Master Thesis

Machine-Translation Evaluation: Comparing Traditional and Neural Machine-Translation Evaluation Metrics for English→Russian

Natalia Khaidanova

a thesis submitted in partial fulfilment of the requirements for the degree of

MA Linguistics
(Text Mining)

Vrije Universiteit Amsterdam

Computational Lexicology and Terminology Lab
Department of Language and Communication
Faculty of Humanities

Supervised by: Sophie Arnoult
2^{nd} reader: Angel Daza Arevalo

Submitted: June 30, 2023

Abstract

The research investigates traditional and neural (reference-based and reference-free) machine-translation evaluation metrics utilized to estimate the quality of machine-generated translations. Specifically, the study replicates and validates selected findings presented at the WMT21 Metrics Task. Additionally, the work explores the suitability of reference-free neural metrics for professional human translators. The research questions are the following: Are the results of the WMT21 Metrics Task reproducible? Can the findings be fully confirmed? Are reference-free neural metrics relevant for professional human translators?

The research methodology involves computing the scores for both traditional (SacreBLEU, TER, chrF2) and neural (BLEURT-20, COMET-MQM_2021, COMET-QE-MQM_2021) machine-translation evaluation metrics. The computations utilize data provided at the WMT21 Metrics Task with a focus on the English→Russian news and TED talks domains. Furthermore, the computational time is evaluated for each metric, providing insights into the feasibility of employing neural metrics in real-world business scenarios. In addition to the metric analysis, the study categorizes linguistic features that influence metric performance. Furthermore, the applicability of the reference-free COMET-QE-MQM_2021 metric for professional human translators is evaluated by utilizing a new dataset comprised of scientific articles.

The main findings of this study only partially confirm the official results of the WMT21 Metrics Task, thereby raising doubts about the reproducibility of the results. Moreover, it is verified that neural metrics require significant computational costs, which makes them more suitable for final evaluation rather than evaluation during system development. Among linguistic features evaluated, sentence length is found to predominantly impact metric performance, with lengthier sentences causing inferior performance. Finally, the study concludes that current reference-free neural metrics are not relevant for professional human translators as the evaluated reference-free COMET-QE-MQM_2021 metric demonstrated a prominent inclination to assign excessively high scores to poor translations, leading to complications in translation evaluation. The overall findings of the research contribute to a deeper understanding of the effectiveness and relevance of the traditional and neural machine-translation evaluation metrics.

Declaration of Authorship

I, Natalia Khaidanova, declare that this thesis, titled *Machine-Translation Evaluation: Comparing Traditional and Neural Machine-Translation Evaluation Metrics for English→Russian* and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a Master degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Date: June 30, 2023

Signed: 

Acknowledgments

I would like to express my gratitude to the following individuals who have contributed to the completion of this thesis:

First and foremost, I would like to extend my sincere appreciation to my university supervisor, Sophie Arnoult, for her guidance and support throughout the entire research process. Her valuable insights, constructive feedback, and constant encouragement have been instrumental in shaping this project.

I would also like to express my gratitude to the members of the Computational Linguistics and Text Mining Lab (CLTL) — Pia Sommerauer, Ilia Markov, Luís Morgado da Costa, and Isa Maks — for their valuable feedback during the thesis presentation.

In conclusion, I am grateful to all those mentioned above, as well as to the numerous individuals whose names may not appear here but who have supported me in various ways.

List of Figures

List of Tables

Contents

Chapter 1

Introduction

Machine translation, the automated translation of text from one language to another, has become increasingly popular in recent years due to advances in technology and growing globalization. As the quality of machine translation continues to improve, more and more companies are turning to this method over human translation to save time and money. However, the increasing reliance on machine translation has also highlighted the need for automatic evaluation algorithms that can accurately measure the quality of machine translations. Developing such algorithms is essential in ensuring that machine translation can effectively meet the needs of businesses and individuals in the global marketplace, as well as in comparing different machine-translation systems against each other and tracking their improvements over time.

Metrics, i.e., computerized methods of quantitative assessment, are an indispensable component of these automatic evaluation algorithms.

One of the earliest automatic evaluation metrics was developed by the speech recognition research community that proposed using the method of *word error rate (WER)*[1] to evaluate the performance of large vocabulary continuous speech recognition (LVCSR) systems. The metric was later adopted for automatic machine-translation evaluation by comparing the output of machine-translation systems (candidate or hypothesis translation) to a human reference translation. In WER, the number of errors (also known as string edit distance or Levenshtein distance) is computed as the sum of word substitutions (S), deletions (D), and insertions (I). If there are N total words in the reference translation, then WER is calculated as follows (Ali and Renals, 2018)

$$\text{WER} = \frac{S + D + I}{N} \times 100 \tag{1.1}$$

Despite being a simple quantitative measure of performance that is effortless to calculate and interpret, WER has several limitations. One major limitation is that it does not take into account the order or context of the errors, and thus two sentences with the same WER can have different levels of readability or comprehensibility. This highlighted the necessity of refining and expanding WER and led to the development of newer metrics such as BLEU (Papineni et al., 2002).

BLEU and other traditional metrics, e.g. NIST (Doddington, 2002), ROUGE (Lin, 2004), METEOR (Banerjee and Lavie, 2005), TER (Snover et al., 2006), and CHRF (Popovic, 2015, 2016) are based on the idea of comparing a candidate translation with one or several reference translations in terms of the statistics of short sequences of words

[1] https://benjaminmarie.com/traditional-versus-neural-metrics-for-machine-translation-evaluation/

or characters (word or character n-grams). The more n-grams a candidate translation shares with the reference translations, the better it is judged to be. Such simple heuristic algorithms make traditional metrics efficient and language- and domain-independent. Nevertheless, the same characteristics also cause them to be sensitive to changes in word ordering and sentence structure. This can lead to inaccuracies in evaluation as one sentence can have multiple acceptable translations and the more complex the morphological structure of the target language is, the more there are appropriate translation variants.

Despite being the leaders in the marketplace due to low computational costs, traditional machine-translation evaluation metrics are getting substituted for more advanced and accurate neural metrics, e.g., ReVal (Gupta et al., 2015b), YiSi (Lo, 2019), BERTscore (Zhang et al., 2020), BLEURT (Sellam et al., 2020), Prism (Thompson and Post, 2020), and COMET (Rei et al., 2020b), that are based on neural networks or pretrained language models. Neural metrics also have limitations, including the requirement for significant computational power to produce scores, restrictions to particular languages and domains, and difficulties in the interpretation of the results. Nevertheless, neural metrics offer the benefit of encompassing intricate connections between machine and reference translations.

In contrast to traditional metrics that solely rely on reference translations, not all neural metrics employ such references for generating scores. As a result, neural metrics can be classified into two categories: reference-based and reference-free metrics. The latter exclusively utilizes machine translation and its corresponding source text. This characteristic makes it valuable in the field of Machine-Translation Quality Estimation (MT QE), which involves developing systems that can automatically estimate the quality of machine-translated texts without relying on reference translations. Furthermore, reference-free neural metrics hold potential utility for professional human translators as they have the capacity to greatly facilitate the translation process. By utilizing machine-translation algorithms and employing these metrics, translators may effectively address sentences or segments of translated text that exhibit lower quality without the need to evaluate the entire text.

Although reference-free neural metrics offer the advantage of increased efficiency compared to reference-based neural metrics by eliminating the necessity for reference translations, they may fall behind since they do not leverage the benefits provided by the references. Therefore, reference-free neural metrics may exhibit a weaker correlation with human judgments of translation quality than their counterparts.

Human judgments play a crucial role in the evaluation of metric performance, serving as a fundamental benchmark for estimating translation quality. They involve the assessment of machine-translated segments, typically consisting of one or two sentences, whereby human evaluators assign scores indicating the overall quality of the translation. By averaging the scores given to individual segments (referred to as segment-level scores), the system-level score can be obtained. This score represents an aggregate measure that indicates the overall performance of the machine-translation system.

There are multiple frameworks for obtaining human judgment scores, including Relative Ranking (RR), Direct Assessment (DA), and Multidimensional Quality Metrics (MQM) (Mariana et al., 2015). RR is a basic approach that involves ranking each system's translation in order of preference. DA, on the other hand, entails annotating translations based on simple linguistic criteria, such as fluency and adequacy. MQM represents the most advanced framework, incorporating a comprehensive set of evalua-

tion criteria that are more detailed and nuanced compared to DA.

Due to the potential difference in the score ranges generated by metrics and human ratings, statistical measures such as Pearson's r, Spearman's ρ, and Kendall's τ are commonly utilized to compare and assess their relationship. A high system-level correlation denotes the excellent overall performance of the metric for a certain machine-translation system and domain, whereas segment-level correlation is utilized to evaluate the accuracy of the metric at a lower, typically sentence, level.

1.1 Aim and Methods of Research

This thesis project focuses on replicating and reproducing selected research conducted at the WMT21 Metrics Task (Freitag et al., 2021).

The WMT (Workshop on Machine Translation) Shared Task[2] is an annual competition that aims to foster the development of state-of-the-art machine-translation systems by providing a common benchmark for researchers and practitioners to evaluate their models. The WMT Metrics Task concentrates on estimating the quality of automatic machine-translation evaluation metrics. The replication work presented in this study involves evaluating the traditional (BLEU (Papineni et al., 2002), TER (Snover et al., 2006), CHRF2 (Popovic, 2016)), and best-performing reference-based (C-SPECPN (Takahashi et al., 2021), BLEURT-20, COMET-MQM_2021 (Rei et al., 2021, 2020b)) and reference-free (COMET-QE-MQM_2021 (Rei et al., 2021) and OpenKiwi-MQM (Kepler et al., 2019b; Rei et al., 2021)) neural metrics. The evaluation process encompasses determining the correlation between the metrics and MQM human judgments for the English→Russian language pair on the news and TED talks domains.

The main findings of the WMT21 Metrics Task indicate that the range of metrics achieving high-level performance is wider at the system level with the surface-level baselines (BLEU, TER, and CHRF2) joining the winners. The reference-free COMET-QE-MQM_2021 and OpenKiwi-MQM exhibit notable overall performance but perform poorly at the segment level.

The primary objective of this project is to validate the latest findings, verify the reproducibility of the results, as well as specify the applicability of reference-free neural metrics for professional human translators. Therefore, the study aims at answering the following research questions:

1. Are the results of the WMT21 Metrics Task reproducible? Can the findings be fully confirmed?

2. Are reference-free neural metrics relevant for professional human translators?

The research sets a hypothesis that the results of the WMT21 Metrics Task for all the traditional metrics can be fully reproduced. However, the reproducibility of neural metrics' scores is debatable due to the underlying stochastic nature of neural networks. Nevertheless, a notable difference between the obtained and official results is not expected. The study also attempts to establish the superiority of neural metrics over traditional ones with respect to their efficiency in real-world business scenarios of machine-translation evaluation. The research further anticipates a raise in performance from both traditional and neural metrics in the news domain compared to the TED

[2]https://www.statmt.org/wmt22/

talks. This expectation is based on the presence of two reference translations in the news test set and the tendency of news texts to exhibit more predictable language patterns. Besides, the research hypothesizes that the automatic evaluation metrics are not expected to exhibit any bias towards particular machine-translation systems. At the same time, the utility of reference-free neural metrics in their current stage of development is uncertain for professional human translators. Their performance at the segment level holds greater significance in this case compared to the system-level evaluation. However, it is expected to be inferior. As a result, the suitability of these metrics for the specific needs of professional human translators may be called into question.

The study employs various methods, including calculating metric scores on the WMT21 Metrics Task data for the English→Russian machine translations, documenting the computation time, reporting on the system- and segment-level correlations between the metrics and MQM scores of translation quality, performing a qualitative linguistic analysis of metric performance, and evaluating the efficiency of reference-free neural metrics on translations conducted by professional human translators in a domain distinct from news and TED talks.

Chapter 2

Related Work

The inception of the WMT in 2006 marked a significant milestone in the field of machine translation and evaluation metrics. The early editions of the task only focused on the Shared Translation Task (Koehn and Monz, 2006) without incorporating the Metrics Task. The primary objective of the Shared Translation Task was to improve methods of building a phrase translation table, which was a fundamental component used in statistical machine translation, augment the existing systems, or build entirely new translation systems. The assessment of translation quality for the submitted systems during this period relied on the innovative (at that moment) BLEU score, which measured word overlap with a reference translation and manual evaluation conducted with the RR annotation framework.

In 2007, Callison-Burch et al. (2007) began to explore alternative evaluation metrics. This pursuit was driven by the need for a more reliable algorithm that could estimate the quality of machine-translated texts. Consequently, in 2008, a new task called the Shared Evaluation Task (Callison-Burch et al., 2008) was introduced in the WMT. The participants of this task were requested to submit automatic machine-translation evaluation metrics, which were subsequently assessed based on two criteria: the ability to rank systems according to their overall performance (system-level evaluation) and the capability to rank systems on a sentence-by-sentence level (segment-level evaluation). The human judgments were obtained with the RR framework and the correlation between the metrics and system-level human assessment was calculated using Spearman's ρ, a statistical measure of correlation applicable to ranks. Since automatic metrics typically provided scores rather than ranks, the raw scores were converted into ranks prior to computing the system-level Spearman's ρ. At the segment level, rather than calculating a correlation coefficient, it was measured how consistent the automatic metrics were with human judgments. Consistency was calculated as follows: For each individual sentence, a pairwise comparison was conducted between the outputs of two machine-translation systems. It was then counted how many times the relative scores assigned by the metric aligned with the human judgments for that sentence, i.e., how many times the metric assigned a higher score to a higher-ranked system. As the metrics generally produced real numbers as scores, pairs that the annotators ranked as ties were excluded.

The initial goals of the Shared Evaluation Task were to achieve the strongest correlation with human judgments, illustrate the suitability of automatic evaluation metrics as surrogates for human evaluation, address the problems associated with comparing a candidate translation against a single reference, and move automatic evaluation beyond system-level ranking to a more fine-grained segment-level ranking.

In 2008, when the evaluation of metrics was first introduced in the WMT, among translations into English, the METEOR metric (Agarwal and Lavie, 2008) exhibited the strongest correlation with human ratings. METEOR measured precision and recall for token unigrams and applied a fragmentation penalty with flexible word matching based on stemming and WordNet-synonymy. Regarding translations out of English, the part-of-speech variant of BLEU (Popovic and Ney, 2007), which counted the overlap in parts-of-speech sequences rather than words, emerged as the most effective metric. Remarkably, that year a lot of metrics already surpassed the original BLEU.

In 2011, Kendall's τ was employed for the first time to measure the correlation between the metrics and human judgments at the segment level. During that period, a novel metric, MTeRaterPlus developed by Columbia and ETS, appeared to be the best-performing. MTeRaterPlus utilized a machine-learning approach and incorporated both sentence-level and document-level features extracted from ETS's e-rater, an automated essay-scoring engine designed to evaluate writing proficiency (Attali and Burstein, 2006).

In 2012, the Shared Evaluation Task got divided into Metrics and Quality Estimation Tasks, allowing for a more focused evaluation approach. Subsequently, in 2014, Pearson's r replaced Spearman's ρ as a system-level evaluation measure and offered a more refined assessment methodology.

In 2016, it was observed for the first time that character-level metrics demonstrated remarkable performance. Additionally, trained metrics exhibited superior performance compared to non-trained metrics, particularly for translations into English.

During that same year, UoW.ReVal (Gupta et al., 2015a) was the top-performing metric based on system-level correlation for translations into English. For translations out of English, the CharacTer metric (Wang et al., 2016) (UoW.ReVal did not participate in that translation direction) emerged as the most effective.

The UoW.ReVal metric employed a dependency-tree Long Short-Term Memory (LSTM) network to represent both a hypothesis and a reference translation using dense vectors. Therefore, UoW.ReVal stood as one of the pioneering neural metrics.

CharacTer represented a novel character-level metric inspired by the widely utilized TER metric. It was defined as the minimum number of character edits required to adjust a hypothesis until it completely matched the reference, normalized by the length of the hypothesis sentence.

In 2017, the inclusion of metric speed (limited to system-level evaluation) became standard practice in all submissions aiming to facilitate the examination of metrics' ability to establish a strong correlation with human judgments while also considering the potential trade-off in terms of speed reduction. Furthermore, the use of RR for generating human judgments was discontinued and replaced by the DA annotation framework.

In 2019, the Quality Estimation Task (QE as a metric) was conducted in conjunction with the Metrics Task within the framework of WMT, marking the emergence of reference-free metrics.

In the following year, in addition to the original goals set forth at the inception of the WMT, new objectives were introduced. These included moving automatic evaluation beyond segment level by incorporating contextual information, analyzing the influence of reference translations on machine-translation system evaluation, and assessing the efficiency of metrics in evaluating human translations.

During that year, a total of 27 metrics were submitted by 10 research groups with 4 of them being reference-free metrics. Overall, there was no definitive metric that

stood out across all language pairs as the best performer. However, there was a notable improvement in the performance of reference-free metrics compared to the previous year. The correlations achieved by those metrics were competitive with reference-based metrics. Notably, COMET-QE (Rei et al., 2020a) demonstrated effectiveness in recognizing the high quality of human translations, whereas BLEU fell short in capturing such nuances.

In 2021, the MQM annotation framework was first utilized to obtain human judgment scores. DA was considered outdated as numerous metrics had already surpassed its capabilities by that point in time.

Therefore, the WMT Shared Evaluation and Metrics Tasks have undergone significant evolution over the years, reflecting advancements across various aspects. This evolution is observable across multiple levels, such as the methodologies employed in acquiring human judgments, the approaches utilized to evaluate the quality of the metrics, and the design principles guiding the development of the metrics themselves.

Chapter 3

Evaluation of Metrics

While automatic evaluation metrics are designed to predict the quality of a machine-translated text, the effectiveness of the metrics themselves needs to be evaluated as well. Such evaluation typically takes place at the system and segment levels.

