Master Thesis

Lost in Translation: Analyzing Machine Translation Quality Estimation with Synthetic Challenges

Selin Acikel

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, Amir Kamran
2^{nd} reader: Pia Sommerauer

Submitted: June 28, 2024

Abstract

This thesis investigated the effectiveness of Machine Translation Quality Estimation (MTQE) models, explicitly focusing on Dutch-English translation pairs, using synthetic datasets tailored to challenge these models with various error types. This study employs synthetic dataset generation techniques using large language models, GPT-3.5-turbo and GPT-4-turbo, that introduce controlled common machine translation errors such as inaccuracies in named entities, numbers, and negation.

The synthetic datasets were utilized to assess the performance of several multilingual MTQE models: CometKiwi, TransQuest, LASER, and LaBSE. Each model was evaluated based on its ability to detect and quantify introduced errors. Results indicate varied sensitivity to different error types across models, highlighting specific strengths and weaknesses in the context of synthetic distortions. For instance, some models showed higher precision in detecting named entity errors, while others were better at identifying number discrepancies.

Declaration of Authorship

I, Ayse Selin Acikel, declare that this thesis, titled *Lost in Translation: Analyzing Machine Translation Quality Estimation with Synthetic Challenges* and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a degree 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: 28 June, 2024

Signed: 

Acknowledgments

First and foremost, I extend my deepest gratitude to my thesis supervisor, Sophie Arnoult, from Vrije Universiteit Amsterdam, for her invaluable guidance and insightful feedback throughout the development of this thesis. Her expertise and support have been fundamental to my research progress and personal growth.

I am equally grateful to Amir, my internship supervisor at TAUS, who has been instrumental in shaping my practical skills and understanding of NLP applications. His mentorship has significantly enriched my learning experience. I would also like to acknowledge the efforts of the rest of the NLP team: David, Lahorka, and Lisa at TAUS, whose assistance and advice were crucial in managing and analyzing the data effectively. Their contributions were key to the success of my internship. Furthermore, I thank the entire team at TAUS for their kindness and hospitality, which made my time there both enjoyable and educational.

My heartfelt thanks go to my peers in the master's program. Their presence and collaborative spirit greatly enhanced our learning experience and made the challenging moments of academic pursuit much more manageable.

I also appreciate all the professors who challenged me to reach my full potential. Their deep commitment to academic excellence has significantly improved my learning experience, driving me to achieve more than I thought possible.

Lastly, I am thankful for the ongoing support of my friends and family, whose encouragement and belief in my abilities continue to inspire and motivate me. Their unwavering support has been my backbone throughout this journey.

List of Figures

List of Tables

Contents

Chapter 1

Introduction

Machine Translation (MT) is an essential component of computational linguistics aimed at converting text from one language to another using computer models. The historical development of MT spans from rule-based systems to the current state-of-the-art neural machine translation (NMT) systems (Sutskever et al., 2014). These advancements have significantly enhanced the fluency and contextual accuracy of translations, making MT integral to global communication, content localization, and information accessibility across language barriers (Koehn, 2020).

As the reliance on Machine Translation (MT) grows, so does the necessity for robust Machine Translation Quality Estimation (MTQE). MTQE is a subfield that focuses on predicting the quality of machine-translated text without reference translations, offering a crucial feedback mechanism for improving MT systems and their applications in real-world scenarios (Specia and Shah, 2018) (See table 1.1 for an example). Historically, the evaluation of translated texts depended on reference translations, which required significant human effort and expertise to produce (Hutchins and Somers, 1992). These reference translations, often created by bilingual experts, were used as a standard against machine-translated text comparisons. While providing a measurable standard for quality, this approach was costly, time-consuming, and limited in its ability to measure alongside the fast advancements in MT technologies.

Type	**Translation**	**QE Score**
Source	"Dit jaar zal asbest meer dan 3 000 mensen in het Verenigd Koninkrijk doden"	
Machine Translation	"This year, asbestos will kill more than 3000 people in the United Kingdom."	0.90
Distorted Machine Translation	"This year, asbestos will kill more than 3000 people in New Zealand."	0.75

Table 1.1: Example of MTQE from Dutch Source Sentences to English Target Sentences. (The QE scores are determined by CometKiwi)

Traditional evaluation methods like BLEU scores (Papineni et al., 2002), while useful, often fall short in capturing the nuanced grammatical and contextual appropriateness of translated texts (Callison-Burch et al., 2006). The limitations of these methods become particularly evident as they struggle to reflect the true semantic and syntactic quality of translations in the absence of identical lexical choices between the reference and the translated text (Doddington, 2002). Consequently, there was a need for more

advanced and nuanced quality assessment methods that could operate independently of human-made reference texts.

MTQE addresses these limitations by employing models that evaluate translation quality based on linguistic features and error patterns rather than direct comparison to a reference text (Specia and Shah, 2018). This shift reduces the dependency on exhaustive reference translations. It also aligns more closely with the dynamic and varied use of language in real-world scenarios. Therefore, it enhances machine translation's practical utility in global communication.

