine behind it, but since 2016 it is an NMT system that consists of a deep LSTM encoder-decoder network with attention (Wu et al., 2016). It is trained on parallel data collected from the public web, as well as synthetic parallel data and monolingual data for under-resourced languages (Caswell and Bapna, 2022; Caswell and Liang, 2020). One potential downfall of GT is that it supposedly pivots through English for most of all its language pairs, with a few exceptions that are supported directly (Benjamin, 2019), which could affect the quality of Spanish–Basque translations; hence our decision to use it only for English–Maltese. We obtained translations from GT with the `deep-translator` pip package.[16]

- **UM-IWSLT**: The first open-source NMT system for English–Maltese translation, developed at the University of Malta (Abela, 2023).[17] It was trained on a series of English↔Maltese corpora available on OPUS, including EU documents, Maltese government documents and laws, and the Tatoeba[18] and MaCoCu[19] corpora. We used a version that had been improved and re-trained for submission to the 2023 International Conference on Spoken Language Translation (IWSLT) (Williams et al., 2023).

And this is our choice of systems for the Spanish–Basque language pair:

- **Itzuli**: NMT system provided by the Basque Country government through their official website.[20] It translates between Basque and Spanish, English, and French, and can be set to use either the "general" or the "legal-administrative" domain. This system does not provide an API, so we obtained translations in batches on the web interface, and we used the "general" domain since our texts are mostly not legal-administrative.

- **UPV-CMBT**: A model developed at HiTZ (the Basque Center for Language Technology), based on the Transformer architecture (Vaswani et al., 2017) and built with the Marian toolkit (Junczys-Dowmunt et al., 2018). It employs inline case markers

[16] `https://pypi.org/project/deep-translator`
[17] `https://github.com/kurtabela/MSc-Thesis`
[18] `https://tatoeba.org`
[19] `https://macocu.eu`
[20] `https://www.euskadi.eus/traductor`

(Berard et al., 2019) to handle text casing, and uses 251 million Basque tokens obtained from monolingual data for back translation (Sennrich et al., 2016). Its training data contains both public and proprietary parallel corpora, consisting mostly of news articles and administrative documents, but also including web-crawled data, literary texts and film subtitles. This model has not been published or made available to the public yet, so we thank Gorka Labaka (UPV/EHU) for providing us with the translations.

Table 4.2 provides a summary of the systems chosen for each language pair.

Language pair	**System**	**Type**
English–Maltese	Google Translate	Proprietary
	NLLB	Open-source
	UM-IWSLT	In-house
Spanish–Basque	Itzuli	Proprietary
	NLLB	Open-source
	UPV-CMBT	In-house

Table 4.2: Machine translation systems selected for evaluation.

4.6 | Quality Control

In any evaluation campaign, but especially crowd-based ones, there is the concern that some participants might not be performing the task very well, even if acting in good faith (Graham et al., 2013). They might have misunderstood the instructions, might not be paying full attention, or trying to go through the items as fast as possible without thinking too much.

In order to identify these assessors, control tasks can be included amongst the evaluation tasks. These tasks are designed to have an expected response, so that if a judge's score diverges too much, it means they are not doing the task properly and their scores should be discarded. Additionally, control tasks allow reliable evaluators a moment of rest—if a translation is easy to judge, one way or another, it gives them a break from other tasks which they have to think longer about (Aranberri et al., 2017).

We employed two types of control tasks, `REF` items and `BAD` items. `REF` items are human-translated references from our parallel datasets, passed as if they were MT outputs, thus expected to have a high score. `BAD` items are created by damaging a regular MT output, and thus they are expected to have low scores.

There are multiple strategies to produce damaged MT outputs, in a way so that their low quality is not immediately obvious and the sentence still appears legible to the innatentive eye. Graham et al. (2013), for example, created them by duplicating two random words from the output in random positions. We followed another strategy implemented by Appraise, which consists of replacing a random part in the middle of an MT output with a segment of similar length taken from a randomly chosen reference, in the same language, resulting in an output that has a nonsensical piece of another text in the middle, but superficially looks like a grammatical segment. Table 4.3 exemplifies this procedure with one `TGT`–`REF`–`BAD` item triplet from the English–Maltese set.

Source segment	Nowadays air travel is only rarely booked directly through the airline without first searching and comparing prices.
Reference (`REF` item)	Illum il-ġurnata l-ivvjaġġar bl-ajru rari jiġi bbukkjat direttament permezz tal-linja tal-ajru minħajr ma l-ewwel isir tiftix u paragunar tal-prezzijiet.
MT output (`TGT` item)	Illum il-ġurnata l-ivvjaġġar bl-ajru huwa rari biss ibbukkjat direttament permezz tal-linja tal-ajru minħajr ma l-ewwel wieħed ifittex u jqabbel il-prezzijiet.
Another reference, picked randomly	Dan is-servizz huwa b'xejn, u b'mod faċli, jista' jkollok aċċess għall-aħbarijiet riċenti, fatti u figuri, dokumenti legali u għadd kbir ta' informazzjoni prattika.
Damaged output (`BAD` item)	Illum il-ġurnata l-ivvjaġġar bl-ajru huwa rari biss ibbukkjat direttament permezz tal-linja tal-ajru minħajr jkollok aċċess għall-aħbarijiet riċenti, fatti jqabbel il-prezzijiet.

Table 4.3: Example of `TGT`, `REF` and `BAD` items for a given source segment, taken from the English–Maltese evaluation set.

These are strategies for quality control that we have seen in the literature and we deemed reasonable for our campaign, but of course, there are potential drawbacks. The human references, for example, are not necessarily all worth the highest rating; their "quality" is still debatable, as for all translations, especially if compared to the highest-performing systems, and so we might find a large variation of scores. With the damaged outputs, we treaded a fine line to damage them just enough so that they would deserve lower scores than regular outputs, but also that they would not be too obviously horrible, in order to try and identify whether assessors were really paying attention; on the other hand, there might be assessors who end up rating the segment proportionally to how much of it is well translated. Therefore, there are many variables to consider, which will

be discussed further when we present the results (Section 5.1.2).

For each language pair, with 400 sentences in the evaluation set and 3 systems, the total amount of source–hypothesis pairs is 1200. This number could be reduced if two hypotheses from different systems happened to be equal, in which case they would be evaluated as one, meaning that all source–hypothesis pairs are unique; this, however, did not happen with any of our segments, likely because we tested systems from which we expected very different levels of translation quality.

The amount of control tasks is defined per batch of 100 items on a ratio of 80 : 10 : 10, so out of each 100 items, 80 are `TGT`, 10 are `REF` and 10 are `BAD` items.

4.7 | Dissemination

Once the campaign website was ready, we started spreading the word in channels that seemed appropriate to reach bilingual speakers to participate. We shared it internally first, in channels for the respective NLP research groups inside the UM and the UPV/EHU, and then also posted about it on Facebook groups related to the minoritary languages, on local mailing lists for linguists and for Maltese and Basque translators, and asked for it to be shared on internal university communication channels to students and faculty.

Below is the English version of the statement that was included in the front page of the campaign website, and also used in the posts and messages that we wrote to share the campaign:

> *Are you a bilingual speaker of Maltese and English, or Spanish and Basque? If so, we need your help! We are researchers from the University of Malta (UM) and from the University of the Basque Country (UPV/EHU), and we are conducting this study in order to evaluate the quality of translations into two under-resourced languages, Maltese and Basque. Translation is not an easy task, so we would like to hear from you: if you wish to participate, we will show you some sentences and ask you to rate the translations, and that would help us a lot with our research to improve them in the future. If you want to read more and try it out, go on and create a profile — we won't ask you for any private information, just tell us what languages you speak and you're good to go!*

4.8 | Turnout

We started the evaluation campaign on August 1st, 2023. For the purpose of this dissertation, the results we will discuss below were collected from the beginning

English–Maltese	
Total evaluations	992
Per item type	
↪ `TGT`	811
↪ `BAD`	101
↪ `REF`	80
Per system (`TGT` only)	
↪ Google Translate	274
↪ NLLB	252
↪ UM-IWSLT	285
Total participants	41
Avg. evaluations per user	24

Spanish–Basque	
Total evaluations	1215
Per item type	
↪ `TGT`	996
↪ `BAD`	133
↪ `REF`	86
Per system (`TGT` only)	
↪ Itzuli	354
↪ NLLB	293
↪ UPV-CMBT	349
Total participants	44
Avg. evaluations per user	27

Table 4.4: Statistics of the human evaluation campaign.

up to October 3rd, 2023, thus spanning 63 days.

To recapitulate, our ultimate goal was to obtain at least 1500 annotations for each language pair, 1200 of which would be actual MT outputs to evaluate (400 segments × 3 systems), and the remaining 300 would be our quality control tasks.

We were not able to meet this goal, and thus we do not have exactly equal amounts of evaluations per language pair and per system, which will be taken into consideration when we discuss our results. Table 4.4 describes how many scores we received for each language pair, divided by system and by item type, and also the numbers of participants.

4.9 | Summary

In this chapter we have covered the design and preparation of our own human evaluation campaign, where we asked bilingual speakers to rate translations from English into Maltese and from Spanish into Basque, in order to collect direct assessments for our meta-evaluation analysis. The decisions we made in the design of this campain are based on previous meta-evaluation studies that we discussed in Chapter 2, regarding the types of human evaluation tasks that have typically been done in MT and the application of each of them depending on the needs of the project and the availability of resources.

The results of this evaluation campaign, as well as the analysis that we have carried out with these results as data, will be detailed in the next chapter.

5

Meta-Evaluation Analysis

Once we collected a number of evaluations for English-Maltese and Spanish-Basque translations in the form of Direct Assessments (DA), we were able to perform a small-scale meta-evaluation study on the performance of COMET models for evaluating these language pairs. The present chapter describes this analysis, as a case study on the usability of COMET for unsupported languages.

First, in Section 5.1 we describe our data pre-processing steps: standardizing the raw DA scores according to each user's averages, and filtering these scores based on our quality control procedures.

We start our analysis with a system-level evaluation of our 3 MT systems (Section 5.2), according to human judgements from the campaign and to COMET-22, as well as a few lexical metrics for comparison. We look at how each method ranks the systems, and also at the magnitude of difference between each system score.

Then, as we knew beforehand that COMET technically does not support Maltese and Basque, we also explore potential routes of improvement and discuss two strategies we devised (Sec. 5.3); the idea was to introduce assessment data in these languages to try and obtain more accurate scores from the models. To do this, we tried fine-tuning COMET-22 on our DA scores, and also trained new COMET-DA models from scratch, exploiting the possibility of switching XLM-R for custom encoders developed specifically for Maltese and Basque.

We finish this chapter with a discussion on the results obtained from this study, and what they say about how COMET functions and what might be done in the future to extend its language support.

5.1 | Data Pre-Processing

This section details the pre-processing steps we performed on the raw scores from our evaluation campaign, to prepare them for analysis.

5.1.1 | Standardization

Different assessors tend to have very different scoring strategies; some of them might stick to the middle of the scale, adding or subtracting only a few points to express that each translation is better or worse, while others might go straight for the extremes. In order to iron out these differences, the raw scores are converted to z-scores, a measure which expresses each value's relation to the mean (Kreyszig, 1979). Standardization is applied by WMT organizers to all the DA scores obtained through their campaigns (Bojar et al., 2016b), which are used to train COMET and were also used by us for evaluation and fine-tuning in Chapter 3, so we did the same with our own campaign results.