Segment-level evaluation focuses on the metric assessment at a low, generally sentence basis. This evaluation involves computing metric scores for individual segments (source-target sentence pairs), which are subsequently compared against segment-level human judgments of translation quality.

In contrast, system-level evaluation entails the comprehensive assessment of the overall performance of a metric across an entire dataset. System-level evaluation involves measuring the correlation between system-level metric scores and the corresponding system-level human judgments. In the context of evaluation metrics, system-level scores typically represent segment-level scores that are averaged across the entire dataset. However, it is important to note that certain metrics, e.g., BLEU, incorporate additional normalization algorithms or penalties when calculating the system-level scores, deviating from a straightforward average. In the case of human ratings, the scores may also be normalized using a z-score transformation where each data point is subtracted by the mean of the dataset and divided by the standard deviation. The purpose of this normalization is to account for any variations in scoring tendencies among the judges.

Both system-level and segment-level evaluations hold significance in assessing automatic evaluation metrics as they offer distinct perspectives on their performance. System-level evaluation provides a holistic outlook on the effectiveness of the metric as a whole, while segment-level evaluation allows for more nuanced and detailed analysis.

3.1 Human Judgments of Translation Quality

The evaluation of metrics relies on human judgments, which serve as the benchmark for estimating the quality of machine translation, enabling the assessment and identification of areas for metric improvement. These judgments involve human evaluators rating segment-level machine translations and assigning a score to each translation based on its quality. There are several annotation frameworks for human evaluation of translation quality, including *Relative Rankings (RR)*, *Direct Assessment (DA)*, and *Multidimensional Quality Metrics (MQM)* (Mariana et al., 2015).

RR involves asking evaluators to rate translations in order of preference, i.e. to assess whether A is better than B, worse than B, or equal to B. It is a simple and straightforward approach, which is useful when there are several translations available

for a given source text. However, there are also some potential limitations of RR, such as this approach requires bilingual annotators who are proficient in both the source and target languages and it does not provide detailed information on specific aspects of the translation that should be improved. An example of the human judgment scores obtained with the RR annotation framework is presented in Table 3.1.

System name	RR score
uedin	2
columbia	4
bbn-combo	1
cambridge	5
cmu-heafield-combo	4

Table 3.1: Example of RR human judgments. The example is taken from the WMT10 Metrics Task. The RR scores (between 1 and 5, where a lower rank value indicates a better output) are given for the machine translations of the 13th segment of the Spanish→English dataset.

The DA annotation framework is used to evaluate the quality of machine translations by comparing them with their reference translations. The assessment is based on a number of criteria, such as fluency and adequacy. The evaluators rate the translation based on each criterion, using a rating scale or rubric, with higher scores indicating better quality. Compared to the RR framework, DA provides a more comprehensive evaluation of specific aspects of translation. Moreover, DA can be used to evaluate translations without requiring annotators to have knowledge of the source language, making it adaptable for use across different languages and domains. However, there are also limitations to using DA, such as the proneness to human bias as different annotators may have different interpretations of what constitutes high- or low-quality translations. An example of the DA scores is given in Table 3.2.

System name	DA score
Facebook-AI	85.0
ICL	84.33333333333333
NiuTrans	92.25
SMU	91.66666666666667
WeChat-AI	43.0

Table 3.2: Example of DA human judgments. The example is taken from the WMT21 Metrics Task. The DA scores (between 0 and 100) are given for the machine translations (five out of 31 participating systems were randomly selected for representation) of the 21st segment of the English→Chinese newstest2021 dataset.

MQM[1] is a relatively new annotation framework created by the Quality Translation Launch Pad group (QTLP 2013). It presents a variety of hierarchical error categories that can be drawn on to create customised metrics based on the end user's needs, and those error categories can be used to evaluate the quality of a machine translation. The original MQM error typology contains seven high-level dimensions, i.e., Terminology, Accuracy, Linguistic conventions (Fluency), Style, Locale conventions, Audience

[1] https://themqm.org/

appropriateness (Verity), and Design and markup. Each dimension comprises more specific error subtypes. For example, Accuracy contains subtypes such as Addition, Mistranslation, and Omission; Audience appropriateness contains Completeness, Legal requirements, and Locale-specific content. When evaluators identify an error instance in a machine translation, they assign the error to an error type and the appropriate severity level, i.e., Neutral, Minor, Major, or Critical. The MQM score is then computed automatically based on the provided data.

Generally, the MQM approach reduces subjectivity, enhances comparability, and greatly improves communication and cooperation among the evaluators proving to outperform other evaluation frameworks. However, despite the current reliability of the MQM framework in generating human ratings, it is crucial to acknowledge that it may not be perfect. The primary drawback of this methodology is the comparatively long time required for annotating data. Consequently, there is a potential for MQM to be substituted by alternative algorithms in the future. This can also be attributed to the continuous evolution of both evaluation metrics and assessment frameworks, suggesting the possibility of advancements that could surpass the capabilities of MQM.

3.2 Statistical Measures of Correlation

Given that machine-translation evaluation metrics may generate scores within a distinct range from human judgments statistical measures of correlation, such as Pearson's r, Spearman's ρ, and Kendall's τ are frequently employed to compare the two. These measures allow the researchers to determine the strength and direction (positive or negative) of the relationship between the metric scores and the human judgment scores of translation quality assigned with any of the annotation frameworks, such as RR, DA, or MQM. In the case of machine-translation evaluation, a correlation value of 1 generally indicates a perfect correlation between the metric and human ratings, a score between 0 and 1 shows a positive relationship between the two assessments, whereas a score of 0 or between 0 and -1 indicates no correlation or a negative relationship. The exceptions are TER and other error metrics, which due to their nature show a negative correlation when performing well. Therefore, a correlation value of -1 for TER indicates a perfect correlation between the metric and human ratings, a score between 0 and -1 shows a negative relationship between the two assessments, whereas a score of 0 or between 0 and 1 indicates no correlation or a positive relationship.

3.2.1 Pearson's r

Pearson's r (Pearson correlation coefficient, Bivariate correlation, or Pearson product-moment correlation coefficient (PPMCC)) is considered a measure of linear association between quantitative random variables, e.g., human (H) and metric (M) scores. The measure is an inferential and descriptive statistic, meaning that it is used to assess whether there is a significant relationship between two variables, as well as describe the strength and direction of the linear relationship between two variables.

The Pearson correlation coefficient ranges from -1 to 1, where -1 indicates a perfect negative correlation, 0 indicates no correlation, and 1 indicates a perfect positive correlation. In general, a correlation coefficient closer to 1 or -1 indicates a stronger linear relationship between the two variables, whereas a correlation coefficient closer to 0 indicates a weaker relationship between the two variables. The formula used to

calculate the Pearson r correlation can be defined as follows:

$$r = \frac{\sum_{i=1}^{n}(H_i - \overline{H})(M_i - \overline{M})}{\sqrt{\sum_{i=1}^{n}(H_i - \overline{H})^2}\sqrt{\sum_{i=1}^{n}(M_i - \overline{M})^2}} \tag{3.1}$$

where H are human assessment scores in a given translation direction, M are corresponding scores as predicted by a given metric. $\overline{H}$ and $\overline{M}$ are their means respectively (Bojar et al., 2016).

The Pearson correlation coefficient should be utilized when all of the following criteria are met: both variables are quantitative, the variables are normally distributed, the data have no outliers, i.e., observations that do not follow the same patterns as the rest of the data, and the relationship is linear, meaning that the relationship between the two variables can be defined by a straight line.[2] Therefore, Pearson's r is a suitable statistical measure to estimate the correlation between the system-level DA or MQM human ratings and metric scores.

3.2.2 Spearman's ρ

Spearman's ρ (Spearman's rank correlation coefficient) is a measure of monotonic relationship between two sets of data. Such an association does not make any assumptions about the distribution of the data. Therefore, it is the appropriate correlation measure when the variables are estimated on a scale that is at least ordinal. The ordinal level of measurement has ordered categories. However, the distances between the categories are not known and cannot be assumed to be equal (van den Heuvel and Zhan, 2022).

Spearman's rank correlation coefficient can range from -1 to 1, where a value of -1 indicates a perfect negative monotonic correlation, 0 indicates no monotonic correlation, and 1 indicates a perfect positive monotonic correlation. Similar to the Pearson correlation coefficient, Spearman's ρ closer to 1 or -1 indicates a stronger monotonic relationship between the two variables, whereas a coefficient closer to 0 indicates a weaker monotonic relationship between the variables. Spearman's ρ, where $d_i = rank_i^H - rank_i^M$, can be calculated as follows:

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2 - 1)} \tag{3.2}$$

Therefore, Spearman's rank correlation coefficient should be utilized when one or more of the following criteria is satisfied: the variables are ordinal, the variables are not normally distributed, the data includes outliers, and the relationship between the variables is non-linear and monotonic.[3] Consequently, Spearman's ρ is most often employed for system-level evaluation when the human judgment scores were obtained with the RR annotation framework.

3.2.3 Kendall's τ

Kendall's τ (Kendall's rank correlation coefficient) is also a measure of monotonic association between two sets of data. It is based on counting the number of concordant and discordant pairs of observations between the two variables (van den Heuvel and Zhan,

[2] https://www.scribbr.com/statistics/pearson-correlation-coefficient
[3] https://www.scribbr.com/statistics/pearson-correlation-coefficient

2022). Concordant pairs are those where both variables increase or both decrease, while discordant pairs are those where one variable increases while the other decreases.

Kendall's τ ranges from -1 to 1, where -1 indicates a perfect negative monotonic correlation, 0 indicates no ordinal correlation, and 1 indicates a perfect positive monotonic correlation. Its formula can be defined as follows:

$$\tau = \frac{n_c - n_d}{\frac{1}{2}n(n-1)} \tag{3.3}$$

where n_c is the number of concordant pairs of observations, n_d is the number of discordant pairs of observations, and n is the total number of pairs of observations.

Therefore, Kendall's τ coefficient is typically utilized for segment-level evaluation, aiming to measure the relationship between the segment-level metric scores and the corresponding human judgments.

Chapter 4

Machine-Translation Evaluation Metrics

Machine translation is the task of automatically translating text from one natural language to another. Machine-translation evaluation metrics are an integral component of this process as they are used to measure the quality of a machine-translation output during the development and evaluation stages of a machine-translation system. As compared to human evaluators, these metrics are cheaper and more reproducible. Besides, they can handle large volumes of data in a short amount of time, which makes them suitable for evaluating machine-translation systems at scale. Furthermore, automated evaluation oftentimes represents the sole feasible approach for assessing the effectiveness of machine-translation systems. For human evaluation, expert translators are required, which can be a challenge for many language pairs due to the rarity of such experts. Moreover, human evaluation is time-consuming and expensive, which makes it impractical for large-scale and fast evaluation of machine-translation algorithms, as required by the very dynamic research area of machine translation. For these reasons, the use of automatic evaluation metrics has been a very active and productive research area in the field of machine translation for over 20 years.

4.1 Traditional Metrics

Most of the metrics created before 2016 are traditional metrics. Traditional metrics for machine-translation evaluation can be viewed as metrics that measure the difference between two strings, i.e., machine translation (candidate translation or hypothesis) and reference translation, based on the word- or/and character n-grams they contain. Traditional metrics do not exploit the source text translated by the system.

One of the main advantages of such evaluation metrics is their low computational costs. Some traditional metrics need to perform shifting of n-grams, which increases the time required to generate the metric scores, particularly for longer strings. Nevertheless, their computation does not require a GPU. Moreover, most of the traditional machine-translation evaluation metrics are language-independent, i.e., the same metric can be applied regardless of the source and target languages. Some traditional metrics may not, however, perform well for certain languages or language pairs, particularly those with different word order or complex morphology as they rely on word- or character-level matching and do not account for differences in grammar or syntax. Subsequently, traditional metrics tend to show a poor correlation with human judgments of translation

quality. Nevertheless, 99% of the research papers rely on traditional metrics (primarily BLEU) to estimate translation quality, rank machine-translation systems, and track their improvements over time (Marie et al., 2021).

In this study, we implement and compare three traditional machine-translation evaluation metrics, namely BLEU (Papineni et al., 2002), TER (Snover et al., 2006), and chrF2 (Popovic, 2016).

4.1.1 BLEU

The central idea behind BLEU (Bilingual Evaluation Understudy), the metric proposed by the IBM MT research group, is defined by Papineni et al. (2002) as follows: the closer a machine translation is to a professional human translation, the better it is. Following this idea, BLEU uses a weighted average of variable length phrase matches against the reference translations, i.e., it compares word n-grams of the candidate translation with the word n-grams of the reference translation and counts the number of matches. The matches are position-independent. The more the matches, the better the candidate translation is.

The crucial aspect of the BLEU metric is its precision measure, which calculates the ratio of unigrams in the candidate translation that also appear in the reference translation. Therefore, BLEU is a precision-based metric. However, the precision measure can be misleading in cases where the candidate translation contains matches but cannot be viewed as acceptable. To address this issue, BLEU uses the modified unigram precision, which accounts for the fact that a reference word should not be considered again after a matching candidate word is found. Corpus-based modified unigram precision is computed after text normalization, i.e., case folding, as follows:

$$p_n = \frac{\sum_{C \in \{Candidates\}} \sum_{n-gram \in C} Count_{clip}(n-gram)}{\sum_{C' \in \{Candidates\}} \sum_{n-gram' \in C'} Count_{clip}(n-gram')} \tag{4.1}$$

In normal settings, precision has to be paired with recall since modified n-gram precision alone fails to enforce the proper translation length. However, naive recall computed over the set of all reference words is not a good measure as BLEU uses multiple reference translations with each potentially choosing a different word to denote the same source concept. Papineni et al. (2002) acknowledges that it would be beneficial to align the reference translations to discover synonymous unigrams and compute recall on concepts rather than words. Nevertheless, given that reference translations differ in length, word order, and syntax, such a computation is complicated to realize in the BLEU metric.

Apart from the modified unigram precision, BLEU has another component called a multiplicative brevity penalty factor that aims at addressing the issue of shorter translations being favored by the precision metric as such translations have fewer opportunities for errors. The penalty reduces the score for the candidate translation if it is significantly shorter than any of the reference translations by making use of the following formula:

$$\text{BP} = \begin{cases} 1 & \text{if} \quad c > r \\ e^{(1-r/c)} & \text{if} \quad c \leq r \end{cases} \tag{4.2}$$

where c is the length of the candidate translation and r is the effective reference corpus length.

Once the geometric average of the modified n-gram precisions (p_n) with n-grams up to length N and positive weights w_n, summing to one (in the baseline, $N = 4$ and $w_n = 1/N$), and the brevity penalty (BP) are calculated, the overall BLEU score is computed as follows:

$$\text{BLEU} = \text{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \tag{4.3}$$

The ranking behavior is more apparent and interpretable in the log domain:

$$\log \text{BLEU} = \min(1 - \frac{r}{c}, 0) + \sum_{n=1}^{N} w_n \log p_n \tag{4.4}$$

The reason for this is that the distribution of scores tends to be skewed, with most translations receiving low scores and a few receiving high scores. In the logarithmic scale, this skewness is reduced making the differences between scores more apparent.

The BLEU metric is frequently used to assess the quality of machine translation, particularly at the system level. Nevertheless, it has certain limitations, e.g., it places no explicit constraints on the order, in which matching n-grams occur, leading to numerous variations of a candidate translation (hypothesis) receiving the same BLEU score. However, not all of these variations are equally grammatically or semantically correct, which means that some translations with the same BLEU score may be less favored by humans (Callison-Burch et al., 2006). Besides, the metric permits the use of multiple references to represent legitimate differences in word choice and word order. Unfortunately, multiple references are rarely available due to the high cost and effort of producing them (Bawden et al., 2020).

4.1.2 TER

TER (Translation Edit Rate) (Snover et al., 2006) is a metric derived from WER (Word Error Rate) and introduced in 2005 by the GALE (Global Autonomous Language Exploitation) research program (Olive, 2005). It estimates the quality of a machine translation by measuring the number of edits (including phrasal shifts) needed to fix a system output so that it exactly matches the closest reference translation. The resulting score is normalized by the average length of the reference. Specifically:

$$\text{TER} = \frac{\text{number of edits}}{\text{average number of reference words}} \tag{4.5}$$

The metric aims to mitigate the high cognitive demands of meaning-based methodologies and the laboriousness of human judgments. It adopts a less complex assessment framework that does not require meaning-based features but still delivers better correlations with human judgments than the BLEU metric.

Apart from WER, the TER metric is also conceptually similar to the machine-translation scoring measure that uses the notion of maximum matching string (MMS) as it only allows a string to be matched once and also permits string reordering. The MMS method has been demonstrated to yield high correlations with human judges (Turian et al., 2003). However, in contrast to it, TER does not explicitly favor longer matching strings and assigns a lower cost to phrasal shifts than MMS and the n-gram-based approaches.

Possible edits needed to match a candidate translation to the closest reference include insertion, deletion, substitution of single words, as well as shifts of word sequences. A shift moves a contiguous sequence of words within the hypothesis to another location within the hypothesis. All edits, including shifts of any number of words, by any distance, have equal cost. In addition, punctuation tokens are treated as normal words and miscapitalization is counted as an edit.

The number of edits for TER is calculated in two phases:

1. A greedy search is used to find the set of shifts by repeatedly selecting the shift that reduces the number of insertions, deletions, and substitutions the most until no more beneficial shifts remain. It is important to note that a shift that reduces the number of insertions, deletions, and substitutions by just one has no reduction in the overall cost, due to the cost of one for the shift itself. However, in this case, the shift is still adopted since the alignment gets more correct and often results in slightly lower edit distance later on.

2. Dynamic programming is used to optimally calculate the remaining edit distance using a minimum-edit distance (where insertions, deletions and substitutions all have a cost of one). The minimum-edit-distance algorithm is $O(n2)$ in the number of words. Therefore, TER uses a beam search, which reduces the computation to $O(n)$ to make the evaluation of long sentences more efficient. The number of edits is calculated for all of the references, and the best (lowest) score is used.