Recent developments in AI and machine learning, particularly the introduction of transformer models (Vaswani et al., 2017) and large language models (LLMs) like GPT-3.5 and GPT-4, have opened new doors for enhancing MTQE (Brown et al., 2020). These models have demonstrated exceptional capabilities in generating human-like text and are used in this project to create synthetic datasets with specific error patterns. Such datasets can serve as challenge sets to evaluate the efficiency of various MTQE models like CometKiwi (Rei et al., 2022), TransQuest (Ranasinghe et al., 2020), LASER (Artetxe and Schwenk, 2019), and LaBSE (Feng et al., 2020) against controlled error categories.

The challenge lies not only in accurately assessing translation quality but also in understanding the specific limitations and strengths of different MTQE models when confronted with systematically introduced errors. This understanding could lead to significant improvements in MTQE systems, ensuring they are more reliable and effective across diverse linguistic contexts.

1.1 Related Works

The rapid growth of digital communications and reliance on MT highlights the urgent need for linguistically accurate translations (Koehn, 2010). With the rise of demand for real-time translation across different industries, the key challenge is to guarantee that these translations are devoid of accuracy errors that could compromise the integrity of information. This research is motivated by the need to enhance MT techniques to detect and address accuracy-related errors, thereby ensuring that MT systems consistently produce precise translations.

Quality Estimation (QE) is key in tackling these issues and offers a mechanism to evaluate translations dynamically without the need for reference texts. This method is beneficial when quick decision-making is critical and traditional translation evaluation methods need to be faster or more manageable Specia and Shah (2018). However, despite significant advancements in natural language processing (NLP) technologies that have enhanced the capabilities of QE systems, the models still struggle with accurately identifying and quantifying errors (Sharou and Specia, 2022).

The Workshop on Machine Translation (WMT) has contributed significantly in the advencement of the MTQE field. WMT is an annual academic event that emphasizes the evaluation of machine translation systems through comparative testing and benchmarking. WMT was established due to the growing need within the computational linguistics community to systematically compare the performance of machine translations across different language pairs and translation approaches. Ever since, they have contributed significantly to the field by providing datasets and benchmarks.

1.1.1 Evolution of MTQE in WMT

During the period from 2016 to 2018, there was a significant change at WMT as the organization started to adopt and integrate neural network-based methods for MTQE (Bojar et al., 2016; Specia et al., 2018). This shift was primarily influenced by the adoption of deep learning technologies, which introduced new capabilities and methodologies to the field of machine translation and its evaluation.

Deep Learning Integration

The integration of deep learning in MTQE tasks highlighted by WMT during these years was a key movement from traditional feature-based quality estimation models. Traditional models heavily relied on manually crafted features, such as lexical, syntactic, and semantic information extracted from both the source text and its translation feature sets (Koehn, 2010). However, these models often struggled with capturing deeper contextual meanings and were limited by the fixed nature of their feature sets.

In contrast, neural approaches, particularly those employing deep neural networks, have the ability to learn these features implicitly from large amounts of data. This capability allows them to better understand and interpret the complexities and nuances of language, which are crucial for assessing translation quality accurately. Models such as CNNs (Kalchbrenner and Blunsom, 2013) and RNNs (Cho et al., 2014) , and later transformers (Vaswani et al., 2017) began to be explored and adopted for their superior performance in capturing sequential data and their ability to maintain context over longer texts (Specia et al., 2018)).

Shared Tasks and Benchmarks

A significant contribution of WMT during this period was the establishment and advancement of shared tasks specifically focused on MTQE. These tasks provided researchers with a platform to test and benchmark their models using standardized datasets annotated with quality scores. For instance, the introduction of sentence-level and word-level quality estimation tasks allowed a more nuanced analysis of translation output, allowing researchers to pinpoint specific areas of strength and weakness in translation models (Fonseca et al., 2019).

These benchmarks were essential not only for advancing the state of the art but also for understanding how different models performed under similar conditions. The datasets used in these tasks often included a variety of language pairs and translation domains, providing a comprehensive testing ground for new MTQE methodologies.

1.1.2 Innovations and Outcomes

The advancements in MTQE during this period led to several key innovations. For example, the use of attention mechanisms within neural models provided a way to focus on specific parts of the input when predicting quality, which was particularly useful for identifying mistranslations or subtle errors that could impact the overall translation quality (Tiedemann and Scherrer, 2018). Additionally, the use of transfer learning and multi-task learning methods has started to become more common. This involves adapting models trained on similar tasks, like machine translation or text summarization, to improve the quality estimation process and make their performance and applicability more effective.