The z-score for a raw score x is calculated by Equation 5.1, based on the user's mean score (μ_{user}) and standard deviation (σ_{user}) over all segments that they scored.

$$z_x = \frac{x - \mu_{user}}{\sigma_{user}} \tag{5.1}$$

Figure 5.1 shows the distributions of raw scores and the corresponding z-scores from two randomly chosen participants, to demonstrate the effect of the standardization. It is noticeable how standardization shrinks the peaks in user B's distribution, so that both users' standardized scores are now centered around 0, which is the z-score value when the raw score corresponds to the mean.

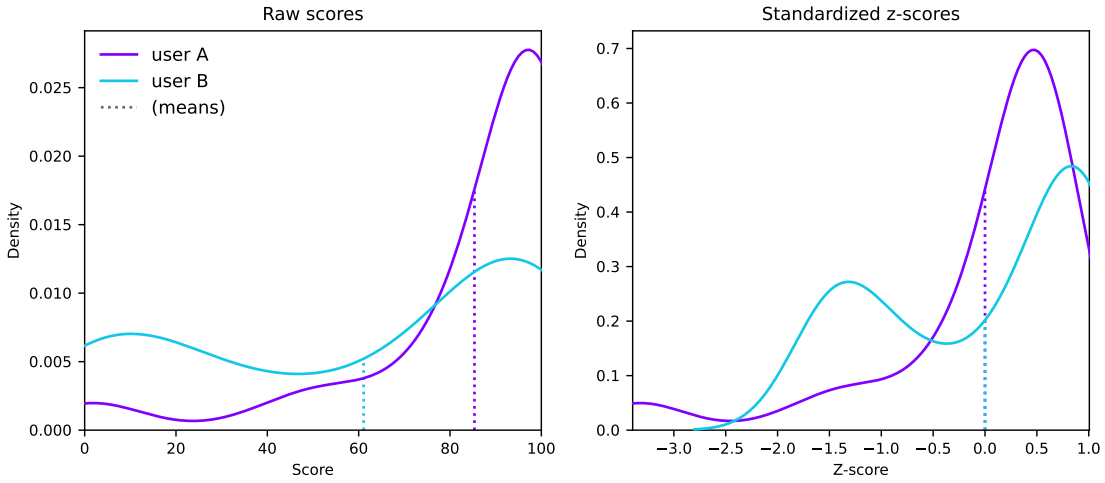

Figure 5.1: Density plots of two participants' raw scores and the corresponding z-scores.

5.1.2 | Quality Control

As we described in Section 4.6, we included a number of control tasks amongst the items in our campaign, in order to identify participants who might not have performed the evaluation properly. There were two types of control tasks: `BAD` items, which were damaged MT outputs, intended to receive low scores, and `REF` items, human translations which should be given high scores. Fig. 5.2 show the distribution of the raw scores assigned to both types of control tasks across the English–Maltese and Spanish–Basque evaluations.

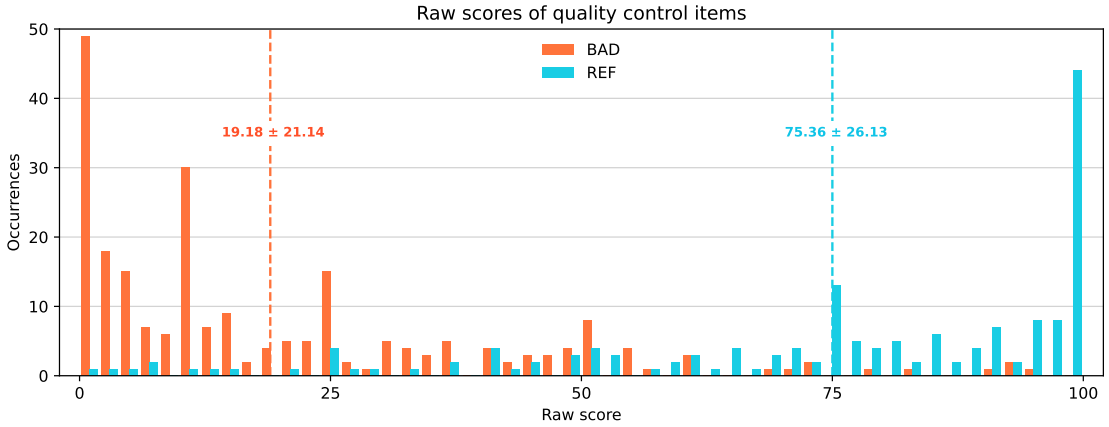

Figure 5.2: Distribution of raw scores received for quality control items, showing the average and standard deviation for each item type.

We tried as best as possible to follow the best practices in quality control as recommended by the literature in MT meta-evaluation (Freitag et al., 2021a; Graham et al., 2013); however, we had to adapt them to the particular circumstances of our campaign. In these large-scale studies, there is a pre-defined pool of assessors who receive batches of items to evaluate; therefore, each batch could include pairs of `TGT-BAD` or `TGT-REF` items (return to Table 4.3 for an example). In simplified terms, each `REF/BAD` item is then considered "passed" if the `REF` score is higher than the `TGT` score, and the `BAD` score is lower.

We, however, left it up to each participant to evaluate as many sentences as they wished, which made it complicated to decide when to show them the quality control task pairs. We could not ensure that each participant would evaluate the corresponding `TGT` item of each control item, so we defined another method for filtering our scores. For the purposes of judging the control item scores, we decided to simplify the 0-100 DA scale as one of binary choice, as if a translation could only be voted as "good" or "bad", which would correspond to values above or below

50. Therefore, a BAD item should be rated lower than 50 to be considered passing, and a REF item should receive a score of at least 50.

However, as we discussed in the previous section, different participants have different scoring strategies. In extreme cases, there were participants with dozens of evaluations within a range of high scores (80±20), but who still rated the BAD and REF items lower or higher than their TGT items, which is what we were looking for. To cover such cases as well, we looked at the standardized z-scores, considering that a z-score of 0.0 corresponds to the user's average; a given control task was also considered "passed" if the z-score was below 0.0 for BAD items, and above for REF items.

Table 5.1 shows the amount of quality control items that were evaluated, per language pair and per type, and how many failures occurred. It also shows the number of participants responsible for the failures, and how many TGT items were discarded as a consequence, since such participants were deemed unreliable.

	en-mt	es-eu
Item occurrences	181	219
↪ BAD	101	133
↪ REF	80	86
Failures	8	11
↪ BAD	4	2
↪ REF	4	9
Unique participants	5	8
Discarded evaluations	183	361

Table 5.1: Statistics of the quality control procedures.

It is notable that there is a large number of discarded items submitted by a small number of "unreliable" participants, which indicates that these few participants completed quite a lot of evaluations each; this might corroborate that they failed because they were doing the evaluations too fast and/or without paying full attention.

We also note that there were more failures for REF items than for BAD items. It might be that the BAD items were only seen as partially bad, as they only contained a segment out of context inserted in the middle of a regular MT output, so these translations were not complete nonsense. Moreover, the quality of the BAD items could also vary amongst them since they were based on MT outputs from different systems.

It could also be, on the other hand, that the `REF` items were underestimated by the participants. Freitag et al. (2021a) pointed out that human crowd-workers tend to underestimate human translations (references), or overestimate MT outputs, and that professional translators are better at distinguishing them appropriately. Although our `REF` items received a high mean score of 75.36±26.13, this is not much higher than the mean of 69.16±29.04 over all `TGT` items. Similarly, Freitag et al. (2021b) saw that human crowd-sourced participants ranked human translations below multiple MT systems in DA evaluation. They hypothesized that a method like DA might lead assessors to prefer more literal translations which are easy to compare to the source and judge, while a more fine-grained method of error analysis (like MQM) forces them to mark specific errors to justify their scores; thus, error analysis encourages assessors to be accepting of non-literal human translations, and to penalize errors in MT outputs more fairly (Freitag et al., 2021a).

Table 5.2 summarizes the amount of `TGT` item evaluations we had, per language pair and system, before and after filtering out the scores from unreliable participants. We retained 77% of the evaluations for English–Maltese, and 63% for Spanish–Basque.

en-mt			es-eu		
System	**Before**	**After**	**System**	**Before**	**After**
Google Translate	274	214	Itzuli	354	228
NLLB	252	189	NLLB	293	192
UM-IWSLT	285	225	UPV-CMBT	349	215
Total	811	628	Total	996	635

Table 5.2: Number of `TGT` items per system, before and after filtering based on quality control.

5.2 | System-Level Evaluation

Using the filtered DA scores from our evaluation campaign, we performed a system-level evaluation of the 3 MT systems we considered for each language pair. Our aim is to compare the results from different automatic methods with human judgements. In the following subsections we will describe our approach to calculate system scores from each method, the results from each, and discuss our conclusions.

5.2.1 | System Scores From Automatic Metrics

We scored each system's outputs using COMET-22[1] out of the box, and for comparison, we also obtained scores from BLEU, TER and CHRF. For these lexical metrics, we used the SacreBLEU implementation, with the default parameters for each metric; below are the corresponding SacreBLEU signatures:

- BLEU: `nrefs:1|case:mixed|eff:yes|tok:13a|smooth:exp|version:2.3.1`
- CHRF: `nrefs:1|case:mixed|eff:yes|nc:6|nw:0|space:no|version:2.3.1`
- TER: `nrefs:1|case:lc|tok:tercom|norm:no|punct:yes|asian:no|version:2.3.1`

For all of the lexical metrics, as well as COMET-22, the system-level score is the average of all segment-level scores, assigned to each of the 400 segments translated by each of the 3 systems. Therefore, we used automatic metrics to rate our whole evaluation set, and the results are reported in Table 5.3.

	System	COMET-22	BLEU	TER*	CHRF
en-mt	Google Translate	0.7434 #1	43.95 #1	39.49 #1	73.66 #1
	NLLB	0.6938 #2	24.73 #2	64.47 #3	62.95 #2
	UM-IWSLT	0.6885 #3	23.82 #3	61.09 #2	59.09 #3
es-eu	Itzuli	0.8367 #2	15.35 #3	79.20 #3	54.17 #3
	NLLB	0.8282 #3	27.19 #1	69.36 #1	56.80 #1
	UPV-CMBT	0.8371 #1	15.61 #2	78.50 #2	54.36 #2

Table 5.3: System scores assigned by each metric, based on the evaluation sets of 400 segments. (*As opposed to the other metrics, TER is a measure of *error*, so the lower the value, the better.)

5.2.2 | System Scores from Human Evaluation

To calculate the system-level scores based on the human evaluations, first we averaged out the available scores per segment, in case of redundancy, and then calculated the average over these segment-level averages. We did this procedure for both the raw DA scores and the standardized z-scores; therefore, the resulting system level scores are presented in the 0–100 range and also in standardized form, calculated from the segment-level z-scores. The z-score measures the amount of

[1] `https://huggingface.co/Unbabel/wmt22-comet-da`

standard deviations between each segment-level raw score and the user's mean score, so the system-level z-score can be interpreted as "above average" if greater than zero, and "below average" if lower than zero.

Table 5.4 shows the system scores alongside their respective rankings; since we were not able to obtain evaluations for all the segments from all systems, and these evaluations have undergone quality control, the table also shows the number of evaluations taken into account for each system.