In order to further reduce the space of possible shifts, to allow for efficient computation, several other constraints are used:

1. The shifted words must precisely correspond to the reference words in the destination position.

2. The word sequence in the initial location of the hypothesis and its corresponding reference words must not match exactly.

3. The word sequence of the reference that corresponds to the destination position must be misaligned prior to the shift.

Snover et al. (2006) acknowledges that the TER measure has some limitations, e.g., it ignores notions of semantic equivalence in the candidate translation and assigns a cost of one to all shifts, regardless of their length or distance which seems arbitrary. Exploring other cost measures for shifts could lead to stronger correlations with human judgments. Nevertheless, it was observed that the TER metric is significantly less affected by the number of references compared to BLEU. The study conducted by Snover et al. (2006) shows that the single-reference variant of TER correlates similarly with human judgments of translation quality as the four-reference variant of BLEU. The automatic TER score with four references correlates with a single human judgment as another human judgment does.

4.1.3 chrF

Character n-grams have been an important component of more complex traditional automatic evaluation metrics such as MTeRater (Parton et al., 2011) and BEER (Stanojevic and Sima'an, 2014a,b). However, the individual potential of character n-grams has

not been investigated until 2015 when the CHRF metric (Popovic, 2015, 2016) was introduced. The metric was developed with the aim of examining the impact of character n-grams apart from word n-grams on the machine-transition evaluation. The motivation behind the development of CHRF was to address the limitations of word-based metrics, such as BLEU and TER, that are not able to adequately capture the quality of a machine translation if the latter involves changes in word order or lexical choices as compared with its reference translations. Besides, CHRF is tokenization-independent in contrast to the vast majority of other traditional metrics, which makes it more suitable for the evaluation of morphologically complex or agglutinative languages where words can be formed by concatenating multiple morphemes or affixes.

The general formula for the CHRF score can be denoted as follows:

$$\text{CHRF}\beta = (1+\beta^2)\frac{\text{CHRP} \cdot \text{CHRR}}{\beta^2 \cdot \text{CHRP} + \text{CHRR}} \tag{4.6}$$

where CHRP and CHRR stand for character n-gram precision and recall arithmetically averaged over all n-grams from n = 1 to n = 6:

1. CHRP is the percentage of n-grams in the candidate translation, which have a counterpart in the reference translation;

2. CHRR is the percentage of character n-grams in the reference translation, which are also present in the candidate translation;

and β is a parameter which assigns β times more weight to recall than to precision. If $\beta = 1$, both recall and precision have the same weight; if $\beta = 4$, recall has four times more importance than precision; if $\beta = 1/4$, precision has four times more importance than recall.

The CHRFβ score was calculated for all available machine-translation outputs from the WMT14 (Bojar et al., 2014) and WMT15 (Bojar et al., 2015) shared tasks and then compared with human rankings obtained with the RR annotation framework at the segment level using Kendall's rank correlation coefficient τ. The scores were analyzed for all available target languages. i.e. English, French, German, Czech, Russian, Hindi, and Finnish. The values of the β parameter were investigated in range from 1/6 to 6 resulting in CHRF2 being the most promising version of the CHRF measure.

Experiments with different n-gram weights for the CHRF2 measure, i.e., removing the first n-gram and keeping uniform weights for the rest of the n-grams, assigning doubled weight to the n-grams following the first n-gram, and distributing n-gram weights according to individual n-gram correlations, showed that uniform weights give the best results.

In addition, the CHRF score was systematically compared with the WORDF score based on word n-grams. The experiments on small datasets showed that CHRF outperforms WORDF when dealing with high-quality candidate translations since it does not overly penalise acceptable morpho-syntactic variations of the same translation.

4.2 Neural Metrics

For about eight years neural metrics have been the state of the art in machine-translation evaluation. These metrics differ from traditional ones in their methodology. Instead of comparing the translation output to the reference text through a string-matching

algorithm, neural metrics use pre-trained language models to assess translation quality. This approach has the advantage of being independent of tokenization, providing higher recall, and facilitating fine-tuning for a particular application. Although neural metrics are superior to traditional ones, the research community still overwhelmingly prefers the latter due to high computational costs associated with neural metrics. Moreover, older versions of neural metrics may not function properly if they were not well-maintained, which can be attributed to changes in nVidia CUDA and frameworks such as (py)Torch and Tensorflow. It is possible that the current version of neural metrics will not be functional in the future. Furthermore, neural metrics often come with a great number of hyperparameters, which are often unspecified. Therefore, reproducing a particular score for a particular dataset, as well as explaining the rationale behind the metric assigning a specific score to a machine translation, may be impossible.

In this research, we utilize neural metrics that showed the best performance at the WMT21 Metrics Task (Freitag et al., 2021), i.e., C-SPECPN (Takahashi et al., 2021), BLEURT-20 (Freitag et al., 2021), and COMET-MQM_2021 (Rei et al., 2021, 2020b). These metrics outperformed other algorithms presented at the WMT21 in terms of their correlation with human ratings at the segment level. Although Freitag et al. (2021) acknowledge that there is a clear difference in performance between these metrics that make use of a reference translation and reference-free COMET-QE-MQM_2021 and OpenKiwi-MQM, the latter still show promising results. Therefore, an evaluation of them will also be conducted.

4.2.1 C-SPECPN

The objective of Takahashi et al. (2021) was to design a metric able to detect a significant error that cannot be missed in real practice cases of evaluation. Therefore, in order to achieve it, the C-SPECPN metric was created using pseudo-negative examples, in which attributes of some words in the translation are transferred to the reversed attribute words based on a Word Attribute Transfer (Ishibashi et al., 2020). The metric model was built to handle such severe translation mistakes.

C-SPECPN* makes use of XLM-RoBERTa (Conneau et al., 2020), which is a cross-lingual language model. The model is fine-tuned on the corpus of WMT15-20 DA scores and fine-tuned further again with the pseudo-negative examples derived from the same data and with the WMT20 MQM segment-level scores using a classification algorithm. To generate the pseudo-negative examples, word attribute transfer was applied to all words in an input sentence, and the words having a target attribute were rewritten into their transferred counterparts while those that were not related to the target attribute were kept unchanged. Figure 2.1 illustrates the architecture of the resulting C-SPECPN metric.

*Cross-lingual Sentence Pair Embedding Concatenation, PN: psuedo-negative

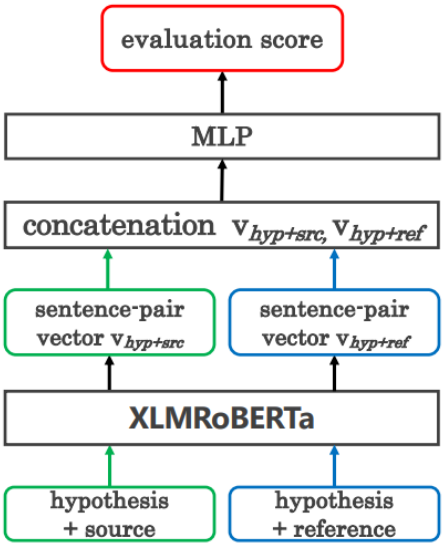

Figure 4.1: C-SPECPN architecture.

The metric uses two sets of input pairs: hypothesis + source and hypothesis + reference. During the evaluation process, the XLM-RoBERTa model encodes the input sentences into sentence-pair vectors (sentence-pair vector $v_{hyp+src}$ and sentence-pair vector $v_{hyp+ref}$). These sentence-pair vectors are concatenated and used to predict the final evaluation score in a multi-layer perceptron (MLP) through a regression algorithm.

Fine-tuning the metric with pseudo-negative examples appeared to improve its performance making it one of the best metrics at the WMT21 Metrics Task. In fact, evaluation results on the WMT21 development dataset showed that fine-tuning the metric on pseudo-negative examples led to a better correlation with human judgment scores compared to fine-tuning without these examples. Nevertheless, based on the findings of the WMT21 Metrics Task, C-SPECPN has certain limitations, e.g., it seems to struggle with word omission and punctuation removal. Consequently, the metric may provide inaccurate estimations of machine-translation quality when certain words or punctuation marks are missing from the reference translation.

4.2.2 BLEURT-20

BLEURT-20 is another embedding-based metric. Due to the absence of an official research paper detailing its inner workings, it is challenging to provide an exact description of its functioning. However, available information suggests that BLEURT-20 makes use of Rebalanced m*BERT (RemBERT) (Chung et al., 2021), which is a BERT-based pre-trained language model designed to balance the representation of different languages in a multilingual setting. The model has 995M parameters during pre-training and 575M parameters during fine-tuning. Further details are provided in Appendix A.1. RemBERT is pre-trained on a large unlabeled text corpus using both Wikipedia and Common Crawl data, covering 110 languages. The fine-tuning process encompasses the tasks from the the Cross-lingual TRansfer Evaluation of Multilingual Encoders (XTREME) benchmark (Hu et al., 2020). The benchmark is used to evaluate the cross-lingual generalization capabilities of the model across 40 languages and 9 tasks. The tasks include *Sentence-*

*multilingual

pair Classification on the XNLI (Conneau et al., 2018) and PAWS-X (Yang et al., 2019) corpora, *Structured Prediction* using the POS (Nivre et al., 2018) and NER (Pan et al., 2017) data, *Question Answering* with the XQuAD (Artetxe et al., 2020), MLQA (Lewis et al., 2020), and TyDi QA (Clark et al., 2020) datasets, and *Information Retrieval* on the BUCC (Zweigenbaum et al., 2018) and Tatoeba data (Artetxe and Schwenk, 2019). The full names of the datasets are listed in Appendix A.2.

To obtain the BLEURT-20 metric, the RemBERT model was fine-tuned on a combination of two datasets: human judgments from the WMT15 - WMT19 Metrics Tasks (z-scores) and generated data. The generated data consist of "perfect" sentence pairs obtained by copying the reference translation into the hypothesis, as well as "catastrophic" sentence pairs, created by randomly sampling tokens for each language pair. The WMT20 data were used for testing. The suffix *-20* in the metric's name denotes the year of the WMT human ratings that were used to build the test set.

During the WMT21 Metrics Task, the evaluation of each metric included the use of challenge sets. These sets comprised two machine-translation outputs, along with their respective source and reference texts. One of the outputs contained a specific type of translation error, while the other did not. The objective was to assess the ability of the metrics to assign a lower score to the machine-translation output that contained the error. The findings indicate that BLEURT-20 demonstrates lower sensitivity to subordination, named entities, terminology, and punctuation when compared to the majority of other neural metrics presented in the shared task. However, the precise implications of this observation were not clarified.

4.2.3 COMET-MQM_2021

COMET-MQM_2021 (Rei et al., 2021, 2020b) is an MQM adaptation of the COMET-DA_2021 model that was further trained for one additional epoch on MQM z-scores extracted from the MQM human ratings for the news dataset of the WMT20 Metrics Task. Figure 2.2 illustrates the metric's architecture [3].

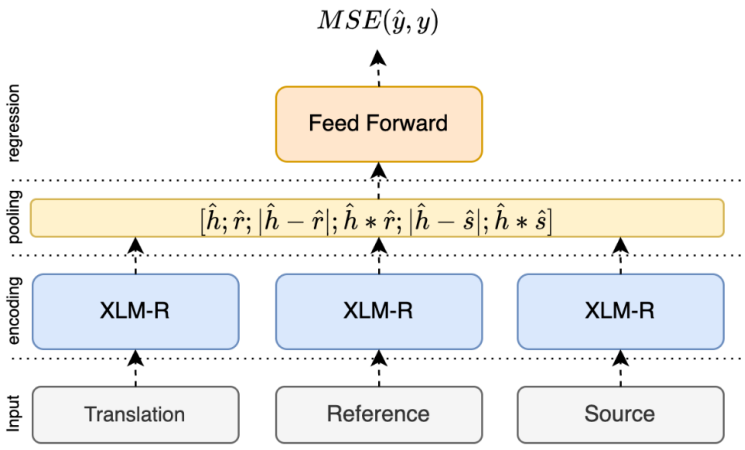

Figure 4.2: COMET-MQM_2021 architecture.

COMET-MQM_2021 makes use of the regression model built on top of XLM-RoBERTa (XLM-R; Conneau et al. 2020) as an encoder model and exploits information

[3] https://unbabel.github.io/COMET/html/models.html

from the hypothesis ($\hat{h}$) and reference translation ($\hat{r}$), as well as from the source sentence ($\hat{s}$). The embeddings of all three inputs are mapped into a shared multilingual feature space. The estimator architecture obtains combined features, i.e., element-wise source product ($\hat{h} * \hat{s}$), element-wise reference product ($\hat{h} * \hat{r}$), absolute element-wise source difference ($|\hat{h} - \hat{s}|$), and absolute element-wise reference difference ($|\hat{h} - \hat{r}|$) using the three embeddings. These combined features that highlight the differences between the embeddings in the semantic feature space are then concatenated to the reference and hypothesis embeddings creating a single vector. The vector then serves as input to a feed-forward regressor. The entire model is trained to minimize the mean squared error (MSE) between the predicted scores ($\hat{y}$) and human judgments (y).

While showing excellent overall performance, as can be stated from the results of the WMT21 Metrics Task, COMET-MQM_2021 displays potential limitations analogous to the C-SPECPN metric. Specifically, it appears to be more sensitive to word or punctuation omission present in the reference translation than most embedding-based metrics submitted at the WMT21. Nevertheless, similar to BLEURT-20, COMET-MQM_2021 shows reduced sensitivity towards factors such as subordination, named entities, terminology, and punctuation in the hypothesis translation when compared to other metrics.

4.2.4 COMET-QE-MQM_2021

COMET-QE*-MQM_2021 (Rei et al., 2021) is a reference-free version of the COMET-MQM_2021 metric. It follows the dual encoder architecture proposed in RUSE (Shimanaka et al., 2018) and replaces the reference translation with the source sentence.

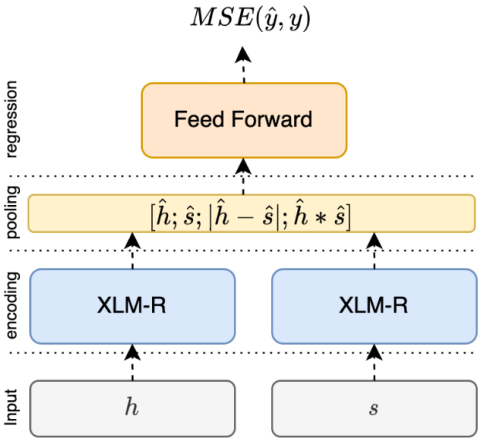

Figure 4.3: COMET-QE-MQM_2021 architecture.

The biggest difference between the COMET-QE-MQM_2021 and COMET-MQM_2021 metrics is that in the reference-free COMET, the combination of features used as input to the feed-forward regressor is different from the COMET that makes use of a reference translation. In the case of COMET-QE-MQM_2021, the combined features are $\hat{h} * \hat{r}$ and $|\hat{h} - \hat{s}|$ and the final vector to the feed-forward regressor is the concatenation of these features together with $\hat{h}$ and $\hat{s}$.

*quality estimation

4.2.5 OpenKiwi-MQM

OpenKiwi-MQM (Kepler et al., 2019b; Rei et al., 2021) is a multitask reference-free QE framework that estimates a sentence-level MQM score along with word-level OK/BAD tags. The goal of word-level QE is to assign quality labels (OK or BAD) to each machine-translated word, as well as to gaps between words (to account for context that needs to be inserted), and source words (to denote the words in the original sentence that have been mistranslated or omitted in the hypothesis translation). Sentence-level QE, on the other hand, aims at determining the quality of the entire translated sentence based on various factors, such as the time taken by a human to edit it or the number of edit operations required to fix it in terms of HTER (Human Translation Error Rate; Snover et al. 2006).

The OpenKiwi-MQM metric is trained on top of XLM-RoBERTa using proprietary MQM data from the customer support domain covering several industries, such as technology and travel industries. The data are comprised of 1.1 million source-hypothesis pairs with corresponding MQM annotations encompassing 38 language pairs, most of which are out-of-English.

In the OpenKiwi architecture, in contrast to COMET-QE-MQM_2021, the input data (source and hypothesis) are jointly encoded. A sentence pair representation is then obtained using average pooling over the hypothesis word embeddings and then used as features to a feed-forward regression layer that learns to produce a sentence-level score. At the same time, the word embeddings from the hypothesis are used to predict OK/BAD tags and therefore, the model is trained in a multitask setting (regression and sequence labeling). Detailed information about the model's hyperparameters is provided in Appendix A.3.

The OpenKiwi framework outperforms other open-source QE toolkits, such as WCE-LIG (Servan et al., 2015), QuEST++ (Specia et al., 2015), Marmot (Logacheva et al., 2016) and deepQuest (Ive et al., 2018), on both word level and sentence level. Since its release, OpenKiwi was adopted as a baseline system for the WMT19 QE Shared Task. Moreover, all the winning systems of the word-, sentence- and document-level tasks of the WMT19 QE Shared Task (Kepler et al., 2019a) used OpenKiwi as their building foundation.

Chapter 5

Comparative Analysis of MT Evaluation Metrics

5.1 Dataset Description

The test sets for metric evaluation are composed of two domains: news articles (newstest2021) and TED talks (tedtalks). While the language style of the TED talks domain is more casual, it covers a diverse range of topics and vocabularies. Two domains allow the evaluation metrics to be thoroughly assessed for their ability to generalize across different types of data.

The newstest2021 test set consists of 1002 source segments (usually one, sometimes two sentences), while the tedtalks set has 512 segments. In both datasets, there are 14 machine-translation outputs per segment. The systems used to generate these outputs are identical in both datasets. Since one major use case for automatic metrics is choosing among different versions of the same system during system development, translations from five other machine-translation systems were also included and named metricsystem{1,..,5}. These additional versions were based on the NMT models trained on unconstrained data and incorporated different variations such as baseline models, fine-tuned models, and models that considered document context. It is important to note that these models were not specifically trained to perform well for either the news or TED talks domain. For system evaluation, there are two reference translations in newstest2021 and only one reference translation in the tedtalks data. All references were created in the same translation direction as the machine-translation systems being evaluated.

The source sentence, reference translation(s), and machine-translation system outputs were mainly derived from the WMT21 News Translation Task (Akhbardeh et al., 2021). The TED talks transcripts were obtained from OPUS[1], based on the corpus released by Reimers and Gurevych (2020). The English transcripts of the TED talks were translated into multiple languages by volunteers. To minimize the effect of translationese (awkwardness or ungrammaticality of translation) in the Chinese→English part of the test set, a native Chinese speaker selected talks with natural-sounding Chinese translations. Then the same talks were extracted from the corpus to create the English→Russian part of the test set where the translation was already available in the corpus and approved by professional translators.

[1] https://opus.nlpl.eu/TED2020.php

The main human rating scores for the English→Russian language pair were obtained via the MQM annotation framework. The annotation was performed by Unbabel[2] that used a single professional native language annotator with several years of experience in translation error detection based on different variations of the MQM framework. The company provided a proprietary variant of MQM specifically tailored for the Russian language annotation. The annotator was given full document context and instructed to highlight spans of errors according to the categories specified in the typology. They were also asked to indicate error severity. The Unbabel severity options included a Critical error severity but did not cover a Neutral category. All error categories were weighted equally within each severity level. MQM scores at a segment level were calculated by summing the number of errors of each severity in a segment and applying a severity weight as described in Table 5.1. As Unbabel did not impose any limits on the number of errors in a segment, the organizers of the Metrics Task applied normalization of the score by segment length.

Severity	Category	Weight
Critical	all	10
Major	all	5
Minor	all	1

Table 5.1: Unbabel's MQM error weighting.

The MQM-based human evaluation of newstest2021 was conducted on a subset of segments with 527 out of 1002 segments being annotated. In the tedtalks test set, all segments were annotated. Freitag et al. (2021) declared that this approach had the advantage of generating more reliable ratings and gave the organizers the opportunity to run the same human evaluation on a different domain (TED talks) on the output generated by the same machine-translation systems in order to test the generalization capabilities of the metrics.

Apart from the MQM annotation, human DA evaluation was conducted for the main submissions in the news domain for all language pairs as part of the WMT evaluation campaign. For translations out of English, segment-level ratings were collected on a 0-100 scale taking into consideration document context and using source-based evaluation that involved a group of researchers and translators. For each machine-translation system, only a subset of documents received ratings, with the rated subset differing across systems. The provided evaluation scores included both raw DA scores and per-rater z-normalized versions of the DA scores.[*]. The exact number of annotated segments per machine-translation system is listed in Appendix B.1 and the examples of the human judgment scores obtained with each annotation type are presented in Table 5.2.

[2]https://unbabel.com/

[*]Standardized scores obtained through the DA evaluation of translation quality by multiple human judges.

MT System	MQM	raw DA	z-normalized DA
Facebook-AI	100.0	90.66666666666667	0.06278439171625057
Manifold	80.0	88.0	0.603750396342278
Nemo	77.77777777777777	67.25	-0.9092686293737025
NiuTrans	61.53846153846153	74.66666666666667	-0.19373107505327758
Online-A	96.15384615384615	72.0	0.16275738939149703

Table 5.2: MQM, raw DA, and per-rater z-normalized DA human judgment scores for the 17th segment of the newstest2021 English→Russian dataset. The five systems were randomly selected for representation.

5.1.1 Linguistic Features of News Domain

Russian is classified as a synthetic language, which implies that it is capable of merging several various linguistic concepts, such as tense, mood, voice, and case, into a single word. Furthermore, Russian does not employ articles. These properties ostensibly make Russian more concise than English, which separates these concepts into distinct words. However, Russian news articles typically contain longer sentences than English news articles. For instance, the average character count of an English newstest2021 source sentence is 145, while the mean length of a Russian reference translation is 158 characters. The primary reason for this phenomenon is attributable to the more formal nature of the Russian news style, whereby a single English term may be translated with multiple Russian words, despite the availability of a one-word substitute. Such instances extracted from the newstest2021 test set are presented in Table 5.3.

Source	Possible Translation	Reference
to self-isolate	самоизолироваться	пройти карантин
global	глобальный	ведущийся во всем мире
ventilators	ИВЛ	искусственная вентиляция легких

Table 5.3: Examples of the instances where a single English term is translated with multiple Russian words, despite the availability of a one-word substitute. The examples are taken from the newstest2021 dataset.

Another factor that contributes to the lengthy nature of Russian news articles is the tendency to fully translate foreign abbreviations, which reflects the formal nomenclature utilized by the Russian governmental and similar entities. This is in keeping with the descriptive style of Russian news reporting. Consequently, the concise names of organizations in English are often rendered in Russian as extended, formal expressions. The demonstrative examples of such instances taken from the newstest2021 dataset are provided in Table 5.4.

Source	Reference
Labour MP	член парламента от Лейбористской партии
CDC	Центры по контролю и профилактике заболеваний
BJP MLA from Indore	член Законодательного собрания от Индаура из партии БДП
the CARES Act	закон «Об оказании помощи и экономической безопасности в связи с распространением коронавирусной инфекции»

Table 5.4: Translation examples of foreign abbreviations from English into Russian. The examples are taken from the newstest2021 dataset.

Furthermore, English news articles often include direct quotes from sources to provide more information and add credibility to the article. Russian news articles, in their turn, often use indirect speech instead of direct quotes, which can make the language sound more formal and objective. Interestingly enough, this characteristic is comparatively scarce in the newstest2021 dataset, wherein direct speech is frequently translated literally in both reference translations. Such an underrepresentation of the phenomenon can potentially lead to evaluation metrics assigning lower quality scores to translations that use indirect speech when rendering direct citations.

Therefore, the fundamental issue lies in the potential of machine-translation systems to preserve the characteristics of the English news style rather than adapting them to Russian. This, in turn, can influence the effectiveness of evaluation metrics, which should be capable of evaluating a machine-translation hypothesis against its reference translation(s) and/or source text. In the case of the newstest2021 test set, the primary objective of a metric is not only to determine whether a machine translation conveys the intended meaning but also whether it adheres to the language style of Russian news articles.

Linguistic Features of newstest2021 Reference Translations

As a result of the importance placed on reference translations, it is worth noting that the translators given newstest2021 data were instructed to perform a comprehensive translation of the entire text instead of translating individual text segments. Such a conclusion was made as reference translations frequently contain contextual information that cannot be derived directly from the sentence being translated but from the sentences surrounding it. In the 50 newstest2021 reference-translation pairs that were examined, there are 5, in which only one reference translation is free and 3, in which both references are free translations*. Table 5.5 contains information about the source and reference pairs in the latter. The aforementioned 5 cases are listed in Appendix B.2.

Table 5.5: Examples of test segments from the newstest2021 dataset, in which both references are free translations.

Source	References
... we expect employers to show those employees who will have to quarantine because of the law the flexibility they need.	... мы ожидаем, что работодатели проявят необходимую гибкость в отношении сотрудников, которые окажутся на карантине из-за *нового* закона.
	... мы ожидаем, что работодатели проявят по отношению к сотрудникам, которые окажутся на карантине из-за *нового* закона, соответствующую гибкость.

*Free translation is a translation approach that prioritizes conveying the meaning and intent of a source text in a way that is natural and idiomatic in the target language, rather than providing a word-for-word or literal translation. Free translation often involves adapting the style and structure of the original text to make it more natural and understandable to the target audience. The goal of free translation is to capture the intended meaning and impact of the original text while maintaining coherence and relevance in the translated version.

Explanation: Both references contain the word *нового* (*new*) referring to law. The conclusion about the newness of the law can only be made when the context of the previous sentences is considered.	
Britain's Recovery Trial programme ... has already pinpointed one promising new drug to tackle the disease ...	Британская программа Recovery ... уже позволила определить одно многообещающее новое лекарство для борьбы с *вирусом* ...
	Британская программа RECOVERY ... уже обнаружила многообещающее лекарство от *вируса* ...
Explanation: In both references, the word *disease* is translated with the words *вирусом* and *вируса*, which are different forms of the same word *вирус* (*virus*). Such a translation is only possible if the context of the previous sentences that mention Covid-19 is taken into account.	
No one knows which they have been given.	Никто не знает, кому из *пациентов* что дали.
	Никто из *пациентов* не знает, кто получил настоящее лекарство.
Explanation: Both references contain the word *пациентов* (*patients*), which is not present in the source sentence and can only be derived from the context of the whole text.	

In normal circumstances, this approach of free translation is typically favored over relying solely on the meaning of a single sentence. However, considering that the majority of evaluation metrics assess the quality of the translated text at a segment level and depend on reference translations as the gold-standard data, free translation should only be applied when no other translation methods are feasible. Although state-of-the-art machine-translation systems take into account the context of the entire text when producing segment-level outputs, in the news domain, a literal translation does not necessarily indicate poor translation quality. As a result, when comparing a literal machine translation to a free reference translation, a metric may assign a lower quality score to the system output, which does not accurately reflect the actual situation.

We also observed that some of the references in the newstest2021 test set contain grammatical and semantic mistakes. The mistakes detected in the 50 examined reference-translation pairs are listed in Tables 5.6 and 5.7.

Reference	Correct Variant	Mistake Type
У нас должн**ы** быть возможность действовать быстро и решительно ...	У нас должн**а** быть возможность действовать быстро и решительно ...	Contradiction in grammatical number
Мы успешно снизили заболеваемость и предотвр**ащаем** повторн**ого** рост**а** вируса ...	Мы успешно снизили заболеваемость и предотвр**атили** повтор**ный** рост вируса ...	Contradiction in tense and case
Уже **были были** использованы большие запасы плазмы ...	Уже **были** использованы большие запасы плазмы ...	Repetition

Table 5.6: Grammatical mistakes detected in the newstest2021 reference translations, a correct variant of the translation, and the typology of the mistakes.

Source	Reference	Translation of Reference	Correct Variant
... random old lady ...	... женщина неопределенного возраста ...	... lady of unknown age ...	... какая-то пожилая женщина ...
The Labour MP told Sophy Ridge ...	Член парламента от Лейбористской партии Софи Ридж заявила ...	The Labour MP, Sophy Ridge, said ...	Парламентарий-лейборист сказал Софи Ридж ...

Table 5.7: Semantic mistakes detected in the newstest2021 reference translations, their literal meaning, and a correct variant of the translation.

While embedding-based neural metrics are more resistant to such a type of noise in the data, traditional metrics are likely to be much more affected by it, which may potentially complicate their evaluation. However, the availability of two reference translations in the newstest2021 dataset, at least one of which is grammatically correct, should compensate for any inconsistencies.

5.1.2 Linguistic Features of TED Talks Domain

There are noticeable dissimilarities in the language styles of news articles and TED talks. While news articles tend to employ a more formal and objective language style characterized by longer sentences, TED talks often feature a more conversational and informal language style with shorter sentences and the use of more concise and accessible language. This pattern is observable in both English and Russian as evidenced by the fact that the average length of the source sentence in the tedtalks dataset is 20 characters, while for the Russian reference, it is 16 characters.

More detailed analysis and comparison of the linguistic characteristics of English and Russian TED talks revealed a remarkable similarity between the language styles employed by these two languages in this particular case. While the language used in news articles exhibits considerable differences, for instance, in the handling of abbreviations, the language style of Russian TED talks is more similar to that of English. Specifically, while English abbreviations in news articles are typically translated in full, they are often left short in TED talks. Furthermore, extended formal terms in English may even be converted into abbreviations in Russian TED talks translations, as exemplified in Table 5.8 below.

Source	Reference
Stephen Hawking warns that "*Artificial intelligence* could end mankind."	Стивен Хокинг предостерегает: «*ИИ* может положить конец человечеству».

Table 5.8: Translation example of an extended term *artificial intelligence* being translated as an abbreviation *ИИ* (*AI*). The example is taken from the tedtalks dataset.

Given that TED talks aim to elicit emotions in their audience rather than simply provide factual information like news articles, the language employed in TED talks is characterized by its emotional intensity and frequent use of epithets and metaphors. The expression of these attributes varies across different languages, primarily between English and Russian. Consequently, the references in the tedtalks dataset are often free translations of the source text. However, in contrast to the reference translations in the newstest2021 dataset, these free translations largely reflect the specific nature of the domain, rather than individual characteristics of the translation methodology.

Furthermore, the extent of free translation in the tedtalks data surpasses mere word-level modifications, with entire sentences often being free translations. Table 5.9 provides illustrations of such cases.

Source	Reference	Translation of Reference
And what a wonderful thing it is.	Удивительно придумано.	Wonderfully thought of.
But basically that's what we're talking about.	Но масштаб примерно таков.	But the scale is approximately this.
It's sort of like ding, ding, ding.	Просто тихие щелчки.	Just quiet clicks.
Yes, they're that, too.	Безусловно.	Certainly.

Table 5.9: Examples of English source sentences, their completely free Russian reference translations, and literal translations of these references. The instances are taken from the tedtalks dataset.

The primary challenge posed by free translations of complete sentences, in contrast to literal translations, lies in the proliferation of numerous potential translation variations, each containing a different set of concepts. Consequently, evaluating the accuracy of machine translation against such free reference translations becomes exceedingly difficult for both traditional and neural metrics, as there is no definitive "ground truth" established. Therefore, considering the prevalence of such instances in the tedtalks dataset, it is reasonable to expect that both traditional and neural metrics will demonstrate poorer performance for this dataset compared to the newstest2021 data.

Linguistic Features of tedtalks Reference Translations

As established in Section 5.1.2, the frequent utilization of free translation in Russian tedtalks references does not stem from the translation methods employed by translators of these particular TED talks, but rather from the unique characteristics inherent to the TED talks domain. Additionally, upon examining approximately 100 reference translations, no grammatical or semantic mistakes were identified. Therefore, despite the availability of only one reference per source sentence in the tedtalks dataset, the quality of the tedtalks reference translations can be regarded as superior to those in the newstest2021 data.

5.2 Metrics Implementation

The implemented traditional metrics include SacreBLEU*, TER, and chrF2. The TorchMetrics[6] Python library was utilized to conduct the evaluation of these metrics.

The implementation of neural metrics involved employing three out of the five metrics described in Sections{4.2.1,...,4.2.5}, namely BLEURT-20, COMET-MQM_2021, and COMET-QE-MQM_2021. In the case of C-SPECpn, the code required to run the metric was not publicly available, which prevented its execution. The OpenKiwi-MQM

*SacreBLEU is an extension and improvement upon the original BLEU metric. It incorporates additional refinements and normalization techniques to address certain limitations of BLEU and applies a more sophisticated tokenization procedure.

[6]https://pypi.org/project/torchmetrics/

metric was not implemented due to a subprocess error with the SentencePiece text tokenizer during the installation process. Moreover, since BLEURT-20 and COMET-MQM_2021 do not support receiving two reference translations at once as input, they were executed for each reference translation in the newstest2021 set separately. The correlation for each pair of scores was then averaged to obtain the ultimate result.

The scores for every implemented traditional and neural metric were computed for both the newstest2021 and tedtalks datasets, as well as for every machine-translation system. Therefore, the total number of machine-translation sentence segments to process amounted to 14,028 for the newstest2021 and to 7,168 for the tedtalks data. The time required to produce the metric scores is listed in Table 5.10. The numbers were obtained by calculating segment-level metric scores for a subset of systems using one of the following runtime types: Local CPU (AMD A10-9620P RADEON R5), Google Colab CPU, or Google Colab standard T4 GPU. The time averaged across the subset was multiplied by the total number of machine-translation systems to get the approximate time required to compute metric scores for the whole data. Since Google Colab limits the time of GPU usage with no subscription, the scores obtained with the Google Colab standard T4 GPU runtime type are only available for some of the metrics implemented on the newstest2021 and tedtalks datasets. Computing the time necessary to produce the COMET-MQM_2021 and COMET-QE-MQM_2021 scores using local CPU did not appear feasible due to time constraints. Therefore, the scores for the majority of machine-translation systems were calculated with a Google Colab CPU.

Table 5.10: Computational time of each metric on the newstest2021 and tedtalks data for the English→Russian language pair. The runtime types include AMD A10-9620P RADEON R5 (Local CPU), Google Colab CPU (Colab CPU), and Google Colab standard T4 GPU (Colab GPU). Note that the numbers may not be fully reproducible as the time varies with each execution of the code.

newstest2021

Runtime Type	Traditional			Neural		
	SacreBLEU	TER	CHRF2	BLEURT-20	COMET-MQM_2021	COMET-QE-MQM_2021
Local CPU	3 min., 38 sec.	24 min., 46 sec.	1 h., 16 min.	23 h., 41 min.	-	-
Colab CPU	1 min., 8 sec.	5 min., 28 sec.	32 min., 11 sec.	15 h., 6 min.	23 h., 11 min.	8 h., 39 min.
Colab GPU	1 min., 8 sec.	5 min., 17 sec.	27 min., 17 sec.	12 h., 33 min.	19 h., 36 min.	-

tedtalks

Local CPU	55.43 sec.	2 min., 6 sec.	15 min., 29 sec.	9 h., 59 min.	-	-
Colab CPU	24.07 sec.	34.16 sec.	7 min., 16 sec.	2 h., 25 min.	5 h., 3 min.	3 h., 16 min.
Colab GPU	22.84 sec.	27.34 sec.	5 min., 38 sec.	-	-	-

Total (newstest2021 and tedtalks)

Local CPU	4 min., 33 sec.	26 min., 52 sec.	1 h., 31 min.	33 h., 40 min.	-	-
Colab CPU	1 min., 32 sec.	6 min., 2 sec.	39 min., 27 sec.	17 h., 31 min.	28 h., 14 min.	11 h., 55 min.
Colab GPU	1 min., 31 sec.	5 min., 44 sec.	32 min., 55 sec.	-	-	-

As anticipated, the computational time required to execute the traditional metrics is significantly lower compared to that of the neural metrics. Specifically, on the newstest2021 and tedtalks datasets, the former runs approximately 73 and 78 times faster than the latter, respectively, when executed with a Google Colab CPU. Furthermore, a noticeable difference in the computational time of the individual traditional metrics can be observed. SacreBLEU takes just over a minute and a half to produce the scores, while CHRF2 requires approximately 39 minutes and TER falls in the middle. In a similar manner, the computational speed of the COMET-MQM_2021 metric is noticeably slower than that of the other neural metrics being evaluated. This difference can be attributed to the larger number of inputs required by COMET-MQM_2021 as compared to BLEURT-20 and COMET-QE-MQM_2021. The comparison between the BLEURT-20 and COMET-QE-MQM_2021 metrics reveals that their computational efficiency differs depending on the dataset being evaluated. Specifically, the results indicate that the COMET-QE-MQM_2021 metric is more efficient than BLEURT-20 when being executed on the newstest2021 dataset. However, this relationship is reversed for the tedtalks domain. The observed discrepancy can be attributed to the variance in the number of reference translations available for each dataset.