	System	Evaluations	System scores	
			Raw	Standardized
en-mt	Google Translate	214	82.03 #1	0.593 #1
	NLLB	189	65.43 #2	0.116 #2
	UM-IWSLT	225	49.11 #3	-0.425 #3
es-eu	Itzuli	228	82.81 #1	0.439 #1
	NLLB	192	63.60 #3	-0.170 #3
	UPV-CMBT	215	82.42 #2	0.358 #2

Table 5.4: System scores resulting from the human evaluation campaign, and their corresponding rankings. Raw scores are calculated from original DA scores (0–100), and standardized scores are from the z-scores.

5.2.3 | Discussion

In the previous subsections, we presented the results of our system-level evaluation, based on the evaluation set of 400 segments translated by 3 MT systems for each language pair. We obtained quality scores from 4 automatic metrics, on the whole set, which were averaged out to generate system scores (Table 5.3), and then also calculated system scores from human judgements on the segments that we received evaluations for (Table 5.4). Based on these system rankings and the magnitude of difference between each method's scores, we can make a few observations.

Based on these rankings and the magnitude of difference between each metric's scores, we can make a few observations.

For English–Maltese, the ranking of GT > NLLB > UM-IWSLT is agreed on by humans and by nearly all metrics, except for TER, which assigns a slightly higher edit rate for NLLB than UM-IWSLT. However, despite the agreement on the ranking, we note the deltas: COMET-22 rates the best and worst systems quite closely (0.74 vs. 0.68), while human scores are 40% lower (82.03 vs. 49.11). In

this case, it seems that the other metrics better capture the degree to which GT is deemed better than both NLLB and UM-IWSLT.

As for Spanish–Basque, both human scores and COMET-22 indicate that Itzuli and UPV-CMBT are the best performing systems and are very similar to each other in quality (0.83 from COMET-22 and 82 from humans), while all the lexical metrics put NLLB as the best system. The BLEU score for NLLB places it 12 points ahead of both Itzuli and UPV-CMBT, while human participants seem to have found NLLB significantly worse than both other systems. Interestingly, Itzuli and UPV-CMBT being rated so closely by all the methods is an impressive result for UPV-CMBT, a new model developed in the University of the Basque Country (UPV/EHU), which seems to be on par with the translator provided by the Basque Government. Nevertheless, like in the English–Maltese results, COMET-22 underestimates the magnitude of difference between the 3 systems, rating NLLB only 0.01 less than the others (as opposed to a delta of 19 in human scores).

We emphasize that this is a discussion based on preliminary results; ideally, we should compare the metrics with human judgements on the whole evaluation set, and with redundancy on all the segments, for a more reliable evaluation. As we have not been able to obtain so many evaluations yet, we presented the results from what was available.

5.3 | Improvement Strategies

Besides using our DA scores for evaluation of existing metrics, another one of our goals was to investigate whether we could improve COMET's performance on our language pairs by using this data at training time. There are two ways in which we could do this: fine-tuning existing COMET models on our data, or training custom COMET models from scratch.

Fine-tuning, the same way we did with 4 unseen language pairs in Chapter 3, consists in using an existing pre-trained COMET model, namely COMET-22, as initial checkpoint, and then training it for a few more epochs on our data. It was also done by Sai et al. (2023), where the authors fine-tuned COMET-MQM on 5 Indic languages and achieved improvements in correlation scores, as we mentioned previously.

The fine-tuning approach has potential advantages and downsides: COMET-22 is a large model, built on top of `xlm-roberta-large`, and trained on a total of 938,155 samples, covering 32 different translation directions in 18 languages (listed

in Table 3.1). Although Maltese and Basque are not supported by XLM-R nor seen in the training data, COMET-22 and other pre-trained COMET models are much larger than what we can build with our data, and it might be that they are able to generalize to unseen languages. The other languages in each pair, English and Spanish, are both supported, and English is present in many other language pairs, which might help the model learn to evaluate English–Maltese anyway. The idea behind fine-tuning is to try and improve this model by showing it a few hundred examples of the language pairs we wish to evaluate.

On the other hand, training new models from scratch consists in using the framework's base structure to train a whole new DA Estimator using our DA scores from the campaign. The main advantage of this approach is that we can switch the cross-lingual encoder, and plug in different encoders that support our under-resourced languages, which ensures that they can be encoded better. The downside is that the dataset we have is much smaller than the usual training datasets of COMET; therefore, we are conducting this experiment as a proof-of-concept rather than an actual attempt to generate a reliable, usable model.

COMET-22 (pre-trained model)	
Initial checkpoint	COMET-22
Pre-trained encoder	XLM-RoBERTa (large)
Training data	COMET-22 training data
COMET-22-FT (COMET-22, fine-tuned)	
Initial checkpoint	COMET-22
Pre-trained encoder	XLM-RoBERTa (large)
Training data	COMET-22 training data + our DA scores
COMET-DA (custom model, trained from scratch)	
Initial checkpoint	None
Pre-trained encoder	mBERTu / IXAmBERT
Training data	Our DA scores

Table 5.5: Summary of the three types of COMET models we are using for evaluation.

Table 5.5 sums up the models we compare in this section: the default COMET-22 without any changes, COMET-22-FT, fine-tuned on our DA data, and brand new COMET-DA models trained exclusively on our data, on top of selected encoders. It is worth clarifying that each COMET-22-FT and COMET-DA model was trained and tested on English–Maltese and Spanish–Basque *separately*, generating 4 new models in total.

Both approaches require a number of pre-processing steps and other technical decisions we will detail in the upcoming subsections. First, we describe a different approach to rescale the z-scores to the range of [0,1] (Sec. 5.3.1); Section 5.3.2 explains how we split the data into train, validation and test sets using stratified sampling; in Section 5.3.3, we discuss our choice of custom encoders for the new COMET-DA models. Finally, we present our results in Section 5.3.4, and discuss what we can learn from these experiments in terms of COMET's performance for unsupported languages and the decisions involved in training new models.

5.3.1 | Rescaling

Looking at the distribution of scores in our campaign results, we decided to make a change in the rescaling step of our data pre-processing, in comparison to what we did with our initial experiments in Chapter 3.

We emphasize the difference between "rescaling" and "standardization": standardization is applied to raw DA scores, based on each user's mean raw score, to smooth over any differences in scoring strategies; rescaling is applied to the whole population of (already standardized) z-scores in the dataset, so that they are all in the range of [0,1].

Rescaling is necessary because the latest COMET models are supposed to produce scores in the range of [0,1], which requires the input data to be in this range, and z-scores are originally unbound. Rescaling has not been incorporated into the COMET training pipeline, so it has to be done as a pre-processing step.

For our first experiments with other language pairs, as described in Section 3.3, we rescaled the z-scores according to the procedure recommended by COMET developers, by min-max scaling the scores into a range of $[r_{min}, r_{max}]$ values and then clipping them to [0,1]. The problem is that this range of $[r_{min}, r_{max}]$, following their suggestion, is between the average z-score of segments rated 0 by more than 1 annotator (r_{min}), and the average z-score of segments rated 100 by more than one annotator (r_{max}). This might work well for large datasets, where these averages are meant to be reasonable values and then only outliers are clipped out; with our

small sets, however, the range would be [-1.47, 0.8] for `en-mt`, for example, and then all values scaled to [-1.47, 0.0] would be clipped to 0. Figure 5.3 illustrates the consequence of this: the second graph shows the disproportional amount of scores now clipped to [0.0, 0.05], which is unfaithful to the actual distribution of the data (first graph).

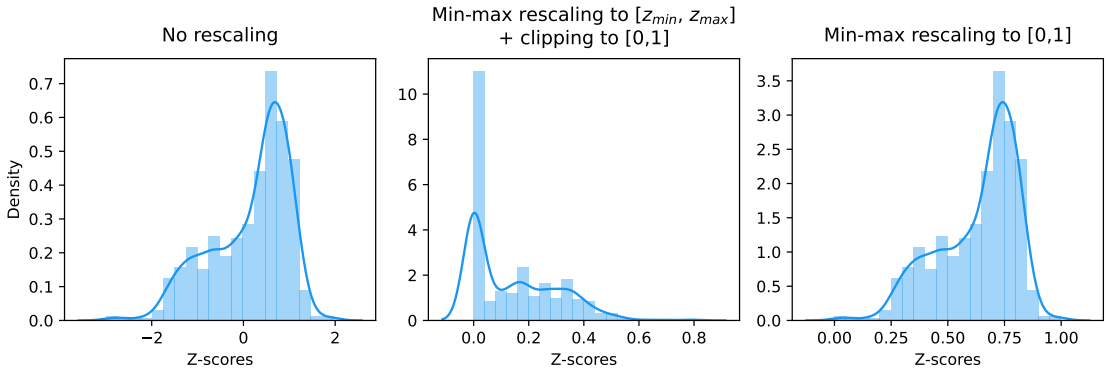

Figure 5.3: Distribution of z-scores before rescaling, and after rescaling by two different methods.

Instead of defining the value range based on the raw DA scores and then clipping, we simply rescaled the data to the range of [0,1] directly.[2] Equation 5.2 shows how each z-score x is scaled to [0,1] based on the extreme z_{min} and z_{max} values[3] in the whole set (Pedregosa et al., 2011). This leaves us with rescaled z-scores that respect their original distributions, as shown in the third graph in Fig. 5.3.

$$x_{scaled} = \frac{x - z_{min}}{z_{max} - z_{min}} \tag{5.2}$$

5.3.2 | Stratified Sampling

Another lesson learnt from the experiments in Chapter 3 was regarding the distribution of quality scores that our fine-tuned COMET models produced: when the training data was concentrated around a restricted range of human z-scores, COMET only learned to produce scores in that same range, and performed poorly in validation and testing. Therefore, the data splits should ideally be representative, as much as possible, of the whole range of scores available.

[2] We still used the `MinMaxScaler` from `scikit-learn`, but with the default `feature_range` of [0,1].

[3] Unlike r_{min} and r_{max}, which are conditioned on the annotator count and the corresponding raw DA scores, z_{min} and z_{max} are simply the minimum and maximum z-scores in the dataset.

In order to avoid that issue in our subsequent experiments, we used stratified sampling, which is a strategy to ensure that a random sample from a dataset is representative of the original data (Thompson, 2012). It is designed for categorical data, where the goal is for the sample to contain a representative amount of items from each class in the dataset. In order to apply it to continuous data, as is the case of our z-scores, the data can be digitized, assigned to a number of bins that correspond to smaller ranges of values and will act as "categories" (Sanders, 2017).

We digitized each of our datasets into 10 bins in order to sample the test and validation sets, and then left the remaining samples as the training set. This way, as can be seen in Figure 5.4, the test and validation sets have similar distributions and cover the range of scores we had available.

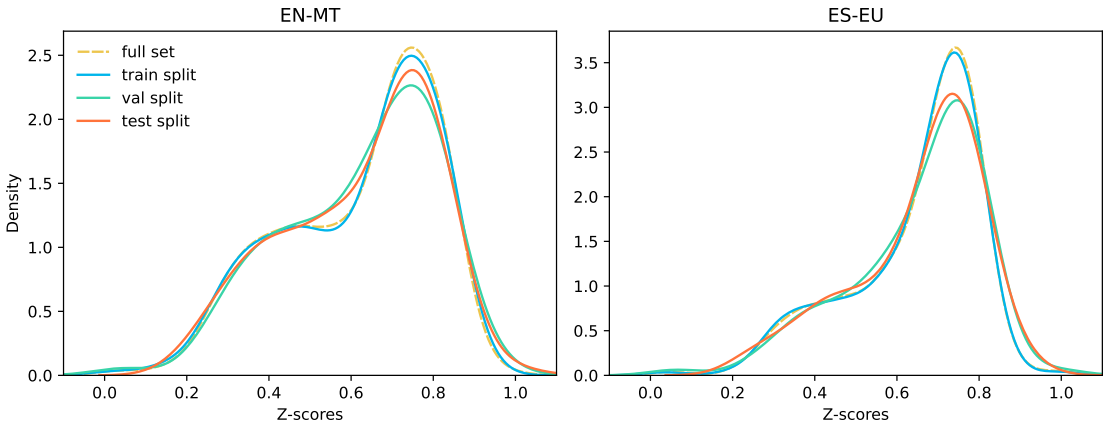

Figure 5.4: Distribution of z-scores in the original datasets and the split sets.