The findings also highlight the advantages of utilizing cloud computing platforms for resource-intensive tasks since it appears more practical to utilize a Google Colab CPU, which demonstrated significantly faster metric execution compared to the AMD A10-9620P RADEON R5 local CPU. This discrepancy can be attributed to the presence of more powerful CPUs, optimized configurations, and ample resources within Google Colab servers. Consequently, these factors enable Google Colab CPUs to handle tasks more efficiently in contrast to individual local CPUs, which may have lower specifications or limited resources.

5.3 Results for the newstest2021 Data

System-level Pearson's r correlation was derived for the traditional metrics by establishing the correlation between the system-level metric scores and MQM system-level scores. The correlation for the neural metrics was obtained by averaging the segment-level metric scores across all segments and calculating the correlation between the resulting system-level metric scores and MQM system-level scores. For BLEURT-20 and COMET-MQM_2021, the correlation was computed for each reference translation individually and then averaged to obtain the ultimate system-level Pearson's r correlation. Segment-level Kendall's rank correlation coefficient τ was obtained by computing the correlation between the segment-level metric and MQM scores for each machine-translation system and averaging the resulting scores across all systems to get the mean value.

5.3.1 System-level Pearson's r Correlation

The results of establishing the correlation between the metrics and human judgments derived with the MQM annotation framework for the newstest2021 data indicate that neural metrics generally exhibit a much stronger system-level correlation with human ratings compared to traditional metrics according to the Pearson correlation coefficient presented in Table 5.11.

Nevertheless, it should be noted that the obtained scores deviate from the official results of the WMT21 Metrics Task. The differences in scores range from a minimum of 0.003 for BLEURT-20 to a maximum of 0.340 for SacreBLEU.

The divergence observed in the traditional metrics can be attributed to potential variations in the computation of the metric scores. It is possible that the organizers of the WMT21 Metrics Task employed a different approach, wherein the metrics were provided solely with the first reference translation. However, attempting to replicate the calculation in a similar manner did not yield significant deviations compared to the results obtained using both reference translations. Due to the lack of explicit information regarding the specific methodology utilized to calculate the correlation between the metrics and human ratings at the WMT21 Metrics Task, replicating the results becomes challenging.

The observed variation in the results for the neural metrics can additionally be explained by considering the underlying stochastic nature of neural networks. As a result, the same input data can propagate differently through the network leading to diverse outputs.

The system-level Pearson's r correlation between the metric, raw DA, and per-rater z-normalized DA human ratings are presented in Appendix B.3, Tables B.3 and B.4.

Baselines			Ref. based		Ref. free
SacreBLEU	TER	CHRF2	BLEURT-20	COMET-MQM_2021	COMET-QE-MQM_2021
0.167	-0.211	0.638	**0.771**	0.546	0.651
Official Results of WMT21 Metrics Task					
0.507	-0.041	**0.783**	0.768	0.659	0.688

Table 5.11: System-level Pearson's r correlation between the metric scores and MQM human ratings for each of the implemented metrics on the newstest2021 data. The best Pearson's r correlation is marked in bold. Since some metrics, such as BLEU, aim to achieve a strong positive correlation with human judgments, while error metrics, such as TER, aim for a strong negative correlation, we compare metrics via the absolute value $|r|$ of a given metric's correlation with human assessment.

The general results of system-level Pearson's r correlation between the metrics and MQM scores for the newstest2021 data indicate that among the metrics evaluated, BLEURT-20 emerges as the most effective, whereas TER exhibits a lack of any association with human judgment scores. CHRF2 displays comparable performance to the neural metrics and even surpasses the performance of COMET-MQM_2021. Furthermore, it is worth noting that the reference-free COMET-QE-MQM_2021 metric exhibits a superior system-level correlation with the MQM scores compared to its both reference- and source-based counterpart, COMET-MQM_2021. This observation suggests that COMET-QE-MQM_2021 may potentially be the most suitable metric for system-level evaluation due to its favorable performance, relatively efficient utilization

of computational resources for score generation, and its ability to alleviate the need for producing expensive and time-consuming human reference translations.

5.3.2 Segment-level Kendall's τ Correlation

Upon examining Kendall's τ correlation provided in Table 5.12, which showcases the relationship between the metrics and MQM human ratings at a segment level for the newstest2021 data, it can be observed that the difference in effectiveness between the traditional and neural metrics is not that significant compared to the system-level evaluation. This can be evidenced by the disparity between the least effective traditional metric (SacreBLEU) and the most effective neural metric (BLEURT-20) being only 0.152 points.

System	Baselines			Ref. based		Ref. free
	SacreBLEU	TER	CHRF2	BLEURT-20	COMET-MQM_2021	COMET-QE-MQM_2021
Facebook-AI	0.110	0.106	0.181	**0.225**	0.196	0.152
Manifold	0.137	0.151	0.206	**0.295**	0.289	0.262
Nemo	0.103	0.120	0.156	0.272	**0.307**	0.273
NiuTrans	0.184	0.182	0.245	**0.317**	0.315	0.279
Online-A	0.095	0.075	0.173	**0.257**	0.253	0.224
Online-B	0.119	0.083	0.183	0.325	**0.345**	0.301
Online-G	0.121	0.141	0.214	**0.296**	0.291	0.273
Online-W	0.153	0.120	0.186	**0.211**	0.159	0.136
Online-Y	0.098	0.103	0.146	**0.372**	0.336	0.293
metricsystem1	0.127	0.153	0.208	0.231	**0.233**	0.199
metricsystem2	0.147	0.167	0.214	0.264	**0.284**	0.254
metricsystem3	0.090	0.114	0.147	**0.322**	0.290	0.253
metricsystem4	0.120	0.167	0.223	0.232	**0.238**	0.194
metricsystem5	0.129	0.114	0.164	**0.240**	0.214	0.180
Average	0.124	0.128	0.189	**0.276**	0.268	0.234
Official Results of WMT21 Metrics Task						
	0.120	0.117	0.193	**0.286**	0.276	0.242

Table 5.12: Segment-level Kendall's τ correlation between the metric scores and MQM human ratings for each of the implemented metrics on the newstest2021 data. The best Kendall's τ correlation is marked in bold. Since some metrics, such as BLEU, aim to achieve a strong positive correlation with human judgments, while error metrics, such as TER, aim for a strong negative correlation, we compare metrics via the absolute value $|\tau|$ of a given metric's correlation with human assessment.

Furthermore, the obtained results are very close to the official results of the WMT21 Metrics Task with BLEURT-20 maintaining its position as the most effective metric on both the system and segment levels. COMET-MQM_2021 falls behind with a negligible gap in performance. COMET-QE-MQM_2021 secures a third place among the evaluated metrics and CHRF2 once again stands out as the top-performing traditional metric. However, SacreBLEU performs worse than the TER metric, which ranked last in the system-level evaluation. Nevertheless, the difference in their performances cannot be considered substantial.

Among all the neural metrics assessed, there is a notable pattern where the worst correlation with the MQM scores occurs consistently for one specific machine-translation

system, namely Online-W. This observation suggests a potential bias towards this particular system. As a result, the assumptions regarding the absence of biases in the metrics towards specific systems may be disproved. However, detecting the cause of this bias does not appear feasible within the scope of this study as it requires a very detailed analysis of the segment-level metric scores.

Segment-level Kendall's τ correlation between the metric scores and DA or per-rater z-normalized DA human ratings are presented in Appendix B.3, Tables B.5 and B.6.

5.3.3 Error Analysis

To further investigate the effectiveness of the best-performing metric (BLEURT-20) in assessing the quality of machine translation for the newstest2021 dataset, a comprehensive error analysis was conducted. The analysis focused on specific characteristics of the news domain outlined in Section 5.1.1. These characteristics included the prevalence of extensive translations of single English terms, complete translations of foreign abbreviations, as well as grammatical and semantic mistakes detected in one of the reference translations. In order to avoid introducing additional complexities, the evaluation did not include the assessment of free translations listed in Table 5.5. This decision was made to avoid the need for creating new references and rerunning the metric. The results of the remaining experiments can be found in Tables 5.{13,...,15}. These tables present the BLEURT-20 scores for the segments containing the aforementioned linguistic characteristics in the first reference (Ref-A). The second reference (Ref-B) does not contain these particular features. The original order of references was changed for better representation.

All evaluated machine-translation systems generated perfect outputs for the examined segments, as was indicated by the MQM human judgment scores of 100.0. To ensure a manageable evaluation process, a maximum of four systems were included per feature. If fewer systems produced perfect translations, no additional systems were added to the table, which guaranteed the reliability of the results. Therefore, it is important to note that the first semantic mistake presented in Table 5.7 was not incorporated into Table 5.15. The reason for its exclusion is that none of the machine-translation systems produced a perfect output for this particular segment.

Extensive Translations of Single English Terms

Table 5.13: BLEURT-20 scores for a segment machine translation, which was compared to the two reference translations: Ref-A, containing extended translations of the provided terms, and Ref-B, where these terms were translated with a single word. For further details, please refer to Table 5.3 and Section 5.1.1.

Source	System	Ref-A	Ref-B
to self-isolate	Facebook-AI	0.67	**0.85**
	Online-B	0.70	**0.87**
	Online-G	0.69	**0.82**
	metricsystem1	0.68	**0.89**
global	Facebook-AI	**0.62**	0.59
	Manifold	**0.55**	**0.55**
	Nemo	**0.60**	0.55
	Online-B	**0.64**	0.54

ventilators	Facebook-AI	**0.85**	0.63
	NiuTrans	**0.69**	0.60
	Online-A	**0.77**	0.63
	metricsystem1	**0.78**	0.65

Table 5.13 presents notable findings regarding the impact of longer Russian reference translations of single English terms on the performance of BLEURT-20. It can be stated that despite the existence of concise one-word translations, the metric is not significantly affected by the lengthier references. Moreover, in approximately 64% of cases, BLEURT-20 tends to favor lengthier reference translations over shorter ones. This preference is evident in its higher ranking of machine translations when comparing them to Ref-A.

Complete Translations of Abbreviations

Source	System	Ref-A	Ref-B
Labour MP	Facebook-AI	0.69	**0.78**
	metricsystem4	0.65	**0.76**
CDC	Manifold	**0.78**	0.77
	Online-G	0.76	**0.78**
	Online-W	**0.83**	0.77
	Online-Y	0.72	**0.74**
BJP MLA from Indore	Manifold	0.64	**0.70**
	NiuTrans	0.67	**0.72**
	Online-A	0.65	**0.71**
	Online-B	0.68	**0.73**
the CARES Act	Facebook-AI	0.71	**0.80**
	Manifold	0.72	**0.81**
	Nemo	0.70	**0.81**
	NiuTrans	0.73	**0.80**

Table 5.14: BLEURT-20 scores for a segment machine translation, which was compared to the two reference translations: Ref-A, containing full translations of the provided abbreviations, and Ref-B, where abbreviations were either preserved or translated in a shorter form. For further details, please refer to Table 5.4 and Section 5.1.1.

However, an interesting observation emerges from the data presented in Table 5.14. It becomes apparent that BLEURT-20 does not exhibit a strong inclination towards references that include full translations of abbreviations. In fact, in 86% of cases, the metric assigns higher scores to perfect machine translations when comparing them to Ref-B, which either preserved the abbreviations or rendered them in a shorter form.

Grammatical and Semantic Mistakes

Table 5.15: BLEURT-20 scores for a segment machine translation, which was compared to the two reference translations: Ref-A, containing the grammatical or semantic mistakes listed in Tables 5.6 and 5.7 (Section 5.1.1), and Ref-B, where no mistakes were detected.

Grammatical Mistakes

Reference	System	Ref-A	Ref-B
У нас должн**ы** быть возможность действовать быстро и решительно ...	Facebook-AI	**0.76**	0.70
	Manifold	**0.71**	0.63
	Nemo	**0.74**	0.69
	NiuTrans	0.67	**0.69**
Мы успешно снизили заболеваемость и предотвр**а-щаем** повторн**ого** рост**а** вируса ...	Facebook-AI	0.81	**0.84**
	Manifold	0.80	**0.81**
	Nemo	**0.81**	0.74
	NiuTrans	0.68	**0.76**
Уже **были были** использованы большие запасы плазмы ...	Facebook-AI	0.69	**0.70**
	Manifold	0.59	**0.67**
	Online-G	0.66	**0.71**
	metricsystem1	0.68	**0.70**

Semantic Mistakes

Член парламента от Лейбористской партии Софи Ридж заявила ...	Facebook-AI	0.69	**0.78**
	metricsystem4	0.65	**0.76**

Furthermore, the presence of grammatical or semantic mistake(s) in one of the reference translations has a significant impact on the ability of BLEURT-20 to accurately assess the quality of machine translation. As indicated by the data in Table 5.15, in approximately 72% of cases, BLEURT-20 demonstrates a preference for Ref-B, the reference that is free from mistakes, over Ref-A. This pattern is particularly visible in the case of semantic mistakes. The disparity in metric scores can vary by up to 0.11 points when a machine translation is evaluated against references of different quality.

Conclusion for Section 5.3.3

Therefore, based on the presented findings, it can be inferred that BLEURT-20 exhibits a preference for longer Russian reference translations when it comes to one-word English terms. This bias towards more extensive translations can be attributed to the embedding-based nature of the metric. However, BLEURT-20 performs much worse when faced with references containing full translations of abbreviations. Moreover, the observations underscore the importance of grammatically accurate and semantically coherent reference translations in achieving more accurate metric scores. The findings highlight the need for high-quality reference translations that align closely with the intended meaning of the source text and convey the desired linguistic nuances of the target language to ensure the effectiveness and reliability of a metric in assessing translation quality.

5.4 Results for the tedtalks Data

5.4.1 System-level Pearson's r Correlation

The analysis of system-level Pearson's r correlation between the metric and MQM scores on the tedtalks dataset presented in Table 5.16 reveals that the effectiveness of the metrics, especially traditional ones, is heavily influenced by the domain. In contrast to the newstest2021 dataset where there is a substantial performance disparity between the traditional and neural metrics, the distinction is barely noticeable for the TED talks.

Moreover, it is worth noting that all metrics exhibit a stronger correlation with the tedtalks MQM scores compared to the newstest2021 MQM human ratings. This phenomenon can be attributed to the observation that sentences in TED talks tend to be significantly shorter than those in the news domain. As a result, the evaluation metrics are more capable of accurately assessing the quality of machine translations in the tedtalks dataset. Therefore, the hypothesis suggesting a potential decline in the performance of both traditional and neural metrics in the TED talks domain, when compared to newstest2021, can be disproved.

Baselines			Ref. based		Ref. free
SacreBLEU	TER	CHRF2	BLEURT-20	COMET-MQM_2021	COMET-QE-MQM_2021
0.828	0.695	0.858	**0.867**	0.839	0.817
Official Results of WMT21 Metrics Task					
0.828	0.697	0.825	**0.868**	0.841	0.817

Table 5.16: System-level Pearson's r correlation between the metric scores and MQM human ratings for each of the implemented metrics on the tedtalks data. The best Pearson's r correlation is marked in bold. Since some metrics, such as BLEU, aim to achieve a strong positive correlation with human judgments, while error metrics, such as TER, aim for a strong negative correlation, we compare metrics via the absolute value $|r|$ of a given metric's correlation with human assessment.

BLEURT-20 once again emerges as the best-performing metric when evaluated using system-level Pearson's r, exhibiting an almost perfect positive relationship with the system-level MQM human ratings. The performance of COMET-QE-MQM_2021 can be considered inferior to that of its counterpart, COMET-MQM_2021, although the difference is not significant. Among the traditional metrics, CHRF2 achieves the highest correlation with human assessment, surpassing all neural metrics except for BLEURT-20. Nevertheless, the distinction in their performance is minimal (0.009 points). SacreBLEU also outperforms one neural metric, specifically COMET-QE-MQM_2021, securing the fourth position among all the evaluated metrics. However, TER exhibits a weaker correlation with the MQM scores compared to other metrics. Nevertheless, its performance remains remarkable, particularly when compared to the news domain.

The obtained results exhibit minimal deviations from the official results of the WMT21 Metrics Task, with the majority of metrics achieving identical scores or differing by a maximum of 0.002 points. The notable exception is the CHRF2 metric, which displays a more substantial difference. However, when compared to the results for the newstest2021 dataset, the disparities are insignificant. This can be attributed to the fact that the tedtalks dataset includes only one reference translation, which increases the possibility of replicating the scores without knowing the exact computation methodology employed.

5.4.2 Segment-level Kendall's τ Correlation

Table 5.17 presents the segment-level Kendall's τ correlation between the metrics and MQM human judgment scores for the tedtalks dataset. Unlike the system-level correlation, where BLEURT-20 showcases the highest performance among all the evaluated metrics, it yields its position to COMET-MQM_2021 in the segment-level evaluation. Furthermore, at the segment level, the disparity in performance between the traditional

and neural metrics is significantly more pronounced compared to the system level with all neural metrics outperforming the traditional ones.