5.3.3 | Encoders

When training new COMET models from scratch, we can plug in custom encoders instead of using XLM-RoBERTa; therefore, we used encoders built specifically with our under-resourced languages in mind.

Micallef et al. (2022) proposed BERTu[4], a monolingual model for Maltese based on the BERT architecture, as well as mBERTu[5], a multilingual model based on mBERT and further trained on Maltese data. Both models were trained on the Korpus Malti v4.0[6], a corpus of web-crawled documents in various domains, made

[4] `https://huggingface.co/MLRS/BERTu`
[5] `https://huggingface.co/MLRS/mBERTu`
[6] `https://huggingface.co/datasets/MLRS/korpus_malti`

available on the Maltese Language Resource Server.[7]

For Basque, a couple of models have been released in the past few years. IXAmBERT (Otegi et al., 2020)[8] is a multilingual model for Basque, Spanish and English, trained on Wikipedia corpora and web-crawled news articles in Basque. BERTeus (Agerri et al., 2020)[9] is a monolingual model for Basque, trained also on Wikipedia and news articles and shown to perform better than standard mBERT in a variety of downstream tasks in Basque. Finally, RoBERTa-eus (Artetxe et al., 2022)[10] is a more recent family of RoBERTa-based language models for Basque, trained on the EusCrawl corpus[11], which consists of texts crawled from websites in Basque with high-quality content, as well as on the Basque portions of the mC4 and CC100 datasets.

Moreover, Basque and Spanish are both included in the mBERT models, which are trained on Wikipedia content in the top 100 languages with the largest Wikipedias.[12] However, as Otegi et al. (2020) and Agerri et al. (2020) point out, Basque texts make up a much smaller fraction of the training data of mBERT than the major languages included, which means it may be "overshadowed" by other languages and under-perform for several reasons, such as the sub-word tokenization module failing to split Basque morphemes properly. All Basque-specific models we mentioned were tested on different downstream tasks in Basque and performed better than mBERT. Therefore, as we already have a relatively small dataset and would rather benefit from encoding our sentences as reliably as possible, we decided to use one of the Basque-specific models.

We trained our COMET models for Spanish–Basque on top of IXAmBERT, the only one that includes both Spanish and Basque, as the other LMs we mentioned above (BERTeus and RoBERTa-eus) are monolingual Basque models; in any case, it would not be possible to use RoBERTa-eus because COMET does not support RoBERTa-based encoders. For English–Maltese, we used the multilingual mBERTu model.

[7] `https://mlrs.research.um.edu.mt`

[8] `https://huggingface.co/ixa-ehu/ixambert-base-cased`

[9] `https://huggingface.co/ixa-ehu/berteus-base-cased`

[10] `https://huggingface.co/ixa-ehu/roberta-eus-cc100-base-cased`

[11] `https://www.ixa.eus/euscrawl`

[12] `https://github.com/google-research/bert/blob/master/multilingual.md`, last accessed 27/09/2023.

5.3.4 | Results and Discussion

Our new models for each language pair, COMET-DA and COMET-22-FT, were trained using the same train and validation sets, and then evaluated on the test set of 100 samples. The train sets for English–Maltese and Spanish–Basque contained 428 and 435 samples respectively.

We then computed the correlation coefficients between the quality scores generated by these models and the human scores in the test set. In Table 5.6 we report Kendall's Tau (τ), Spearman (ρ) and Pearson (r) correlations. We decided to report Pearson correlation here as well for redundancy; Kendall's Tau and Spearman are both rank correlations, so they are more robust to outliers, as values are only represented by their ranks, but Pearson considers the magnitude of difference between scores, and thus, it should be more stable for segments/systems of similar quality (Kocmi et al., 2021; Mathur et al., 2020a; Osborne et al., 2022).

The correlation scores for all models can be seen in Table 5.6.

	en-mt			es-eu		
Model	τ	ρ	r	τ	ρ	r
COMET-22	0.292	0.421	0.399	0.223	0.326	0.214
COMET-DA	0.375	0.527	**0.527**	0.119	0.172	0.169
COMET-22-FT	**0.391**	**0.542**	0.525	**0.245**	**0.354**	**0.242**

Table 5.6: Kendall's Tau (τ), Spearman (ρ) and Pearson (r) correlation scores for 3 COMET models, evaluated on our test set. Scores in red were deemed statistically insignificant (p-values < 0.05).

The fine-tuned models (COMET-22-FT) resulted in the highest correlation scores for both language pairs, with an improvement of 0.10–0.13 in all correlation coefficients over COMET-22 for English-Maltese. For Spanish-Basque, the differences are smaller, but still notable, given a training set of only 435 samples.

The models trained from scratch (COMET-DA) performed very differently across our language pairs. The one for English–Maltese performs better than COMET-22 on our test set. Unfortunately, the same did not hold for Spanish-Basque; COMET-DA obtained low, statistically insignificant correlations on the `es-eu` test set.

Upon closer investigation, we found that all of these COMET models—the pretrained COMET-22 as well as our new models—only produced scores in narrow ranges: the scatter plots in Figure 5.5 show that, although the values in the test set—the human scores—are distributed across the y-axis (albeit unevenly), automatic scores are almost all above 0.5, mainly between 0.6 and 0.8.

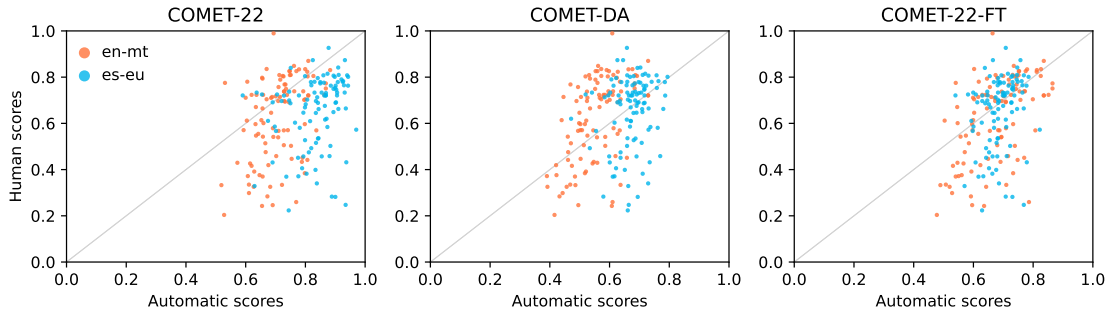

Figure 5.5: Quality scores produced by COMET-22 and our new models (x-axis), and the corresponding human scores in the test set (y-axis).

When it comes to our new models, COMET-DA and COMET-22-FT, we can hypothesize that the distribution of results is heavily influenced by the distribution of the training data. Figure 5.6 shows the density plots of the scores in the training sets, the test sets, and the model results side by side. The training scores for `es-eu` peak between 0.6–0.8, and the models almost entirely produce scores within this same range; for `en-mt`, where the training data is a little more balanced, the resulting scores are also in a wider range, around 0.5–0.8.

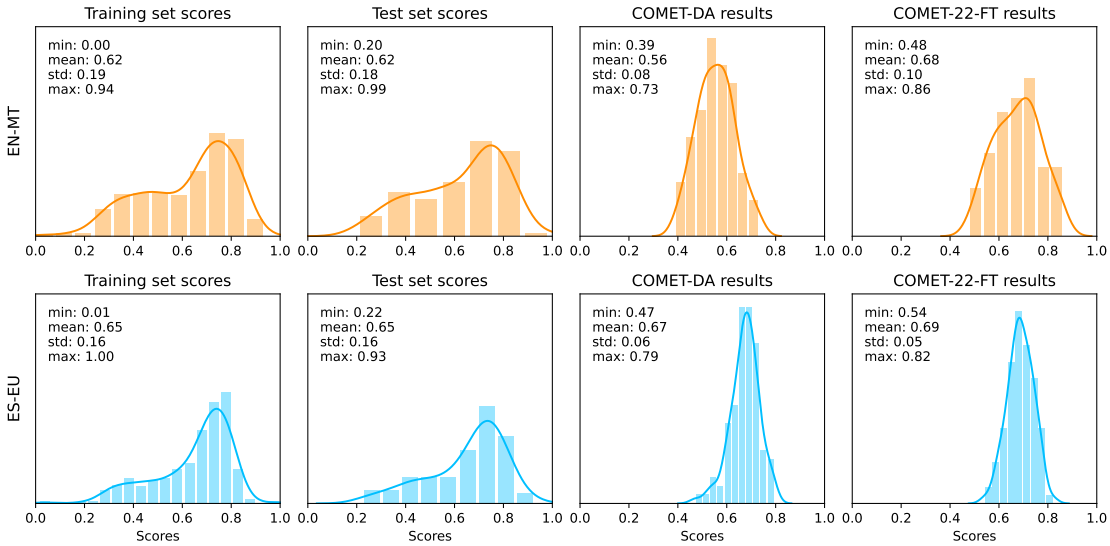

Figure 5.6: Distributions of scores in each training set, test set, and the results from our COMET-DA and COMET-22-FT models.

The fact that COMET-22 behaves similarly, producing scores concentrated between 0.6–0.8 on our test set despite not having seen any of our data at training time, led us to look closely at its training data as well. The scores in its dataset are

only available as unbound z-scores, before rescaling, so we simulate how they would have been rescaled before training, following the process described by COMET developers: applying min-max scaling to [r_{min}, r_{max}] and clipping the results to [0,1]. We also apply the alternative rescaling process as we did for our data (see Sec. 5.3.1), by min-max scaling directly to [0,1], for comparison. We show the 3 distributions in Fig. 5.7 (z-scores before and after both rescaling processes).

The z-scores in the COMET-22 training set are largely concentrated within [-2.0, 2.0], but outliers bring r_{min} down to -1.9, and then all z-scores rescaled to [-1.9, 0], roughly 38% of the full training set, are clipped to 0, yielding a very unbalanced distribution of rescaled scores that will be used for training (the second graph in Fig. 5.7). It appears to be the same issue we saw with our dataset, but in the scale of a hundred thousand scores. Therefore, if COMET-22 has decent correlation scores when tested on our data for languages it technically does not support, it might be out of sheer "luck", as our test data is also mostly within the range of scores the model is most likely to produce.