It is also interesting to note that, contrary to expectations based on the system-level performances of the metrics, all metrics except for TER exhibit lower segment-level correlation in the TED talks domain compared to the news. This can potentially be attributed to the difference in the number of reference translations available in the datasets. Consequently, the number of accessible references plays a more critical role in the segment- than in the system-level evaluation of reference-based metrics. Nevertheless, this explanation fails to elucidate the reason behind the superior segment-level performance of the reference-free COMET-QE-MQM_2021 in the news domain as opposed to the TED talks.

It is also worth mentioning that the results for the evaluated metrics deviate from those obtained in the WMT21 Metrics Task. Furthermore, at the segment level, the disparities in correlation scores are more noticeable compared to the system level. However, it is important to acknowledge that the distinctions are still not as severe as those observed in the news domain. Despite the differences, the metrics can still be ranked in the same order of performance as stated by the WMT21 Metrics Task.

System	Baselines			Ref. based		Ref. free
	SacreBLEU	TER	CHRF2	BLEURT-20	COMET-MQM_2021	COMET-QE-MQM_2021
Facebook-AI	0.030	0.063	0.108	0.165	**0.182**	0.132
Manifold	0.094	0.142	0.204	0.263	**0.275**	0.209
Nemo	0.137	0.148	0.199	0.264	**0.273**	0.214
NiuTrans	0.089	0.144	0.213	0.299	**0.312**	0.253
Online-A	0.151	0.208	0.232	0.252	**0.270**	0.189
Online-B	0.057	0.075	0.133	0.203	**0.234**	0.193
Online-G	0.073	0.088	0.127	0.163	**0.193**	0.175
Online-W	0.040	0.063	0.104	**0.145**	0.141	0.124
Online-Y	0.133	0.158	0.224	0.281	**0.301**	0.270
metricsystem1	0.096	0.135	0.190	0.240	**0.258**	0.189
metricsystem2	0.129	0.135	0.202	0.235	**0.258**	0.199
metricsystem3	0.184	0.209	0.256	**0.321**	0.303	0.225
metricsystem4	0.140	0.178	0.220	**0.303**	0.281	0.201
metricsystem5	0.122	0.152	0.183	**0.289**	0.231	0.175
Average	0.105	0.136	0.185	0.244	**0.251**	0.196
Official Results of WMT21 Metrics Task						
	0.112	0.142	0.189	0.255	**0.258**	0.204

Table 5.17: Kendall's τ correlation between the metric scores and MQM human ratings for each of the implemented metrics on the tedtalks data. The best Kendall's τ correlation is marked in bold. Since some metrics, such as BLEU, aim to achieve a strong positive correlation with human judgments, while error metrics, such as TER, aim for a strong negative correlation, we compare metrics via the absolute value $|\tau|$ of a given metric's correlation with human assessment.

Interestingly enough, all the evaluated metrics except for SacreBLEU seem to be biased towards the same machine-translation system (Online-W), as also detected in the segment-level evaluation of the newstest2021 data where the correlation scores for this system were the lowest. However, similarly to the previous findings, it does not appear possible to detect the cause of this bias within the scope of this study.

5.4.3 Error Analysis

In the case of the tedtalks dataset, BLEURT-20 demonstrates superior performance only when considering its system-level correlation with the human judgments of translation quality. Despite this, it has been considered appropriate to select this metric for further error analysis due to its inherent characteristics. Specifically, BLEURT-20 relies solely on the machine-translation hypothesis and a single reference translation when producing output. This allows for a meaningful assessment of how the reference translation specifically influences the resulting metric score. When it comes to COMET-MQM_2021, which requires three inputs comprising the source sentence, candidate translation, and reference, evaluating this impact becomes more complex as it appears challenging to determine whether it is the source sentence or the reference translation that primarily affects the final score.

Table 5.18 shows the impact of completely free reference translations on the BLEURT-20 quality scores. As in Section 5.3.3, all systems presented in the table produced perfect translations of the source text. To facilitate the evaluation process, a maximum of four systems were included in the table. If there were fewer than four systems that yielded a perfect output for this particular segment, no additional systems were added to the table.

Source	Translation of Reference	System	Score
And what a wonderful thing it is.	Wonderfully thought of.	Facebook-AI	0.54
		NiuTrans	0.47
		Online-W	0.43
But basically that's what we're talking about.	But the scale is approximately this.	Facebook-AI	0.43
		Manifold	0.43
		NiuTrans	0.44
		Online-W	0.44
It's sort of like ding, ding, ding.	Just quiet clicks.	Facebook-AI	0.34
		Nemo	0.28
		NiuTrans	0.45
		Online-A	0.32
Yes, they're that, too.	Certainly.	Facebook-AI	0.76
		Online-G	0.45
		Online-W	0.62
		Online-Y	0.75

Table 5.18: BLEURT-20 scores for a segment machine translation, which was compared to a completely free Russian reference translation. For further details, please refer to Table 5.9 and Section 5.1.2.

The results indicate that completely free reference translations affect the evaluation metrics that make use of them very severely. This statement is supported by the significantly low quality scores assigned by BLEURT-20 to perfect machine translations based on such references. Furthermore, a comparison of these scores with those assigned by this metric to the newstest2021 translation hypotheses by using references that incorporate either extended source sentence translations or grammatical and semantic mistakes reveals a considerable decrease in scores. Therefore, it can be concluded that completely free reference translations have a more pronounced impact on the evaluation metrics compared to semantic mistakes detected in the newstest2021 data, where they were observed to have a substantial effect on the performance of BLEURT-20.

Chapter 6

Reference-free Metrics for Human Translators

The process of professional human translation can be extremely time-consuming, particularly when translators are confronted with lengthy texts. At the same time, state-of-the-art machine-translation systems have reached a level of proficiency where their output often exhibits an indistinguishable quality compared to that of human translation. Furthermore, machine-translation systems possess a significant speed advantage over human translators. Exploiting this advantage could be beneficial not only for large companies but for human translators as well. A translator could employ a machine-translation system to generate translations of the source text and subsequently utilize reference-free neural metrics, e.g., COMET-QE-MQM_2021, to evaluate the quality of each translated sentence. If these metrics demonstrate a reasonably reliable performance, the human translator would only need to address sentences with low metric scores without the need to read through the whole translated text. This streamlined approach would substantially facilitate the translation process resulting in an expedited production of high-quality translations.

In this chapter, we further evaluate the performance of the reference-free neural metric COMET-QE-MQM_2021, specifically with regard to its applicability for professional human translators.

6.1 Dataset Description

To assess the suitability of reference-free neural metrics in the context of professional human translation, we generated our own datasets. The primary objective behind constructing these datasets was to mimic a real-world translation scenario. Therefore, we decided to incorporate complex sentences into our test sets. This decision stemmed from the observation that professional human translators seldom deal with simple texts. In the course of their regular professional activities, translators are exposed to a broad spectrum of textual genres, encompassing but not limited to legal contracts and agreements, scientific research papers and articles, financial statements, technical manuals, product specifications, and medical texts. With this in mind, it was decided to take texts of comparable genres to build the test sets.

Our datasets comprise concise scientific articles. Each set consists of one text, either *Baby K* or *A Beautiful Mind*. Both texts including their Russian translations were taken

from the enrutext.com[1] website. This website is targeted at Russian speakers who are learning English as a foreign language. It provides texts suitable for different levels of English proficiency accompanied by corresponding Russian translations that have been carefully assessed and considered to be of very high quality. The *Baby K* dataset consists of 24 segments (single sentences), including the title, whereas *A Beautiful Mind* comprises 27 sentences. Therefore, the combined length of both datasets amounts to 51 segments. The average length of the source sentence is estimated to be 121 characters for the *Baby K* article and 132 characters for *A Beautiful Mind*. When considering the human translation, the mean length of that in the former is found to be 124 characters, while for the latter it is 139 characters. It might have been reasonable to merge the two articles into a single data file. However, it was decided to keep them separately in order to facilitate the evaluation process.

When taking into consideration the content of the chosen articles, *Baby K* combines scientific and medical information with a narrative depiction. It provides factual details regarding neural tube defects (NTDs), with a specific emphasis on a single case study involving an infant named Stephanie Keene, also known as Baby K, who was diagnosed with anencephaly, a type of NTD. The text includes medical terminology, e.g., central nervous system, cranium, fetus, cerebrum, cardiorespiratory arrest, etc., descriptions of the condition, and discussions about legal and ethical issues surrounding the care of Baby K. Overall, the text can be classified as a scientific narrative or a medical case description.

The article *A Beautiful Mind* offers an overview of game theory encompassing its definitions, concepts, and practical applications. The text explains the underlying principles behind cooperative and non-cooperative games, including the notion of Nash equilibrium, and presents an example of the Prisoner's Dilemma. It also mentions the historical context of game theory and highlights the achievements of mathematician John Nash, the author of Nash equilibrium, including his Nobel Prize in Economic Sciences and Abel Prize in mathematics. Therefore, the article can be regarded as a scientific narrative with a focus on economics and legal issues, which can be stated from the use of the following terms: coalition, optimal outcome, maximized profit, finite game, to interrogate, etc.

Given the research objective of assessing reference-free neural metrics in their ability to differentiate between inferior and superior translations, the datasets were also supplemented with poorer translations of the source sentences. These translations were obtained with opus-mt-en-ru[2] (Tiedemann and Thottingal, 2020), which is a pre-trained machine-translation system specifically designed to convert English text into Russian. It is trained on publicly available parallel corpora collected in the large bitext Open Parallel Universal Sampler (OPUS) repository[3]. The system utilizes state-of-the-art transformer-based neural machine translation (NMT) to generate translations between the two languages.

The structure of the final datasets is presented in Table 6.1.

[1] https://enrutext.com/
[2] https://huggingface.co/Helsinki-NLP/opus-mt-en-ru
[3] https://opus.nlpl.eu/

Source	Human Translation	Machine Translation
Neural tube defects affect either the development of the brain, or spine, or both.	Дефекты нервной трубки воздействуют на развитие либо головного мозга, либо спинного мозга, либо обоих участков ЦНС одновременно.	Дефекты нервной трубы влияют либо на развитие мозга, либо позвоночника, либо и того, и другого.
Each prisoner is sentenced to one year in prison.	Каждый заключённый приговорён к одному году тюрьмы.	Каждый заключенный приговаривается к одному году тюремного заключения.

Table 6.1: Structure of the *Baby K* and *A Beautiful Mind* datasets. The first example is taken from *Baby K*, whereas the second instance is derived from *A Beautiful Mind*.

6.2 Implementation

An evaluation was conducted using the reference-free COMET-QE-MQM_2021. This involved requesting the metric to evaluate both perfect human translations and lower-quality machine translations. Through this process, we aimed to determine whether the metric can effectively differentiate between human and machine translations and accurately evaluate their quality. By drawing precise conclusions from these evaluations, we can make judgments regarding the applicability of reference-free neural metrics for professional human translators.

6.3 Results

The lack of a well-defined score range in COMET-QE-MQM_2021 poses a challenge when evaluating translations. The metric potentially generates quality scores ranging from 0 to 0.2, with 0 representing the lowest possible translation quality and 0.2 indicating the highest quality.* Nevertheless, the absence of documented information and supporting evidence regarding this score range complicates the evaluation of a single machine-translation system (or human translation), particularly when it is not being compared to some other system (or human translation).

When evaluating the results, the COMET-QE-MQM_2021 metric generally assigned higher quality scores to human translations compared to machine translations generated by opus-mt-en-ru. This is evidenced by the higher system-level scores obtained for the former, specifically 0.124 for both the *Baby K* and *A Beautiful Mind* datasets, in contrast to the system-level scores of 0.118 and 0.119, respectively, given to the machine-translation system.

6.3.1 Cases of Metric Disregarding Human Translations

Out of the 51 segments evaluated, COMET-QE-MQM_2021 determined that the machine translation outperformed the human translation in 18 segments. Nevertheless,

*In contrast to other COMET metrics, which yield scores ranging from 0 to 1, the scoring range varies for COMET-MQM_2021 and COMET-QE-MQM_2021. This distinction was further evidenced through an examination of the official metric scores submitted for the WMT21 Metrics Task. The official scores can be accessed at the following link: https://github.com/WMT-Metrics-task/wmt21-metrics-data/tree/main.

the difference is typically insignificant and varies from 0.001 to 0.008 points. The two sentences with the larger difference in metric segment scores between the two available translations are presented in Table 6.2:

		Score
Source	One of the most common types of birth defects that afflict yet unborn children are referred to as neural tube defects NTDs).	
Human Translation	Один из наиболее распространённых пороков развития – так называемые дефекты нервной трубки (ДНТ, ДЗНТ, дефекты заращения нервной трубки).	0.098
Machine Translation	Один из наиболее распространенных видов врожденных дефектов, от которых страдают еще не родившиеся дети, называется дефектами нервной трубки (НТР).	0.136
Source	Game theory can be explained broadly as a study of behaviour of rational beings in cooperative and non-cooperative and non-cooperative decision making.	
Human Translation	Теорию игр можно в широком смысле объяснить как учение о поведении рациональный существ в кооперативном и некооперативном принятии решений.	0.121
Machine Translation	Теория игр может быть широко объяснена как исследование поведения рациональных существ в процессе принятия решений на основе сотрудничества и без сотрудничества.	0.137

Table 6.2: Segments with a larger than 0.008 points difference in COMET-QE-MQM_2021 quality scores between their human and machine translations. The 1st segment is the 3rd sentence of the *Baby K* dataset. The 2nd segment is the 2nd sentence of *A Beautiful Mind*.

In the first example provided in Table 6.2, the metric encountered confusion due to an extensive translation of the term *NTDs* into Russian, including variations such as *ДНТ*, *ДЗНТ*, and *дефекты заращения нервной трубки*. These three translations essentially represent different interpretations of the same birth defect. Such a translation approach is commonly employed when there are multiple possible translations of the same scientific term available. Subsequently, it is considered more professional to list all these translations upon the initial mention of the term, rather than providing only one variant. Moreover, the human translation did not include the word *types* and neglected the phrase *that afflict yet unborn children*. While these omissions do not significantly affect the content of the translation and make it more natural-sounding in Russian, they could potentially have a negative impact on the resulting quality score assigned by COMET-QE-MQM_2021 to the entire human translation.

The machine translation of the same source segment is grammatically correct and includes all the information present in the initial sentence. However, it is less formal and does not align well with the writing style commonly employed in Russian scientific articles. Additionally, opus-mt-en-ru incorrectly rendered the term *NTDs* as *НТР* instead of *ДНТ*. Since this particular error is rather serious, the entire machine translation can be regarded as significantly inferior to the human translation, which is, however, not reflected in the metric scores.

In contrast to the first example, the human translation of the second instance listed in Table 6.2 captures all the concepts present in the source sentence, thereby reflecting its complete meaning. The same can be said for the machine translation. However, it

once again lacks the desired level of formality and contains some minor errors. For instance, the word *broadly* is translated as *широко* instead of *в широком смысле*, which is a correct translation of the word itself. Nevertheless, this translation is not suitable in the given context. Furthermore, the phrase *на основе сотрудничества и без сотрудничества* is grammatically correct but sounds somewhat too colloquial, more suitable for spoken language rather than written form. Consequently, the machine translation of the second example can also be considered inferior to the human translation. However, these flaws were not detected by the metric either.

Nevertheless, it should be noted that not all of the 18 instances where COMET-QE-MQM_2021 assigned higher scores to a machine translation than to a human translation can be considered incorrect. Among these instances, there are 3 cases where the opus-mt-en-ru translation can be viewed as somewhat superior to the human translation. In these cases, the machine translation captures the meaning and structure of the source sentence with greater accuracy compared to the human translation, while still maintaining a suitable language style. These specific instances are listed in Appendix C.1, Table C.1.

6.3.2 Cases of Metric Disregarding Machine Translations

Based on this observation, a decision was made to investigate the presence of other instances where machine translation surpasses human translation, yet its excellence remains undetected by COMET-QE-MQM_2021. Given the high quality of the human translations in both our datasets, determining whether the machine translations outperform them poses considerable challenges as the best machine translations are often on a par with human translations. We identified a total of 5 cases, in which COMET-QE-MQM_2021 assigned lower scores to machine translations despite their equivalence in quality to human translations. Table 6.3 presents the instances featuring the largest disparities in scores. Other instances are listed in Appendix C.1, Table C.2

Table 6.3: Segments where the machine translation is equivalent in quality to the human translation, yet COMET-QE-MQM_2021 was not able to detect it. The 1st segment is the 26th sentence of *A Beautiful Mind*. The 2nd segment is the 4th sentence of the same dataset.

		Score
Source	Finding it might be hard, but the willingness to do that, perhaps, can make us able to stop the wars and other major threats to our society.	
Human Translation	Найти такое решение может быть сложно, но открытость к возможности сделать это, может быть, даст нам возможность остановить войны и другие глобальные угрозы нашему обществу.	0.143
Machine Translation	Найти его может быть трудно, но готовность сделать это, возможно, позволит нам остановить войны и другие серьезные угрозы нашему обществу.	0.132

Source	It's a relatively new field of science that emerged in the second half of the 20th century.	
Human Translation	Это достаточно новая отрасль науки, появившаяся во второй половине XX века.	0.178

Machine Translation	Это относительно новая область науки, которая появилась во второй половине XX века.	0.170

In the first example presented in Table 6.3, the human translation can be interpreted as a free translation of the source sentence. It introduces the word *решение* (decision) in the phrase *Найти такое решение может быть сложно* (Finding this decision might be hard), which is not present in the initial sentence and can only be inferred by considering the context of previous segments. Moreover, the phrase *the willingness to do that* is translated as *открытость к возможности сделать это* (the openness to the opportunities to do that) in the human translation, which cannot be considered a common translation of this phrase. On the other hand, the machine translation of this sentence adheres more closely to the source sentence without introducing any concepts that deviate from it.