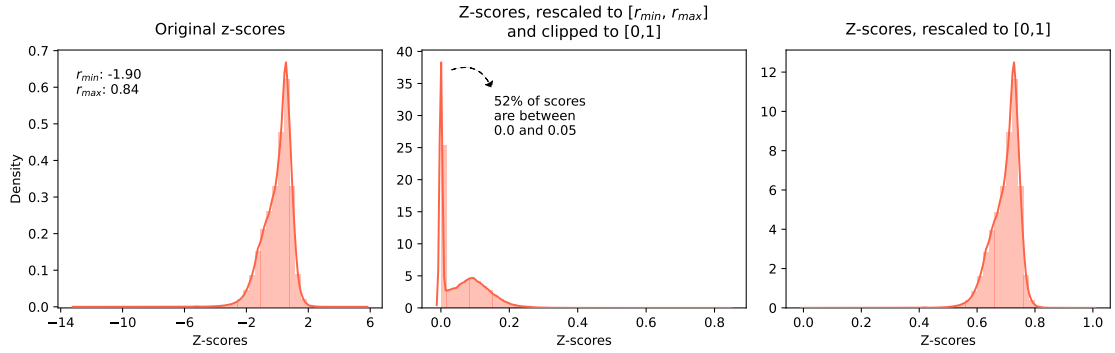

Figure 5.7: Distribution of z-scores (original and rescaled) in the training data of COMET-22.

In order to investigate this possibility, we made "adversarial" datasets to test only COMET-22 again: we randomly sampled 100 segments, exclusively with scores ≤ 0.6, from each of our train sets. The idea is that these MT outputs have been judged as "below average" by participants of our evaluation campaign, and the z-scores lie outside the range where the training data of COMET-22 is concentrated.

We report the correlation scores in Table 5.7, and also plot the quality scores produced by COMET-22 against the human scores in the test set in Fig. 5.8. Results show that the performance of COMET-22 drops significantly in comparison to the regular test set, and all the correlations on the adversarial test set are

statistically insignificant. Based on this test, we suggest that, in the case of our language pairs, the performance of COMET-22 is unstable, and might be influenced by the unbalanced distribution of scores in its training data. This is also in line with our findings from the system-level evaluation (Sec. 5.2), where COMET-22 overestimated the performance of the worst systems, in comparison with human judgements.

		en-mt			es-eu		
Model	**Test set**	τ	ρ	r	τ	ρ	r
COMET-22	Regular	0.292	0.421	0.399	0.223	0.326	0.214
COMET-22	Adversarial	0.099	0.137	0.077	-0.010	-0.011	0.140

Table 5.7: Correlation scores for COMET-22 on the adversarial test set of 100 segments with z-scores ≤ 0.6. Results on the regular test set are reproduced from Table 5.6 for comparison.

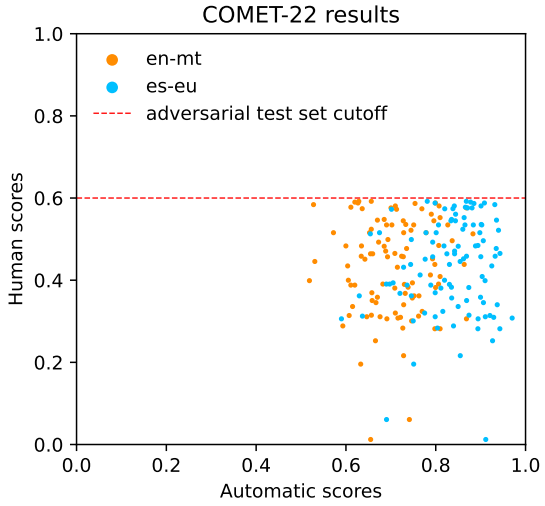

Figure 5.8: Quality scores produced by COMET-22 for the adversarial test set (x-axis), and the corresponding human scores (y-axis).

Therefore, the most important lesson we take from these experiments—and also from the ones described in Chapter 3—is that COMET can be highly susceptible to the distribution of scores present in its training set, and as a potential solution, we believe that new COMET models should ideally be trained on balanced data across its full range of scores [0.0–1.0]. We do not have this kind of data available for our language pairs yet, but it would be highly interesting to perform experiments with balanced datasets for unsupported languages in the future. It might even be better to train on less data, but balanced, rather than larger but very unbalanced datasets.

6

Conclusions

Translation quality, being a subjective concept, is hard to define and thus hard to measure. As Machine Translation evolved from rule-based systems to huge neural networks, and the demand for quick evaluation methods only grew, the field turned away from human evaluations and towards automatic metrics, mostly based on superficial lexical matching, but researchers quickly realized that these simple metrics were an inadequate substitute. In the search for better methods, trainable metrics have made an impact in the field in recent years, as they directly predict human judgements of translation quality and have been shown to evaluate translations more accurately.

In this dissertation, we have discussed the advantages and limitations of the main types of evaluation methods for MT, and placed our focus on COMET, a trainable evaluation system proposed in 2020 that has been gaining traction in MT research and development. We specifically discussed concerns about its limited language support, and what it means for under-resourced languages—which are not supported by large cross-lingual encoders, and do not have large quantities of annotated data available.

To investigate COMET's performance on unseen language pairs, we started with some preliminary experiments with COMET-22 using WMT Direct Assessment datasets that had not been included in its training. We found that COMET-22 performed quite poorly on 3 out of 4 language pairs, even though all languages were supported by XLM-R. We also explored the possibility of improving this performance by fine-tuning the model on more DA data, and it led to minor improvements in some of the cases.

For our case study on English–Maltese and Spanish–Basque, as annotated data was not yet available, we set out to collect Direct Assessments from bilingual

speakers through a crowd-based evaluation campaign. We were able to obtain evaluations for a portion of our dataset of translations from 3 MT systems, and with these scores, we performed a small-scale meta-evaluation study. We compared the performance of COMET against a few lexical metrics, and also produced new COMET models by fine-tuning COMET-22 on our data and training custom COMET-DA models from scratch, leveraging specified cross-lingual encoders for our languages.

We believe that this analysis, although based only on a small case study, has been a step in the right direction, and it has allowed us to pinpoint some potential gaps in the structure of COMET that may hinder its generalization capabilities. In our tests, it appeared that the scores produced by the models were influenced by the unbalanced distributions of scores in their training datasets. This is a complex issue to diagnose and pinpoint the root cause; a lot more experiments would be needed, but our observations can ultimately aid in the development of more robust models, especially in a low-resource scenario, where the quality of the data matters even more.

6.1 | Achieved Aims and Objectives

Looking back on the initial aims and objectives of the present work, we believe that we have been able to achieve what we set out to accomplish. It is not possible to provide one answer to the question of whether COMET can generalize to unseen and unsupported languages; it can only be determined on a case-by-case basis. However, we have performed experiments with 6 language pairs, with varying degrees of success, and have explored a number of factors that may have influenced our results one way or another. Additionally, we have provided further evidence of the potential of fine-tuning as a means to improve the performance of COMET for certain language pairs; as opposed to training whole new models from scratch, this approach leverages the existing knowledge of pre-trained models, while providing them with new examples in these specific languages.

Through this work, we also intended to shed light on how immensely complicated it is to evaluate translations by means of human judgements *and* by automatic metrics. The process of devising and executing our experiments has involved a multitude of decisions: the human evaluation method and its specific layout and prompting question, what scale of scores to use, who could participate, how to perform quality control, how to rescale the data, how to sample the data

to split it into sets, how to measure the quality of the results, how to interpret the results. Everything depends on a number of factors and we have tried as best as possible to take into consideration the findings from well-established literature on MT evaluation, while also adapting a lot of it to our particular circumstances.

6.2 | Critique and Limitations

The most prominent limitation in our work is the scale of our evaluation campaign; we were only able to obtain evaluations for a part of our segments, which left us with quite a small dataset to evaluate and train COMET models for each language pair. Therefore, we reinforce that our results are preliminary, and experiments would have to be reproduced with larger amounts of data to yield reliable empirical evidence of what we have investigated.

With regards to the evaluation campaign, we were also limited in terms of who could participate. It would have been ideal to hire professional translators to conduct a more reliable human evaluation campaign, and to do it by means of error analysis, in order to obtain more fine-grained insights about specific translation errors in each hypothesis. We chose a different approach, by counting on the voluntary contributions of bilingual speakers, and adapted our campaign design to better fit this scenario.

Lastly, we clarify that we have conducted a *case study* by choosing English-Maltese and Spanish-Basque as our focus, and the preliminary conclusions from this work do not necessarily generalize to any other under-resourced language pairs, nor to other languages unsupported by the current COMET models. In fact, we have conducted the experiments in parallel, with no interference between data for different language pairs, and while we have kept the conditions as similar as possible, this was done more for the sake of consistency than comparability, as our intent has always been to compare *methods*, not languages. In fact, by focusing on two languages that are unrelated, as opposed to different studies which tackle closely related languages, our aim was to diversify our potential results and eliminate any odds that the methods might yield similar results for both language pairs due to confounding factors.

6.3 | Future Work

Throughout this dissertation, we have examined numerous decisions that have to be made along every step of a meta-evaluation study—decisions which highly affect the results. Moreover, translation evaluation is a highly subjective and complex topic in the first place. Therefore, we see a plethora of directions in which we could continue to evolve this research, simply by experimenting with different parameters to explore the outcomes.

We briefly introduce a few ideas below:

- Based on our conclusions from the last chapter, it would be highly interesting to experiment with training new COMET models on balanced datasets—containing balanced amounts of evaluation scores. These models could be evaluated in comparison to pre-trained COMET models, in order to investigate how many of COMET's accurate predictions happen "by chance", due to its uneven distribution of scores in training.
- Exploring the different architectures that the COMET framework provides could also be interesting, for future works that wish to delve deeper into its capabilities; we focused on COMET-DA, but one could also convert DA scores to DARR and evaluate a COMET-RANK model, or to delve into Quality Estimation (QE) and evaluate the performance of COMET-QE and/or COMETKIWI models.
- Conducting an error analysis of MT systems for English–Maltese and Spanish–Basque. Error analysis frameworks such as MQM allow for a very fine-grained study of common translation errors between these languages, and could help guide future research on MT approaches to improve these shortcomings.
- Including other trainable metrics in a larger meta-evaluation study, so that they could be compared against COMET in a scenario of evaluation of unsupported languages.

6.4 | Final Remarks

To the best of our knowledge, this project has been the first meta-evaluation analysis of machine translation focused on Maltese and Basque, and one of few research works that have obtained human evaluations of translations to or from these languages. We believe that machine translation is an essential application

to ensure the survival of minoritary languages in the digital age, and we hope that our work has brought us a step closer to the development of better evaluation methods for under-resourced languages, so that they will not be left behind as machine translation evaluation moves forward.

References

A. Görög. Quality evaluation today: the dynamic quality framework. In *Proceedings of Translating and the Computer 36*, London, UK, November 27-28 2014. AsLing. URL https://aclanthology.org/2014.tc-1.21.

Ahmed Abdelali, Francisco Guzman, Hassan Sajjad, and Stephan Vogel. The AMARA corpus: Building parallel language resources for the educational domain. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1856–1862, Reykjavik, Iceland, May 2014. European Language Resources Association (ELRA). URL http://www.lrec-conf.org/proceedings/lrec2014/pdf/877_Paper.pdf.

Kurt Abela. Machine translation system for low resource languages. Master's thesis, University of Malta, Malta, 2023.