In the second example, the disparity between the human and machine translation is very insignificant. The translations employ different terms, namely *достаточно* and *относительно* for *relatively*, as well as *отрасль* and *область* for *field*. However, both word choices are suitable in the given context. Furthermore, in the human translation, the phrase *that emerged* is rendered as a present participle, *появившаяся* (emerging), whereas the machine translation retains the original verb form.

Therefore, while both the human and machine translations show a high level of quality in the given instances, it is difficult to explain why the metric assigned a higher score to the human translations compared to the machine translations, considering all the aforementioned aspects of these particular cases.

6.3.3 Conclusion for Section 6.3

The COMET-QE-MQM_2021 metric has demonstrated a tendency to assign relatively low scores to translations of high quality. Apart from the aforementioned examples, 2 specific occurrences were identified where the metric assigned a score lower than 0.090 to a human translation. These instances are outlined in Appendix C.1, Table C.3. However, determining the underlying cause of this phenomenon presents a significant challenge as it is not correlated with segment length or the utilization of free translation.

Furthermore, the metric exhibits a prominent inclination to give excessively high scores to poor translations. Nevertheless, COMET-QE-MQM_2021 is not consistently prone to this behavior as it occasionally exhibits the ability to distinguish a poor translation from a good one. However, the frequency of errors made by the metric in this regard, specifically 15 out of 51 segments, can lead to even more severe complications in translation evaluation compared to its tendency to assign low scores to good translations. The reason behind this lies in the fact that translations with high quality scores are unlikely to be examined by translators. Consequently, poor translations may go unnoticed. On the other hand, if translators do examine these segments, the evaluation of translation using reference-free neural metrics becomes meaningless.

Moreover, apart from the aforementioned 15 instances, we identified 2 cases in our datasets where the metric assigned a score of 0.136 or higher to an absolutely unacceptable machine translation. These specific mistakes are outlined in Table 6.4, serving as further evidence of the unsuitability of COMET-QE-MQM_2021 for professional human translators.

		Score
Source	Anencephaly is a NTD that in broadest terms means the complete absence of the cerebrum, the largest part of the brain responsible for senses and cognition.	
Machine Translation	Annestphaly — это NTD, что в самом широком смысле означает полное отсутствие церебральной мышцы, самой большой части мозга, отвечающей за чувства и сознание.	0.140
Source	Abortion is strongly encouraged when anencephaly is detected via ultrasound.	
Machine Translation	Аборты активно поощряются в тех случаях, когда анэнцефалии обнаруживаются с помощью ультразвука.	0.139

Table 6.4: Segments where COMET-QE-MQM_2021 assigned excessively high scores to poor machine translations.

In the first translation from opus-mt-en-ru, the term *cerebrum* was incorrectly translated as *церебральной мышцы* (cerebral muscle), which not only constitutes an inaccurate translation in this context but also represents a nonexistent term. The second instance contains an erroneous grammatical rendering of the source sentence. Specifically, the terms *abortion* and *anencephaly* were translated in their plural forms as *аборты* (abortions) and *анэнцефалии* (anencephalies), thereby failing to accurately convey the intended meaning and adhere to the appropriate language style.

Therefore, it can be concluded that the current state of development of COMET-QE-MQM_2021 does not enable it to differentiate between human and machine translations effectively. While the metric shows promising results in system-level evaluation, its performance at the segment level still requires improvement. A significant drawback of the metric is its disregard for language style, which is a crucial aspect of a good translation and must not be neglected. Typically, the metric emphasizes the accuracy of word-to-word translation without adequately considering the context of the entire segment. Based on these findings, it can be stated that the reference-free COMET-QE-MQM_2021 metric is not suitable for professional human translators as it would not facilitate the translation process but instead introduce complications.

Chapter 7

Discussion

While replicating the scores for the implemented neural metrics may be challenging due to the inherent stochastic nature of neural networks, the scores for the traditional metrics can be fully reproduced. However, our research encountered a problem regarding the computation of scores for the traditional metrics at the WMT21 Metrics Task. Specifically, it remains unclear how exactly the scores were calculated. In an attempt to address this issue, we conducted experiments on the newstest2021 dataset. Particularly, we tried computing the scores by considering only the first reference translation. We also explored an alternative approach where the second reference was treated as a candidate translation. However, neither of these methods yielded an improved correlation between the metric scores and MQM human judgments. By exploring other input combinations, there is a possibility of achieving identical results for the traditional metrics while bringing the scores for the neural metrics closer to the ones presented by the WMT21 Metrics Task. However, such an approach lacks empirical evidence as it only involves systematically exploring the remaining variants without any hypotheses set.

Additionally, it was noted that the performance of the traditional metrics is heavily influenced by the specific Python libraries employed for their computation. For instance, in our case, all the traditional metric scores were initially calculated using the SacreBLEU[1] Python library, which resulted in a much more substantial disparity between the obtained results and the official results of the WMT21 Metrics Task.

This variation in performance may be attributed to the fact that the tokenizers utilized by SacreBLEU may differ from those used in TorchMetrics. Since traditional metrics solely rely on counting the overlap in the number of token or character n-grams between a machine translation and its reference translation, the way the segments are tokenized can significantly impact the resulting metric scores and their correlation with MQM human judgments. Nevertheless, even taking into account this nuance, such a great performance disparity between the traditional metrics computed with different Python libraries was not expected.

Besides this, due to time constraints, it was not possible to identify the underlying cause of the bias observed in certain metrics towards specific machine-translation systems. The task itself presents significant challenges as it entails carefully assessing the segment-level metric scores, comparing them with the corresponding human judgment scores, identifying instances where there is a substantial disparity between the metric and human evaluations, and analyzing the linguistic characteristics of the machine translations to potentially identify any recurring patterns. Conducting such an

[1] https://pypi.org/project/sacrebleu/

analysis would involve evaluating 2,108 segments in the newstest2021 dataset and 2,048 segments in the tedtalks dataset. However, there is no guarantee that these experiments would produce definitive results.

The impact of partially free reference translations remains to be investigated as well. Since this involves editing reference translations to align with the literal meaning of the source text, these experiments were not included in our study. However, considering the relatively frequent utilization of partially free reference translations across various domains, it would be beneficial to explore their influence on the performance of the evaluated metrics.

Moreover, in addition to the extensive translations of single English terms, complete translations of foreign abbreviations, completely free reference translations, as well as grammatical and semantic mistakes detected in one of the reference translations within the newstest2021 dataset, it would be valuable to assess the impact of other linguistic features, which were not prominently represented in the given datasets, on the performance of the evaluated metrics. These features may include the conversion of direct speech into an indirect expression and antonymic translation, which involves replacing a word or phrase in the source text with its antonym in the target language. For instance, in the case of translating the phrase *not good*, an antonymic translation would substitute it with *bad*. Exploring the impact of these linguistic features can provide further insights into how different translation techniques and strategies utilized to produce reference translations affect the effectiveness and accuracy of the evaluated metrics.

Additionally, it is important to note that our main experiments focused solely on examining the impact of peculiarities and errors present within reference translations on the performance of the metrics. We did not specifically investigate whether different metrics have the capability to identify significant errors within the machine translations themselves. This evaluation was exclusively carried out to assess the suitability of COMET-QE-MQM_2021 for professional human translators. However, the ability of the evaluated metrics to identify errors in machine translations has already been addressed by the organizers of the WMT21 Metrics Task through the use of various challenge sets. Nevertheless, it would be worthwhile to conduct similar experiments by employing alternative error typologies.

Although the findings from evaluating the applicability of reference-free neural metrics for professional human translators have provided informative results, it is crucial not to rely solely on them due to the limited assessment conducted on a single reference-free neural metric, specifically COMET-QE-MQM_2021. While this metric may be considered nearly state-of-the-art in machine-translation evaluation, it is essential to reinforce the results by implementing additional metrics such as OpenKiwi-MQM and others. Expanding the evaluation to encompass a variety of metrics will contribute to a more comprehensive understanding of the performance and applicability of reference-free neural metrics in the context of professional human translation.

Chapter 8

Conclusion

The research outcomes demonstrate that the performance of the metrics is greatly influenced by the domain, such as while neural metrics generally exhibit superior system-level performance over traditional ones in the news domain, their system-level performance for the TED talks may be considered equal. This discrepancy can be attributed to the fact that the mean segment length in TED talks is approximately 7 times shorter than that in news. Moreover, all the metrics, particularly neural ones, show poorer performance at the segment level for TED talks compared to the news. This phenomenon could potentially be explained by the presence of only one reference translation in the dataset or the excessive utilization of partially or completely free reference translations.

The conclusions of the study indicate that character-based traditional metrics, specifically CHRF2, exhibit favorable system-level performance. On the other hand, word-level metrics, i.e., TER and SacreBLEU, demonstrate the weakest overall results among all the evaluated metrics. Although SacreBLEU shows a slightly improved correlation with the MQM human judgments compared to TER, the organizers of the WMT22 Metrics Task (Freitag et al., 2022) strongly advise against utilizing BLEU or SacreBLEU any further, emphasizing this recommendation in the title of the paper presenting the official results: *Stop Using BLEU – Neural Metrics Are Better and More Robust.*

The performance of neural metrics significantly surpasses that of traditional ones at the segment level. As a result, prioritizing their utilization is strongly recommended, especially when conducting a more detailed and fine-grained evaluation.

The results also indicate that neural metrics demonstrate equal reproducibility compared to traditional ones, which can be attributed to the absence of additional adjustable parameters in the former. In contrast, traditional metrics have various parameters, and their usage is complicated by the availability of different Python libraries, each potentially implementing distinct tokenization algorithms. These factors contribute to the challenges involved in reproducing specific scores when utilizing traditional metrics. Therefore, in order to fully replicate the results for these metrics, it is important to have access to the precise methodology employed for their computation.

Our scores for the metrics can be called reproducible only at the segment level since the discrepancies between the obtained scores and the official results of the WMT21 Metrics Task are minimal. In the case of newstest2021, the majority of metrics maintain the same performance ranking despite the slight variations in scores. Similarly, in the TED talks domain, all metrics align with the performance order determined by the WMT21 Metrics Task. However, the situation differs when it comes to system-level evaluation, particularly for the traditional metrics in the news domain, where the

reproducibility of metric scores becomes more challenging.

The findings of the WMT21 Metrics Task, which suggest a broader range of metrics achieving high-level performance at the system level, are supported. However, the inclusion of surface-level baselines (SacreBLEU and TER) among the top-performing metrics was not verified. Moreover, contrary to the conclusion drawn in the WMT21 Metrics Task, the observation that COMET-QE-MQM_2021 demonstrates strong overall performance but performs poorly at the segment level only partially applies to the English→Russian language pair as the metric demonstrates quite good performance at both levels when compared to other evaluated neural metrics. Consequently, it is not possible to conclusively state that the results of the WMT21 Metrics Task can be fully reproduced as the findings are only partially confirmed.

Additional findings from our research verify that while neural metrics exhibit state-of-the-art performance, particularly at the system level, they come with a significant computational time requirement. Neural metrics run approximately 76 times slower compared to traditional ones. This characteristic makes the majority of neural metrics, especially reference-based ones, less suitable for evaluation during system development and more appropriate for the final assessment of the resulting machine-translation system. However, this disparity may not apply to reference-free neural metrics, which do not require the creation of reference translations, unlike other metrics. This aspect leads to substantial time and cost savings, making reference-free neural metrics highly desirable for machine-translation evaluation.

Nevertheless, the relevance of reference-free neural metrics for professional human translators raises doubts. When using COMET-QE-MQM_2021 for evaluation, the findings strongly suggest that this particular reference-free neural metric is not suitable for professional human translators as it frequently introduced confusion rather than facilitated the translation process. Specifically, the metric demonstrated a significant inclination to assign high quality scores to poor translations. For this reason, the evaluation conducted by COMET-QE-MQM_2021 cannot be called reliable.

Furthermore, an unexpected discovery was made regarding the potential bias of certain metrics towards specific machine-translation systems since all the neural metrics exhibited the lowest correlation with the MQM human judgments for one particular machine-translation system. However, due to the limitations of the study, it was not feasible to determine the underlying cause of this bias.

Upon examining the linguistic features that significantly influence reference-based neural metrics, it was found that, apart from the mean segment length, the presence of a completely free reference translation has the most pronounced impact on the evaluation metrics. This impact even surpasses the influence of semantic mistakes within the reference translation. On the other hand, grammatical mistakes do not significantly affect the performance of the metrics. Surprisingly, extensive translations of single English terms in the reference translation, despite the availability of a one-word substitute, are generally favored. However, the metrics do not demonstrate a strong inclination towards references that include full translations of abbreviations.

Appendix A

Metrics

A.1 RemBERT details

Hyperparameter	RemBERT
Number of layers	32
Hidden size	1152
Vocabulary size	250,000
Input embedding dimension	256
Output embedding dimension	1536
Number of attention heads	18
Attention head dimension	64
Dropout	0
Learning rate	0.0002
Batch size	2048
Train steps	1.76M
Adam $\beta 1$	0.9
Adam $\beta 2$	0.999
Adam ϵ	10^{-6}
Weight decay	0.01
Gradient clipping norm	1
Warmup steps	15000

Table A.1: Hyperparameters for RemBERT architecture and pre-training used in the BLEURT-20 metric.

A.2 XTREME tasks

XTREME includes the following datasets: The Cross-lingual Natural Language Inference (XNLI; Conneau et al. 2018) corpus, the Cross-lingual Paraphrase Adversaries from Word Scrambling (PAWS-X; Yang et al. 2019) dataset, part-of-speech (POS) tagging data from the Universal Dependencies v2.5 Nivre et al. 2018) treebanks, the Wikiann Pan et al. 2017) dataset for named entity recognition (NER), the Cross-lingual Question Answering Dataset (XQuAD; Artetxe et al. 2020), the Multilingual Question Answering (MLQA; Lewis et al. 2020) dataset, the gold passage version of the Typologically Diverse Question Answering (TyDi QA; Clark et al. 2020) dataset, data from the third shared task of the workshop on Building and Using Parallel Corpora (BUCC; Zweigenbaum

et al. 2018), and the Tatoeba dataset (Artetxe and Schwenk, 2019).

A.3 XLM-RoBERTa Details

System	batch_size	2
Encoder	hidden_size	1024
Decoder	bottleneck_size	1024
	dropout	0.05
	hidden_size	1024
Optimizer	class_name	adam
	encoder_learning_rate	0.0001
	learning_rate_decay	1.0
	learning_rate_decay_start	0
	learning_rate	0.0001
Trainer	training_steps	2180
	early_stop_patience	10
	validation_steps	0.5
	gradient_accumulation_steps	4
	gradient_max_norm	1.0

Table A.2: Hyperparameters for XLM-RoBERTa architecture used in OpenKiwi-MQM.

Appendix B

Comparative Analysis of MT Evaluation Metrics

B.1 Dataset Details

MT System	newstest2021			tedtalks
	MQM	raw DA	z-normalized DA	MQM
Facebook-AI	527	911	911	512
Manifold	527	948	948	512
Nemo	527	924	924	512
NiuTrans	527	882	882	512
Online-A	527	957	957	512
Online-B	527	986	986	512
Online-G	527	892	892	512
Online-W	527	989	989	512
Online-Y	527	996	996	512
metricsystem1	527	-	-	512
metricsystem2	527	-	-	512
metricsystem3	527	-	-	512
metricsystem4	527	-	-	512
metricsystem5	527	-	-	512
	out of 1002			out of 512

Table B.1: Number of annotations for the English→Russian language pair in the newstest2021 and tedtalks datasets per machine-translation (MT) system and annotation type.

B.2 Linguistic Features of newstest2021 Reference Translations

Table B.2: Examples of test segments from the newstest2021 dataset, in which one of the references is a free translation.

Source	Reference
Dominic Raab: Government can't make apologies for Spain quarantine decision	Доминик Рааб: Правительство не может извиниться за решение о карантине *по прибытии из* Испании

Explanation: The reference contains: *upon arrival from Spain* as a translation of *for Spain*.	
Dominic Raab has defended the Government's decision to re-introduce quarantine measures on Spain at short notice.	Доминик Рааб выступил в защиту решения правительства вновь срочно ввести карантин *по прибытии из* Испании.
Explanation: The reference contains: *upon arrival from Spain* as a translation of *on Spain*.	
Shadow Health Secretary Jonathan Ashworth condemned the Government for its "frankly shambolic" handling of the measure.	Заместитель министра здравоохранения Джонатан Эшворт осудил Правительство за «откровенно безалаберный» *подход к введению ограничений*.
Explanation: The reference contains: *approach to imposing restrictions* as a translation of *handling of the measure*.	
He said Downing Street's sudden decision had left holidaymakers "confused and distressed."	Он сказал, что из-за внезапного решения Даннинг-стрит отдыхающие оказались "*застигнуты врасплох в сложной ситуации*".
Explanation: The reference contains: *caught by surprise in a difficult situation* as a translation of *confused and distressed*.	
It also showed that the much-hyped drug hydroxychloroquine - as well as the combined therapy of the drugs lopinavir and ritonavir - had no effect in saving patients' lives.	*Программа* также показала, что ни нашумевший гидроксихлорохин, ни комбинация лопинавира с ритонавиром никак не помогают спасать жизни.
Explanation: The reference contains: *programme* as a translation of *it*.	

B.3 Results for the newstest2021 Data

Pearson's r system-level correlation between the metrics, raw DA, and per-rater z-normalized DA human ratings was computed only for the annotated machine-translation systems. The system-level metric scores were obtained by averaging all segment-level metric scores across all segments, regardless of the annotations available.