Milind Agarwal, Sweta Agrawal, Antonios Anastasopoulos, Luisa Bentivogli, Ondřej Bojar, Claudia Borg, Marine Carpuat, Roldano Cattoni, Mauro Cettolo, Mingda Chen, William Chen, Khalid Choukri, Alexandra Chronopoulou, Anna Currey, Thierry Declerck, Qianqian Dong, Kevin Duh, Yannick Estève, Marcello Federico, Souhir Gahbiche, Barry Haddow, Benjamin Hsu, Phu Mon Htut, Hirofumi Inaguma, Dávid Javorský, John Judge, Yasumasa Kano, Tom Ko, Rishu Kumar, Pengwei Li, Xutai Ma, Prashant Mathur, Evgeny Matusov, Paul McNamee, John P. McCrae, Kenton Murray, Maria Nadejde, Satoshi Nakamura, Matteo Negri, Ha Nguyen, Jan Niehues, Xing Niu, Atul Kr. Ojha, John E. Ortega, Proyag Pal, Juan Pino, Lonneke van der Plas, Peter Polák, Elijah Rippeth, Elizabeth Salesky, Jiatong Shi, Matthias Sperber, Sebastian Stüker, Katsuhito Sudoh, Yun Tang, Brian Thompson, Kevin Tran, Marco Turchi, Alex Waibel, Mingxuan Wang, Shinji Watanabe, and Rodolfo Zevallos. FINDINGS OF THE IWSLT 2023 EVALUATION CAMPAIGN. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, pages 1–61, Toronto, Canada (in-person and online), July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.iwslt-1.1. URL https://aclanthology.org/2023.iwslt-1.1.

Rodrigo Agerri, Iñaki San Vicente, Jon Ander Campos, Ander Barrena, Xabier Saralegi, Aitor Soroa, and Eneko Agirre. Give your text representation models some love: the case for Basque. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4781–4788, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.588.

Dionisius A. Agius. A Semitic Maltese inventory with a possible Siculo-Arabic intervention. *Zeitschrift für Arabische Linguistik*, (6):7–15, 1981. ISSN 0170026X.

Chantal Amrhein, Nikita Moghe, and Liane Guillou. 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), December 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.wmt-1.44.

Nora Aranberri, Gorka Labaka, Arantza Díaz De Ilarraza, and Kepa Sarasola. Ebaluatoia: Crowd evaluation for English—Basque machine translation. *Lang. Resour. Eval.*, 51(4):1053–1084, dec 2017. ISSN 1574-020X. doi: 10.1007/s10579-016-9335-x.

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. Unsupervised neural machine translation. *CoRR*, abs/1710.11041, 2017.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. Unsupervised statistical machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3632–3642, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1399. URL https://aclanthology.org/D18-1399.

Mikel Artetxe, Itziar Aldabe, Rodrigo Agerri, Olatz Perez de Viñaspre, and Aitor Soroa. Does corpus quality really matter for low-resource languages?, 2022.

Mikko Aulamo, Umut Sulubacak, Sami Virpioja, and Jörg Tiedemann. OpusTools and parallel corpus diagnostics. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 3782–3789. European Language Resources Association, May 2020. ISBN 979-10-95546-34-4. URL https://www.aclweb.org/anthology/2020.lrec-1.467.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate, 2014. URL https://arxiv.org/abs/1409.0473.

Rafael E. Banchs, Luis F. D'Haro, and Haizhou Li. Adequacy–fluency metrics: Evaluating MT in the continuous space model framework. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(3):472–482, 2015. doi: 10.1109/TASLP.2015.2405751.

Satanjeev Banerjee and Alon Lavie. 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, June 2005. Association for Computational Linguistics. URL https://aclanthology.org/W05-0909.

Ankur Bapna, Isaac Caswell, Julia Kreutzer, Orhan Firat, Daan van Esch, Aditya Siddhant, Mengmeng Niu, Pallavi Baljekar, Xavier Garcia, Wolfgang Macherey, Theresa Breiner, Vera Axelrod, Jason Riesa, Yuan Cao, Mia Xu Chen, Klaus Macherey, Maxim Krikun, Pidong Wang, Alexander Gutkin, Apurva Shah, Yanping Huang, Zhifeng Chen, Yonghui Wu, and Macduff Hughes. Building machine translation systems for the next thousand languages, 2022.

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. 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, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-5301. URL https://aclanthology.org/W19-5301.

Evarist Bartolo. Our endangered language. Times of Malta, July 2022. URL https://timesofmalta.com/articles/view/endangered-language-evarist-bartolo.969547.

Martin Benjamin. How GT pivots through English, 2019. URL https://www.teachyoubackwards.com/extras/pivot/.

Luisa Bentivogli, Mauro Cettolo, Marcello Federico, and Christian Federmann. Machine translation human evaluation: an investigation of evaluation based on post-editing and its relation with direct assessment. In *Proceedings of the 15th International Conference on Spoken Language Translation*, pages 62–69, Brussels, October 29-30 2018. International Conference on Spoken Language Translation. URL https://aclanthology.org/2018.iwslt-1.9.

Alexandre Berard, Ioan Calapodescu, and Claude Roux. Naver labs Europe's systems for the WMT19 machine translation robustness task. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 526–532, Florence, Italy, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-5361. URL https://aclanthology.org/W19-5361.

Arianna Bisazza, Ahmet Üstün, and Stephan Sportel. On the Difficulty of Translating Free-Order Case-Marking Languages. *Transactions of the Association for Computational Linguistics*, 9:1233–1248, 11 2021. ISSN 2307-387X. doi: 10.1162/tacl_a_00424.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Barry Haddow, Matthias Huck, Chris Hokamp, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Carolina Scarton, Lucia Specia, and Marco Turchi. Findings of the 2015 workshop on statistical machine translation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 1–46, Lisbon, Portugal, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/W15-3001. URL https://aclanthology.org/W15-3001.

Ondrej Bojar, Christian Federmann, Barry Haddow, Philipp Koehn, Matt Post, and Lucia Specia. Ten years of WMT evaluation campaigns: Lessons learnt. In *Proceedings of the LREC 2016 Workshop "Translation Evaluation–From Fragmented Tools and Data Sets to an Integrated Ecosystem*, pages 27–34, 2016a.

Ondřej Bojar, Yvette Graham, Amir Kamran, and Miloš Stanojević. 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, August 2016b. Association for Computational Linguistics. doi: 10.18653/v1/W16-2302. URL https://aclanthology.org/W16-2302.

Ondřej Bojar, Yvette Graham, and Amir Kamran. Results of the WMT17 metrics shared task. In *Proceedings of the Second Conference on Machine Translation*, pages 489–513, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-4755. URL https://aclanthology.org/W17-4755.

Ondřej Bojar, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Philipp Koehn, and Christof Monz. Findings of the 2018 Conference on Machine Translation (WMT18). In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 272–303, Belgium, Brussels, October 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-6401. URL https://aclanthology.org/W18-6401.

Joseph M. Brincat. Maltese – an unusual formula. MED Magazine, issue 27, February 2005. URL http://macmillandictionaries.com/MED-Magazine/February2005/27-LI-Maltese.htm.

P. Brown, J. Cocke, S. Della Pietra, V. Della Pietra, F. Jelinek, R. Mercer, and P. Roossin. A statistical approach to language translation. In *Proceedings of the 12th Conference on Computational Linguistics - Volume 1*, COLING '88, page 71–76, USA, 1988. Association for Computational Linguistics. ISBN 963 8431 56 3. doi: 10.3115/991635.991651. URL https://doi.org/10.3115/991635.991651.

Aljoscha Burchardt. Multidimensional quality metrics: a flexible system for assessing translation quality. In *Proceedings of Translating and the Computer 35*, London, UK, November 28-29 2013. Aslib. URL https://aclanthology.org/2013.tc-1.6.

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. Re-evaluating the role of BLEU in machine translation research. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 249–256, Trento, Italy, April 2006. Association for Computational Linguistics. URL https://aclanthology.org/E06-1032.

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. (Meta-) evaluation of machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 136–158, Prague, Czech Republic, June 2007. Association for Computational Linguistics. URL https://aclanthology.org/W07-0718.

John B. Carroll. An experiment in evaluating the quality of translations. *Mech. Transl. Comput. Linguistics*, 9:55–66, 1966.

Isaac Caswell and Ankur Bapna. Unlocking zero-resource machine translation to support new languages in google translate, 2022. URL https://blog.research.google/2022/05/24-new-languages-google-translate.html.

Isaac Caswell and Bowen Liang. Recent advances in google translate, 2020. URL https://blog.research.google/2020/06/recent-advances-in-google-translate.html.

Niladri Chatterjee, Anish Johnson, and Madhav Krishna. Some improvements over the BLEU metric for measuring translation quality for Hindi. In *2007 International Conference on Computing: Theory and Applications (ICCTA'07)*, pages 485–490, 2007. doi: 10.1109/ICCTA.2007.120.

Eirini Chatzikoumi. How to evaluate machine translation: A review of automated and human metrics. *Natural Language Engineering*, 26(2):137–161, September 2019. doi: 10.1017/s1351324919000469.

Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Si Wei, Hui Jiang, and Diana Inkpen. Enhanced LSTM for natural language inference. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1657–1668, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1152. URL https://aclanthology.org/P17-1152.

Xiaoyu Chen, Daimeng Wei, Hengchao Shang, Zongyao Li, Zhanglin Wu, Zhengzhe Yu, Ting Zhu, Mengli Zhu, Ning Xie, Lizhi Lei, Shimin Tao, Hao Yang, and Ying Qin. Exploring robustness of machine translation metrics: A study of twenty-two automatic metrics in the WMT22 metric task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 530–540, Abu Dhabi, United Arab Emirates (Hybrid), December 2022. Association for Computational Linguistics. URL `https://aclanthology.org/2022.wmt-1.46`.

Hyung Won Chung, Thibault Févry, Henry Tsai, Melvin Johnson, and Sebastian Ruder. Rethinking embedding coupling in pre-trained language models. *CoRR*, abs/2010.12821, 2020.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116, 2019.

Nicola Crichton. Methodological issues in clinical research. *J. Clin. Nurs.*, 10:707–715, 01 2001.

Michael Denkowski and Alon Lavie. Choosing the right evaluation for machine translation: an examination of annotator and automatic metric performance on human judgment tasks. In *Proceedings of the 9th Conference of the Association for Machine Translation in the Americas: Research Papers*, Denver, Colorado, USA, October 31-November 4 2010. Association for Machine Translation in the Americas. URL `https://aclanthology.org/2010.amta-papers.20`.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018.

George Doddington. 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, 2002. Morgan Kaufmann Publishers Inc.

Yadolah Dodge. *Spearman Rank Correlation Coefficient*, pages 502–505. Springer New York, New York, NY, 2008. ISBN 978-0-387-32833-1. doi: 10.1007/978-0-387-32833-1_379. URL `https://doi.org/10.1007/978-0-387-32833-1_379`.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig. Ethnologue: Languages of the world (26th edition), 2023. URL `https://www.ethnologue.com`.

Sergey Edunov, Myle Ott, Marc'Aurelio Ranzato, and Michael Auli. On the evaluation of machine translation systems trained with back-translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2836–2846, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.253. URL `https://aclanthology.org/2020.acl-main.253`.

Marcello Federico, Matteo Negri, Luisa Bentivogli, and Marco Turchi. Assessing the impact of translation errors on machine translation quality with mixed-effects models. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1643–1653, Doha, Qatar, October 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1172. URL `https://aclanthology.org/D14-1172`.

Christian Federmann. 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, August 2018. Association for Computational Linguistics. URL https://aclanthology.org/C18-2019.

David Freedman, Robert Pisani, and Roger Purves. Statistics (international student edition). *Pisani, R. Purves, 4th edn. WW Norton & Company, New York*, 2007.

Markus Freitag, Isaac Caswell, and Scott Roy. APE at scale and its implications on MT evaluation biases. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 34–44, Florence, Italy, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-5204. URL https://aclanthology.org/W19-5204.

Markus Freitag, David Grangier, and Isaac Caswell. 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, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.5. URL https://aclanthology.org/2020.emnlp-main.5.

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

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, George Foster, Alon Lavie, and Ondřej Bojar. Results of the WMT21 metrics shared task: Evaluating metrics with expert-based human evaluations on TED and news domain. In *Proceedings of the Sixth Conference on Machine Translation*, pages 733–774, Online, November 2021b. Association for Computational Linguistics. URL https://aclanthology.org/2021.wmt-1.73.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André FT Martins. 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, 2022.

Paul L. Garvin. The fulcrum approach – twelve years later. International Forum on Information and Documentation 5 (2), pp. 27-29, 1980. URL https://aclanthology.org/www.mt-archive.info/IntForumInfDoc-1980-Garvin.pdf.

Attila Görög. Quantifying and benchmarking quality: the TAUS dynamic quality framework. *Tradumàtica*, (12):0443–454, 2014.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. The FLORES-101 evaluation benchmark for low-resource and multilingual machine translation. *CoRR*, abs/2106.03193, 2021.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 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, August 2013. Association for Computational Linguistics. URL https://aclanthology.org/W13-2305.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. Can machine translation systems be evaluated by the crowd alone. *Natural Language Engineering*, 23(1):3–30, 2017. doi: 10.1017/S1351324915000339.

Alex Graves, Abdel-rahman Mohamed, and Geoffrey E. Hinton. Speech recognition with deep recurrent neural networks. *CoRR*, abs/1303.5778, 2013.

Harald Hammarström, Robert Forkel, Martin Haspelmath, and Sebastian Bank. glottolog/glottolog: Glottolog database 4.8, 2023. URL https://zenodo.org/record/8131084.

John Hutchins. ALPAC: the (in)famous report. MT News International, no. 14, pp 9-12, June 1996. URL https://aclanthology.org/www.mt-archive.info/90/MTNI-1996-Hutchins.pdf.

John Hutchins. The first public demonstration of machine translation : the georgetown-ibm system, 7 th january 1954. 2006a. URL https://api.semanticscholar.org/CorpusID:132677.

John Hutchins. The history of machine translation in a nutshell. 2006b. URL https://api.semanticscholar.org/CorpusID:231780839.

W. J. Hutchins. *Machine Translation: Past, Present, Future*. John Wiley & Sons, Inc., USA, 1986. ISBN 0470203137.

W. John Hutchins. The Georgetown-IBM experiment demonstrated in January 1954. In *Proceedings of the 6th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 102–114, Washington, USA, September 28 - October 2 2004. Springer. URL https://link.springer.com/chapter/10.1007/978-3-540-30194-3_12.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.560. URL https://aclanthology.org/2020.acl-main.560.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-4020. URL https://aclanthology.org/P18-4020.

Nal Kalchbrenner and Phil Blunsom. Recurrent continuous translation models. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1700–1709, Seattle, Washington, USA, October 2013. Association for Computational Linguistics. URL https://aclanthology.org/D13-1176.

M. G. Kendall. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93, 1938. ISSN 00063444.

M. G. Kendall. The treatment of ties in ranking problems. *Biometrika*, 33(3):239–251, 1945. doi: 10.1093/biomet/33.3.239.

M.G. Kendall. *Rank Correlation Methods.* Theory and applications of rank order-statistics. Griffin, 1970. ISBN 9780852641996. URL https://books.google.com.mt/books?id=Mm2jjgEACAAJ.

Fabio Kepler, Jonay Trénous, Marcos Treviso, Miguel Vera, and André F. T. Martins. 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, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-3020. URL https://aclanthology.org/P19-3020.

Hyun Kim, Jong-Hyeok Lee, and Seung-Hoon Na. Predictor-estimator using multilevel task learning with stack propagation for neural quality estimation. In *Proceedings of the Second Conference on Machine Translation*, pages 562–568, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-4763. URL https://aclanthology.org/W17-4763.

Tom Kocmi, Christian Federmann, Roman Grundkiewicz, Marcin Junczys-Dowmunt, Hitokazu Matsushita, and Arul Menezes. To ship or not to ship: An extensive evaluation of automatic metrics for machine translation. *CoRR*, abs/2107.10821, 2021.

Tom Kocmi, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Thamme Gowda, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Rebecca Knowles, Philipp Koehn, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Michal Novák, Martin Popel, and Maja Popović. Findings of the 2022 conference on machine translation (WMT22). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 1–45, Abu Dhabi, United Arab Emirates (Hybrid), December 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.wmt-1.1.

Philipp Koehn. *Statistical Machine Translation.* Cambridge University Press, December 2009. doi: 10.1017/cbo9780511815829. URL https://doi.org/10.1017/cbo9780511815829.

Philipp Koehn and Rebecca Knowles. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver, August 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-3204. URL https://aclanthology.org/W17-3204.

Philipp Koehn and Christof Monz. Manual and automatic evaluation of machine translation between European languages. In *Proceedings on the Workshop on Statistical Machine Translation*, pages 102–121, New York City, June 2006. Association for Computational Linguistics. URL https://aclanthology.org/W06-3114.

David Kolovratník, Natalia Klyueva, and Ondřej Bojar. Statistical machine translation between related and unrelated languages. volume 584, pages 31–36, 01 2009.

Julia Kreutzer, Shigehiko Schamoni, and Stefan Riezler. QUality estimation from ScraTCH (QUETCH): Deep learning for word-level translation quality estimation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 316–322, Lisbon, Portugal, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/W15-3037. URL https://aclanthology.org/W15-3037.

E. Kreyszig. *Advanced Engineering Mathematics.* John Wiley & Sons Inc, 4 edition, 1979. ISBN 0471021407. URL https://books.google.de/books?id=UnN8DpXI74EC.

Gorka Labaka. *EUSMT: incorporating linguistic information to SMT for a morphologically rich language. Its use in SMT-RBMT-EBMT hybridation.* PhD thesis, University of the Basque Country, 2010.

Gorka Labaka, Cristina España-Bonet, Lluís Màrquez, and Kepa Sarasola. A hybrid machine translation architecture guided by syntax. *Machine Translation*, 28(2):91–125, September 2014. doi: 10.1007/s10590-014-9153-0.

Guillaume Lample, Ludovic Denoyer, and Marc'Aurelio Ranzato. Unsupervised machine translation using monolingual corpora only. *CoRR*, abs/1711.00043, 2017.

Samuel Läubli, Sheila Castilho, Graham Neubig, Rico Sennrich, Qinlan Shen, and Antonio Toral. A set of recommendations for assessing human–machine parity in language translation. *Journal of Artificial Intelligence Research*, 67, March 2020. doi: 10.1613/jair.1.11371.

Alon Lavie. Why we built COMET, a new framework and metric for automated machine translation evaluation, 2020. URL `https://resources.unbabel.com/blog/why-we-believe-high-quality-machine-translation-really-is-possible`.

LDC. Linguistic data annotation specification: Assessment of fluency and adequacy in translations. Technical report. Revision 1.5., 2005.

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

Pierre Lison and Jörg Tiedemann. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 923–929, Portorož, Slovenia, May 2016. European Language Resources Association (ELRA). URL `https://aclanthology.org/L16-1147`.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019.

Chi-kiu Lo. 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, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-5358. URL `https://aclanthology.org/W19-5358`.

Lieve Macken, Bram Vanroy, and ArdaTezcan. Adapting machine translation education to the neural era: a case study of MT quality assessment. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 305–314. European Association for Machine Translation (EAMT), 2023. ISBN 9789520329471. URL `https://events.tuni.fi/eamt23/proceedings/`.

Benjamin Marie, Atsushi Fujita, and Raphael Rubino. Scientific credibility of machine translation research: A meta-evaluation of 769 papers. 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 7297–7306, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.566. URL `https://aclanthology.org/2021.acl-long.566`.

André F. T. Martins, Ramón Astudillo, Chris Hokamp, and Fabio Kepler. Unbabel's participation in the WMT16 word-level translation quality estimation shared task. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 806–811, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/W16-2387. URL https://aclanthology.org/W16-2387.

André F. T. Martins, Marcin Junczys-Dowmunt, Fabio N. Kepler, Ramón Astudillo, Chris Hokamp, and Roman Grundkiewicz. Pushing the limits of translation quality estimation. *Transactions of the Association for Computational Linguistics*, 5:205–218, 2017. doi: 10.1162/tacl_a_00056.

Nitika Mathur, Timothy Baldwin, and Trevor Cohn. Putting evaluation in context: Contextual embeddings improve machine translation evaluation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2799–2808, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1269. URL https://aclanthology.org/P19-1269.

Nitika Mathur, Timothy Baldwin, and Trevor Cohn. 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, July 2020a. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.448. URL https://aclanthology.org/2020.acl-main.448.

Nitika Mathur, Johnny Wei, Markus Freitag, Qingsong Ma, and Ondřej Bojar. Results of the WMT20 metrics shared task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 688–725, Online, November 2020b. Association for Computational Linguistics. URL https://aclanthology.org/2020.wmt-1.77.

I. Dan Melamed, Ryan Green, and Joseph P. Turian. Precision and recall of machine translation. In *Companion Volume of the Proceedings of HLT-NAACL 2003 - Short Papers*, pages 61–63, 2003. URL https://aclanthology.org/N03-2021.

Kurt Micallef, Albert Gatt, Marc Tanti, Lonneke van der Plas, and Claudia Borg. Pre-training data quality and quantity for a low-resource language: New corpus and BERT models for Maltese. In *Proceedings of the Third Workshop on Deep Learning for Low-Resource Natural Language Processing*, pages 90–101, Hybrid, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.deeplo-1.10. URL https://aclanthology.org/2022.deeplo-1.10.

Ananya Mukherjee and Manish Shrivastava. REUSE: REference-free UnSupervised quality estimation metric. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 564–568, Abu Dhabi, United Arab Emirates (Hybrid), December 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.wmt-1.50.

Makoto Nagao. A framework of a mechanical translation between Japanese and English by analogy principle. In *Proc. of the International NATO Symposium on Artificial and Human Intelligence*, page 173–180, USA, 1984. Elsevier North-Holland, Inc. ISBN 0444865454.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. CrowS-pairs: A challenge dataset for measuring social biases in masked language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1953–1967, Online, November

2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.154. URL https://aclanthology.org/2020.emnlp-main.154.

Aurélie Névéol, Yoann Dupont, Julien Bezançon, and Karën Fort. French CrowS-pairs: Extending a challenge dataset for measuring social bias in masked language models to a language other than English. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8521–8531, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.583. URL https://aclanthology.org/2022.acl-long.583.

Philip Osborne, Heido Nõmm, and André Freitas. A survey of text games for reinforcement learning informed by natural language. *Transactions of the Association for Computational Linguistics*, 10: 873–887, 2022. doi: 10.1162/tacl_a_00495.

Arantxa Otegi, Aitor Agirre, Jon Ander Campos, Aitor Soroa, and Eneko Agirre. Conversational question answering in low resource scenarios: A dataset and case study for Basque. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 436–442, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.55.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics, 2002.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12: 2825–2830, 2011.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 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, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1202. URL https://aclanthology.org/N18-1202.

John R. Pierce and John B. Carroll. Language and machines: Computers in translation and linguistics. 1966. URL https://api.semanticscholar.org/CorpusID:60353620.