Baselines			Ref. based		Ref. free
SacreBLEU	TER	CHRF2	BLEURT-20	COMET-MQM_2021	COMET-QE-MQM_2021
0.965	0.830	0.934	0.966	0.968	**0.971**

Table B.3: System-level Pearson's r correlation between the metric scores and raw DA human ratings for each of the implemented metrics on the newstest2021 data. The best Pearson's r correlation is marked in bold. Since some metrics, such as BLEU, aim to achieve a strong positive correlation with human judgments, while error metrics, such as TER, aim for a strong negative correlation, we compare metrics via the absolute value $|r|$ of a given metric's correlation with human assessment.

Baselines			Ref. based		Ref. free
SacreBLEU	TER	CHRF2	BLEURT-20	COMET-MQM_2021	COMET-QE-MQM_2021
0.927	0.821	0.930	0.959	0.965	**0.966**

Table B.4: System-level Pearson's r correlation between the metric scores and per-rater z-normalized DA human ratings for each of the implemented metrics on the newstest2021 data. The best Pearson's r correlation is marked in bold. Since some metrics, such as BLEU, aim to achieve a strong positive correlation with human judgments, while error metrics, such as TER, aim for a strong negative correlation, we compare metrics via the absolute value $|r|$ of a given metric's correlation with human assessment.

System	Baselines			Ref. based		Ref. free
	SacreBLEU	TER	CHRF2	BLEURT-20	COMET-MQM_2021	COMET-QE-MQM_2021
Facebook-AI	0.063	0.107	0.115	0.194	**0.202**	0.165
Manifold	0.168	0.186	0.217	**0.307**	0.303	0.287
Nemo	0.154	0.152	0.186	0.289	**0.294**	0.279
NiuTrans	0.147	0.189	0.178	0.271	**0.287**	0.267
Online-A	0.113	0.138	0.167	0.293	**0.299**	0.286
Online-B	0.070	0.098	0.098	0.200	**0.229**	0.209
Online-G	0.141	0.190	0.192	**0.279**	0.257	0.230
Online-W	0.090	0.113	0.111	0.192	**0.209**	0.204
Online-Y	0.097	0.120	0.133	0.318	**0.325**	0.286
Average	0.116	0.144	0.155	0.260	**0.267**	0.246

Table B.5: Segment-level Kendall's τ correlation between the metric scores and raw DA human ratings for each of the implemented metrics on the newstest2021 data. The best Kendall's τ correlation is marked in bold. Since some metrics, such as BLEU, aim to achieve a strong positive correlation with human judgments, while error metrics, such as TER, aim for a strong negative correlation, we compare metrics via the absolute value $|\tau|$ of a given metric's correlation with human assessment.

System	Baselines			Ref. based		Ref. free
	SacreBLEU	TER	CHRF2	BLEURT-20	COMET-MQM_2021	COMET-QE-MQM_2021
Facebook-AI	0.107	0.147	0.163	**0.244**	0.230	0.183
Manifold	0.172	0.196	0.223	0.348	**0.358**	0.333
Nemo	0.124	0.130	0.159	0.297	**0.309**	0.295
NiuTrans	0.130	0.180	0.185	**0.326**	0.316	0.287
Online-A	0.140	0.183	0.203	0.341	**0.351**	0.338
Online-B	0.114	0.155	0.160	0.307	**0.323**	0.299
Online-G	0.136	0.180	0.198	**0.297**	0.285	0.256
Online-W	0.116	0.127	0.146	0.265	0.281	**0.282**
Online-Y	0.113	0.154	0.161	**0.372**	0.361	0.335
Average	0.128	0.161	0.178	0.311	**0.313**	0.290

Table B.6: Segment-level Kendall's τ correlation between the metric scores and per-rater z-normalized DA human ratings for each of the implemented metrics on the newstest2021 data. The best Kendall's τ correlation is marked in bold. Since some metrics, such as BLEU, aim to achieve a strong positive correlation with human judgments, while error metrics, such as TER, aim for a strong negative correlation, we compare metrics via the absolute value $|\tau|$ of a given metric's correlation with human assessment.

Appendix C

Reference-free Metrics for Human Translators

C.1 Results

		Score
Source	Nevertheless, there were some cases of anencephaly that truly stood out from the rest.	
Human Translation	Тем не менее, были случаи анэнцефалии, не похожие на другие.	0.114
Machine Translation	Тем не менее были случаи анэнцефалии, которые действительно отличались от остальных.	0.119

Source	Mathematician John Forbes Nash, who was an author of the concept, proved that this equilibrium is possible to find for any finite game.	
Human Translation	Математик Джон Форбс Нэш, автор этой идеи, доказал, что такое равновесие возможно найти для каждой конечной игры.	0.114
Machine Translation	Математик Джон Форбс Нэш, который был автором этой концепции, доказал, что это равновесие можно найти для любой конечной игры.	0.117

Source	What is the optimal course of action for each prisoner?	
Human Translation	Каким будет оптимальный курс действий для каждого заключённого?	0.153
Machine Translation	Каков оптимальный курс действий для каждого заключенного?	0.156

Table C.1: Instances of COMET-QE-MQM_2021 assigning higher scores to machine translation than to human translation of the same source sentence. In these cases, the machine translation can actually be viewed as somewhat superior to the human translation.

		Score
Source	Her heart had stopped on April 5, 1995.	
Human Translation	Она умерла от остановки сердца 5 апреля 1995 года.	0.155
Machine Translation	Ее сердце остановилось 5 апреля 1995 года.	0.150

Source	The causes of the condition are still unclear, but it is speculated that it can be triggered by a folic acid deficiency and certain types of diabetes in pregnant women.	
Human Translation	Причины возникновения патологии пока что неясны, но считается, что она может возникнуть на фоне дефицита фолиевой кислоты и некоторых типов диабета у беременных.	0.143
Machine Translation	Причины этого заболевания по-прежнему неясны, однако предполагается, что оно может быть вызвано дефицитом фолиевой кислоты и некоторыми видами диабета у беременных женщин.	0.140

Source	However, if both prisoners testify against each other, both of them will get a harder sentence, and both will serve 2 years in prison.	
Human Translation	Однако, если оба заключённых будут свидетельствовать друг против друга, оба получат более тяжелое наказание – по 2 года в тюрьме.	0.124
Machine Translation	Однако, если оба заключенных будут давать показания друг против друга, они оба получат более суровое наказание, и оба будут отбывать два года тюремного заключения.	0.121

Table C.2: Segments where the machine translation is equivalent in quality to the human translation, yet COMET-QE-MQM_2021 was not able to detect it.

		Score
Source	Keeping her heart beating had cost over 500,000$, a sum, as some would argue, that could've been spent on research aimed to prevent NTDs or, possibly, treatment of other newborn children.	
Human Translation	Биение её сердца стоило более 500,000$ – сумма, которую, как могли бы сказать некоторые, можно было бы потратить на исследования по предотвращению ДНТ или, возможно, на лечение других новорожденных детей.	0.085

Source	Anencephaly is a NTD that in broadest terms means the complete absence of the cerebrum, the largest part of the brain responsible for senses and cognition.	
Human Translation	Анэнцефалия – ДНТ, в наиболее общих понятиях характеризуемый как абсолютное отсутствие конечного мозга, части головного мозга, ответственной за чувства и сознание.	0.088

Table C.3: Segments where COMET-QE-MQM_2021 assigned a score lower than 0.090 to a high-quality human translation.

Bibliography

A. Agarwal and A. Lavie. Meteor, M-BLEU and M-TER: Evaluation metrics for high correlation with human rankings of machine translation output. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 115–118, Columbus, Ohio, June 2008. Association for Computational Linguistics.

F. Akhbardeh, A. Arkhangorodsky, M. Biesialska, O. Bojar, R. Chatterjee, V. Chaudhary, M. R. Costa-jussa, C. España-Bonet, A. Fan, C. Federmann, M. Freitag, Y. Graham, R. Grundkiewicz, B. Haddow, L. Harter, K. Heafield, C. Homan, M. Huck, K. Amponsah-Kaakyire, and et al. Findings of the 2021 conference on machine translation (WMT21). In *Proceedings of the Sixth Conference on Machine Translation*, pages 1–88. Association for Computational Linguistics, 2021.

A. Ali and S. Renals. Word error rate estimation for speech recognition: e-wer. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 20–24, Melbourne, Australia, 2018. Association for Computational Linguistics.

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

M. Artetxe, S. Ruder, and D. Yogatama. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020)*, pages 4623–4637, Online, July 5–10 2020. Association for Computational Linguistics.

Y. Attali and J. Burstein. Automated essay scoring with e-rater v.2.0. *Journal of Technology, Learning, and Assessment*, 4(3):159–174, 2006.

S. Banerjee and A. 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, 2005. Association for Computational Linguistics.

R. Bawden, B. Zhang, L. Yankovskaya, A. Tattar, and M. Post. A study in improving BLEU reference coverage with diverse automatic paraphrasing. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 918–932. Association for Computational Linguistics, 2020.

O. Bojar, C. Buck, C. Federmann, B. Haddow, P. Koehn, J. Leveling, C. Monz, P. Pecina, M. Post, H. Saint-Amand, R. Soricut, L. Specia, and A. Tamchyna. Findings of the 2014 workshop on statistical machine translation. In *Proceedings of the*

9th Workshop on Statistical Machine Translation (WMT-14), pages 12–58, Baltimore, Maryland, USA, June 2014. Association for Computational Linguistics.

O. Bojar, R. Chatterjee, C. Federmann, B. Haddow, M. Huck, C. Hokamp, P. Koehn, V. Logacheva, C. Monz, M. Negri, M. Post, C. Scarton, L. Specia, and M. Turchi. Findings of the 2015 workshop on statistical machine translation. In *Proceedings of the 10th Workshop on Statistical Machine Translation (WMT-15)*, pages 1–46, Lisbon, Portugal, September 2015. Association for Computational Linguistics.

O. Bojar, Y. Graham, A. Kamran, and M. Stanojevic. 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 11-12 2016. Association for Computational Linguistics.

C. Callison-Burch, M. Osborne, and P. Koehn. Re-evaluating the role of bleu in machine translation research. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 249–256, Trento, Italy, 2006. Association for Computational Linguistics.

C. Callison-Burch, C. Fordyce, P. Koehn, C. Monz, and J. 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.

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

H. W. Chung, T. Fevry, H. Tsai, M. Johnson, and S. Ruder. Rethinking embedding coupling in pre-trained language models. In *International Conference on Learning Representations*, 2021.

J. H. Clark, E. Choi, M. Collins, D. Garrette, T. Kwiatkowski, V. Nikolaev, and J. Palomaki. Tydi qa: A benchmark for information-seeking question answering in typologically diverse languages. *Transactions of the Association for Computational Linguistics*, 8:454–470, 2020.

A. Conneau, R. Rinott, G. Lample, A. Williams, S. Bowman, H. Schwenk, and V. Stoyanov. Xnli: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium, 2018. Association for Computational Linguistics.

A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzman, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov. Unsupervised cross-lingual representation learning at scale, 2020.

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

M. Freitag, R. Rei, N. Mathur, C.-k. Lo, C. Stewart, G. Foster, A. Lavie, and O. 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 (WMT)*, pages 733–774. Association for Computational Linguistics, November 10–11 2021.

M. Freitag, R. Rei, N. Mathur, C.-k. Lo, C. Stewart, E. Avramidis, T. Kocmi, G. Foster, A. Lavie, and A. F. T. 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, Abu Dhabi, United Arab Emirates (Hybrid), Dec. 2022. Association for Computational Linguistics.

R. Gupta, C. Orasan, and J. van Genabith. ReVal: A simple and effective machine translation evaluation metric based on recurrent neural networks. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1066–1072, Lisbon, Portugal, Sept. 2015a. Association for Computational Linguistics.

R. Gupta, C. Orasan, and J. van Genabith. Reval: A simple and effective machine translation evaluation metric based on recurrent neural networks. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1066–1072, Lisbon, Portugal, 2015b. Association for Computational Linguistics.

J. Hu, S. Ruder, A. Siddhant, G. Neubig, O. Firat, and M. Johnson. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, 2020.

Y. Ishibashi, K. Sudoh, K. Yoshino, and S. Nakamura. Reflection-based word attribute transfer. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 51–58, Online, 2020. Association for Computational Linguistics.

J. Ive, F. Blain, and L. Specia. deepquest: A framework for neural-based quality estimation. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3146–3157, Santa Fe, New Mexico, USA, 2018. Association for Computational Linguistics.

F. Kepler, J. Trenous, M. Treviso, M. Vera, A. Gois, M. A. Farajian, A. V. Lopes, and A. F. T. Martins. Unbabel's participation in the wmt19 translation quality estimation shared task. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 78–84, Florence, Italy, 2019a. Association for Computational Linguistics.

F. Kepler, J. Trenous, M. Treviso, M. Vera, and A. 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, 2019b. Association for Computational Linguistics.

P. Koehn and C. 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.

P. Lewis, B. Oguz, R. Rinott, S. Riedel, and H. Schwenk. Mlqa: Evaluating cross-lingual extractive question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020)*, pages 7315–7330, Online, 2020. Association for Computational Linguistics.

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

C.-k. 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, 2019. Association for Computational Linguistics.

V. Logacheva, C. Hokamp, and L. Specia. Marmot: A toolkit for translation quality estimation at the word level. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3671–3674, Portoroz, Slovenia, 2016. European Language Resources Association (ELRA).

V. Mariana, T. Cox, and A. Melby. The multidimensional quality metrics (mqm) framework: a new framework for translation quality assessment. *The Journal of Specialised Translation*, (23):137–161, 2015.

B. Marie, A. Fujita, and R. 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. Association for Computational Linguistics, 2021.

J. Nivre, M. Abrams, Z. Agic, L. Ahrenberg, L. Antonsen, M. J. Aranzabe, G. Arutie, M. Asahara, L. Ateyah, M. Attia, et al. Universal dependencies 2.2. *LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics, Charles University in Prague*, 2018.

J. Olive. Global autonomous language exploitation (gale). DARPA/IPTO Proposer Information Pamphlet, 2005.

X. Pan, B. Zhang, J. May, J. Nothman, K. Knight, and H. Ji. Cross-lingual name tagging and linking for 282 languages. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1946–1958, Vancouver, Canada, 2017. Association for Computational Linguistics.

K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, 2002. Association for Computational Linguistics.

K. Parton, J. Tetreault, N. Madnani, and M. Chodorow. E-Rating Machine Translation. In *Proceedings of the Sixth Workshop on Statistical Machine Translation (WMT-11)*, pages 108–115, Edinburgh, Scotland, 2011.

M. Popovic. CHRF: Character N-gram F-score for Automatic MT Evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisboa, Portugal, September 2015.

M. Popovic. CHRF Deconstructed: β 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.

M. Popovic and H. Ney. Word error rates: Decomposition over POS classes and applications for error analysis. In *Proceedings of ACL Workshop on Machine Translation*, Prague, Czech Republic, 2007.

R. Rei, C. Stewart, A. C. Farinha, and A. Lavie. Unbabel's participation in the WMT20 metrics shared task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 911–920, Online, Nov. 2020a. Association for Computational Linguistics.

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

R. Rei, A. C. Farinha, C. Zerva, D. van Stigt, C. Stewart, P. Ramos, T. Glushkova, A. F. T. Martins, and A. Lavie. Are references really needed? Unbabel-IST 2021 submission for the metrics shared task. In *Proceedings of the Sixth Conference on Machine Translation*, pages 1030–1040. Association for Computational Linguistics, 2021.

N. Reimers and I. 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. Association for Computational Linguistics, 2020.

T. Sellam, D. Das, and A. 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, 2020. Association for Computational Linguistics.

C. Servan, N. T. Le, N. Q. Luong, B. Lecouteux, and L. Besacier. An open-source toolkit for word-level confidence estimation in machine translation. In *Proceedings of the 12th International Workshop on Spoken Language Translation: Papers*, pages 196–203, Da Nang, Vietnam, 2015.

H. Shimanaka, T. Kajiwara, and M. 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, 2018. Association for Computational Linguistics.

M. Snover, B. Dorr, R. Schwartz, L. Micciulla, and J. 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, 2006. Association for Machine Translation in the Americas.

L. Specia, G. Paetzold, and C. Scarton. Multi-level translation quality prediction with quest++. In *Proceedings of ACL-IJCNLP 2015 System Demonstrations*, pages 115–120, Beijing, China, 2015. Association for Computational Linguistics and The Asian Federation of Natural Language Processing.

M. Stanojevic and K. Sima'an. BEER: BEtter Evaluation as Ranking. In *Proceedings of the Ninth Workshop on Statistical Machine Translation (WMT-14)*, pages 414–419, Baltimore, Maryland, June 2014a.

M. Stanojevic and K. Sima'an. Fitting Sentence Level Translation Evaluation with Many Dense Features. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP-14)*, pages 202–206, Doha, Qatar, Oct. 2014b. Association for Computational Linguistics.

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

B. Thompson and M. 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, 2020. Association for Computational Linguistics.

J. Tiedemann and S. Thottingal. OPUS-MT – Building open translation services for the World. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 479–480, Lisboa, Portugal, 2020. European Association for Machine Translation.

J. P. Turian, L. Shen, and I. D. Melamed. Evaluation of machine translation and its evaluation. In *Proceedings of Machine Translation Summit IX: Papers*, New Orleans, USA, 2003.

E. van den Heuvel and Z. Zhan. Myths about linear and monotonic associations: Pearson's r, spearman's ρ, and kendall's τ. *The American Statistician*, 76(1):44–52, 2022.

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

Y. Yang, Y. Zhang, C. Tar, and J. Baldridge. Paws-x: A cross-lingual adversarial dataset for paraphrase identification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3687–3692. Association for Computational Linguistics, 2019.

T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi. BERTSCORE: Evaluating text generation with BERT. In *International Conference on Learning Representations*, pages 1–43. Department of Computer Science and Cornell Tech, Cornell University, 2020.

P. Zweigenbaum, S. Sharoff, and R. Rapp. Overview of the third bucc shared task: Spotting parallel sentences in comparable corpora. In *Proceedings of 11th Workshop on Building and Using Comparable Corpora*, pages 39–42, 2018.