Thierry Poibeau. *Machine translation*. The MIT Press, 2017. ISBN 9780262534215.

Maja Popović. 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, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/W15-3049. URL https://aclanthology.org/W15-3049.

Maja Popović. chrF deconstructed: beta parameters and n-gram weights. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 499–504, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/W16-2341. URL https://aclanthology.org/W16-2341.

Maja Popović. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-4770. URL https://aclanthology.org/W17-4770.

Maja Popović. On reducing translation shifts in translations intended for MT evaluation. In *Proceedings of Machine Translation Summit XVII: Translator, Project and User Tracks*, pages 80–87, Dublin, Ireland, August 2019. European Association for Machine Translation. URL https://aclanthology.org/W19-6712.

Maja Popović. On nature and causes of observed MT errors. In *Proceedings of Machine Translation Summit XVIII: Research Track*, pages 163–175, Virtual, August 2021. Association for Machine Translation in the Americas. URL https://aclanthology.org/2021.mtsummit-research.14.

Maja Popović, Mihael Arčan, and Filip Klubička. Language related issues for machine translation between closely related South Slavic languages. In *Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial3)*, pages 43–52, Osaka, Japan, December 2016. The COLING 2016 Organizing Committee. URL https://aclanthology.org/W16-4806.

Amy Pu, Hyung Won Chung, Ankur Parikh, Sebastian Gehrmann, and Thibault Sellam. 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, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.58. URL https://aclanthology.org/2021.emnlp-main.58.

Ye Qi, Devendra Sachan, Matthieu Felix, Sarguna Padmanabhan, and Graham Neubig. When and why are pre-trained word embeddings useful for neural machine translation? In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 529–535, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-2084. URL https://aclanthology.org/N18-2084.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon 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, November 2020a. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.213. URL https://aclanthology.org/2020.emnlp-main.213.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon 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, November 2020b. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.213. URL https://aclanthology.org/2020.emnlp-main.213.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. Unbabel's participation in the WMT20 metrics shared task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 911–920, Online, November 2020c. Association for Computational Linguistics. URL https://aclanthology.org/2020.wmt-1.101.

Ricardo Rei, Ana C Farinha, Chrysoula Zerva, Daan van Stigt, Craig Stewart, Pedro Ramos, Taisiya Glushkova, André F. T. Martins, and Alon 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, November 2021a. Association for Computational Linguistics. URL https://aclanthology.org/2021.wmt-1.111.

Ricardo Rei, Ana C Farinha, Chrysoula Zerva, Daan van Stigt, Craig Stewart, Pedro Ramos, Taisiya Glushkova, André F. T. Martins, and Alon 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, November 2021b. Association for Computational Linguistics. URL https://aclanthology.org/2021.wmt-1.111.

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. 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), December 2022a. Association for Computational Linguistics. URL https://aclanthology.org/2022.wmt-1.52.

Ricardo Rei, Ana C Farinha, José G.C. de Souza, Pedro G. Ramos, André F.T. Martins, Luisa Coheur, and Alon Lavie. Searching for COMETINHO: The little metric that could. In *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, pages 61–70, Ghent, Belgium, June 2022b. European Association for Machine Translation. URL https://aclanthology.org/2022.eamt-1.9.

Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C. Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte M. Alves, Alon Lavie, Luisa Coheur, and André F. T. Martins. Cometkiwi: Ist-unbabel 2022 submission for the quality estimation shared task, 2022c.

Ricardo Rei, Nuno M Guerreiro, José Pombal, Daan van Stigt, Marcos Treviso, Luisa Coheur, José GC de Souza, and André FT Martins. Scaling up COMETKIWI: Unbabel-IST 2023 submission for the Quality Estimation Shared Task. *arXiv preprint arXiv:2309.11925*, 2023a.

Ricardo Rei, Nuno M. Guerreiro, Marcos Treviso, Luisa Coheur, Alon Lavie, and André Martins. The inside story: Towards better understanding of machine translation neural evaluation metrics. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1089–1105, Toronto, Canada, July 2023b. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-short.94. URL https://aclanthology.org/2023.acl-short.94.

Nils Reimers and Iryna Gurevych. 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, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.365. URL https://aclanthology.org/2020.emnlp-main.365.

Ehud Reiter. A structured review of the validity of BLEU. *Computational Linguistics*, 44(3):393–401, September 2018. doi: 10.1162/coli_a_00322.

Mike Rosner and Claudia Borg. Deliverable d1.25 report on the Maltese language. European Language Equality (ELE); EU project no. LC-01641480 – 101018166, February 2022. URL https://european-language-equality.eu/reports/language-report-maltese.pdf.

Mike Rosner and Jan Joachimsen. *The Maltese Language in the European Information Society*, pages 51–61. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012. ISBN 978-3-642-30681-5. doi: 10.1007/978-3-642-30681-5_8. URL https://doi.org/10.1007/978-3-642-30681-5_8.

Hadeel Saadany and Constantin Orasan. BLEU, METEOR, BERTScore: Evaluation of metrics performance in assessing critical translation errors in sentiment-oriented text. In *Proceedings of the Translation and Interpreting Technology Online Conference*, pages 48–56, Held Online, July 2021. INCOMA Ltd. URL https://aclanthology.org/2021.triton-1.6.

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

Michael J. Sanders. Stratifying a continuous target variable, March 2017. URL https://michaeljsanders.com/2017/03/24/stratify-continuous-variable.html.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 815–823, 2015. doi: 10.1109/CVPR.2015.7298682.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. WikiMatrix: Mining 135M parallel sentences in 1620 language pairs from wikipedia, 2019.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.704. URL https://aclanthology.org/2020.acl-main.704.

Rico Sennrich, Barry Haddow, and Alexandra Birch. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1009. URL https://aclanthology.org/P16-1009.

Hiroki Shimanaka, Tomoyuki Kajiwara, and Mamoru Komachi. 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, October 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-6456. URL https://aclanthology.org/W18-6456.

Georges Van Slype. *Critical Study of Methods for Evaluating the Quality of Machine Translation Systems*. Bureau Marcel van Dijk and CEC, Brussels, 1979.

Mary Snell-Hornby. The professional translator of tomorrow: language specialist or all-round expert? In *Teaching Translation and Interpreting*, page 9. John Benjamins, 1992.

Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA, August 8-12 2006a. Association for Machine Translation in the Americas. URL https://aclanthology.org/2006.amta-papers.25.

Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA, August 8-12 2006b. Association for Machine Translation in the Americas. URL `https://aclanthology.org/2006.amta-papers.25`.

Harold L. Somers. Review article: Example-based machine translation. *Machine Translation*, 14:113–157, 1999.

Lucia Specia, Marco Turchi, Nicola Cancedda, Nello Cristianini, and Marc Dymetman. 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, May 14–15 2009. European Association for Machine Translation. URL `https://aclanthology.org/2009.eamt-1.5`.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to sequence learning with neural networks. *CoRR*, abs/1409.3215, 2014.

Rachael Tatman. Evaluating text output in NLP: BLEU at your own risk. Towards Data Science, January 2019. URL `https://towardsdatascience.com/evaluating-text-output-in-nlp-bleu-at-your-own-risk-e8609665a213`.

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

Ian Tenney, Dipanjan Das, and Ellie Pavlick. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1452. URL `https://aclanthology.org/P19-1452`.

Brian Thompson and Matt Post. 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, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.8. URL `https://aclanthology.org/2020.emnlp-main.8`.

Steven K Thompson. *Sampling*, volume 755. John Wiley & Sons, 2012.

Jörg Tiedemann. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey, May 2012. European Language Resources Association (ELRA). URL `http://www.lrec-conf.org/proceedings/lrec2012/pdf/463_Paper.pdf`.

R.L. Trask. *The History of Basque*. Taylor & Francis, 2013. ISBN 9781136167560.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *CoRR*, abs/1706.03762, 2017.

Yu Wan, Dayiheng Liu, Baosong Yang, Haibo Zhang, Boxing Chen, Derek Wong, and Lidia Chao. 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, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.558. URL `https://aclanthology.org/2022.acl-long.558`.

Jiayi Wang, Kai Fan, Bo Li, Fengming Zhou, Boxing Chen, Yangbin Shi, and Luo Si. Alibaba submission for WMT18 quality estimation task. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 809–815, Belgium, Brussels, October 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-6465. URL `https://aclanthology.org/W18-6465`.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. MiniLM: Deep self-attention distillation for task-agnostic compression of pre-trained transformers, 2020. URL `https://arxiv.org/abs/2002.10957`.

Arthur Wetzel. Preserving linguistic diversity in the digital world. MultiLingual, September 2018. URL `https://multilingual.com/articles/preserving-linguistic-diversity-in-the-digital-world/`.

John S. White and Theresa A. O'Connell. Evaluation in the ARPA machine translation program: 1993 methodology. In *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*, 1994. URL `https://aclanthology.org/H94-1024`.

John S. White, Theresa A. O'Connell, and Francis E. O'Mara. The ARPA MT evaluation methodologies: Evolution, lessons, and future approaches. In *Proceedings of the First Conference of the Association for Machine Translation in the Americas*, Columbia, Maryland, USA, October 5-8 1994. URL `https://aclanthology.org/1994.amta-1.25`.

Aiden Williams, Kurt Abela, Rishu Kumar, Martin Bär, Hannah Billinghurst, Kurt Micallef, Ahnaf Mozib Samin, Andrea DeMarco, Lonneke van der Plas, and Claudia Borg. UM-DFKI Maltese speech translation. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, pages 433–441, Toronto, Canada (in-person and online), July 2023. Association for Computational Linguistics. URL `https://aclanthology.org/2023.iwslt-1.41`.

Philip Williams and Barry Haddow. The ELITR ECA corpus. *CoRR*, abs/2109.07351, 2021.

Shijie Wu and Mark Dredze. Are all languages created equal in multilingual BERT? In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 120–130, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.repl4nlp-1.16. URL `https://aclanthology.org/2020.repl4nlp-1.16`.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol

Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. Google's Neural Machine Translation System: Bridging the gap between human and machine translation, 2016.

Jitka Zehnalová. Tradition and trends in translation quality assessment. *Tradition and Trends in Trans-Language Communication*, OMLS, Vol 2:41–56, 01 2013.

Xianfeng Zeng, Yijin Liu, Fandong Meng, and Jie Zhou. Towards multiple references era – addressing data leakage and limited reference diversity in nlg evaluation, 2023.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with BERT. *CoRR*, abs/1904.09675, 2019.

Ying Zhang, Stephan Vogel, and Alex Waibel. Interpreting BLEU/NIST scores: How much improvement do we need to have a better system? In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal, May 2004. European Language Resources Association (ELRA). URL `http://www.lrec-conf.org/proceedings/lrec2004/pdf/755.pdf`.

Wei Zhao, Goran Glavaš, Maxime Peyrard, Yang Gao, Robert West, and Steffen Eger. On the limitations of cross-lingual encoders as exposed by reference-free machine translation evaluation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1656–1671, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.151. URL `https://aclanthology.org/2020.acl-main.151`.

Yanqing Zhao, Min Zhang, Xiaoyu Chen, Yadong Deng, Aiju Geng, Limin Liu, Xiaoqin Liu, Wei Li, Yanfei Jiang, Hao Yang, et al. Human evaluation for translation quality of chatgpt: A preliminary study. *HiT-IT 2023*, pages 282–287, 2023.