1.1.3 Recent Findings

The WMT 2022 Quality Estimation shared task revealed some critical challenges that MTQE models continue to face (Fomicheva et al., 2022). These include the accurate handling of named entities and the prediction of semantic errors. These issues are particularly apparent in low-resource languages, where less training data is available. Additionally, the task highlighted a critical need to improve the precision of MTQE models in assessing translation quality, not only at the overall text level but also at the word and sentence levels. This advancement is crucial for identifying subtle linguistic differences that can significantly impact translation quality, highlighting the ongoing need for advanced model training and evaluation methodologies.

1.2 Motivation

This research is mainly motivated to generate synthetic datasets that simulate a variety of accuracy errors. This approach allows for a controlled yet comprehensive evaluation of QE models and provides deep insights into their effectiveness across different accuracy errors.

In exploring the effectiveness of QE, this study will utilize a Dutch-English language pair as the MT output. High-resource languages often benefit from extensive data and research, providing robust translation models that are well-understood (Zoph et al., 2016). In contrast, Dutch English, being relatively less explored, might reveal new challenges and insights. This deviation could lead to broader applications and a deeper understanding of QE methodologies, especially in how they handle low-resource or less-common language scenarios (Guzmán et al., 2019).

While WMT has significantly contributed to advancements in MTQE, it has primarily focused on either high-resource language pairs such as English-Chinese or very low-resource language pairs such as English-Gujarati. My study proposes a novel experiment using challenge sets that include systematic errors to test MTQE models specifically for the Dutch-English pair, which is not currently included in the WMT datasets. This approach not only addresses a gap in the research but also enhances our understanding of MTQE across different linguistic contexts. By studying these developments and applying similar neural network approaches to the Dutch-English language pair, my thesis aims to explore whether the insights gained from previous language pairs can be effectively translated to less commonly studied language pairs.

1.3 Aims and Research Question

This research aims to deepen the understanding of how MTQE performs and its strengths and weaknesses. This project is particularly focused on the capability of MTQE models to handle translations that have been intentionally distorted to simulate common translation errors.

The overall research question guiding this study is:

> *How effectively can current Machine Translation Quality Estimation models identify and quantify different types of translation errors introduced by advanced large language models in a Dutch-English dataset?*

This central question breaks down into several sub-questions to address the various facets of the research:

1. Can we utilize generative models, such as GPT-3.5 and GPT-4, to create challenge test sets by deliberately altering sentences to include specific error patterns?
2. Which QE model—CometKiwi, TransQuest, LASER, or LaBSE—performs best when confronted with these synthetically altered test sets?
3. Are there specific error patterns that consistently challenge the QE models, potentially highlighting areas for future improvement?

In the context of the second subquestion of my research, the term 'best' refers to the effectiveness and efficiency of an MTQE model in accurately identifying and quantifying the translation errors embedded within the test sets. This assessment is multifaceted and includes several performance dimensions, such as accuracy, consistency, granularity, and robustness. This evaluation helps identify the most reliable and helpful model in real-world applications where diverse and complex translation errors occur frequently.

1.4 Thesis Outline

The following chapter 2 provides the background for MTQE and introduces important topics related to MTQE. Chapter **??** is dedicated to related work in the field of MTQE and the similarities and differences between this project and earlier research. Chapter 3 describes the methods employed in this research, detailing the process of creating synthetic datasets and outlining the QE models applied during the experiments. Chapter 4 presents the core of the experiments. This includes a description of the dataset, the results obtained from the experiments, and an extensive error analysis. Lastly, chapter 5 discusses the implications of the findings, shares limitations to the study, reflects on the research question, and mentions future possibilities in the MTQE field.

Chapter 2

Theoretical Framework

In this chapter, several essential topics will be explored to deepen the understanding of my thesis. The discussion begins with the significant advancements in NMT, highlighting its development and impact. This is followed by an examination of the common critical errors encountered in MT. Various evaluation methods of MT are then introduced, emphasizing how these approaches assess translation accuracy and fluency. Lastly, the concept of challenge sets is explained, highlighting their utility in enhancing MTQE processes.

2.1 Neural Machine Translation

Although NMT models were researched between the '80s and '90s (Rumelhart et al., 1986; LeCun et al., 1989; Lecun et al., 1998), computational complexity and data scarcity have made it impossible to implement neural methods for MT during that period (Koehn, 2020). However, around 2013 and 2014, a return in interest led researchers to experiment with end-to-end NMT models. These models included convolutional neural networks (CNN) (Kalchbrenner and Blunsom, 2013) and recurrent neural networks (RNN) (Cho et al., 2014; Sutskever et al., 2014).

In more detail, these models implemented what is known as encoder-decoder architectures. The encoder processes the input sentence from a source language, transforming it into a dense vector representation. The decoder then uses this representation step by step to generate the output sentence in the target language. However, traditional RNNs and CNNs faced challenges with longer sentences because they relied on encoding the entire input sequence into a single fixed-length vector, which could lead to information loss over long distances (Bahdanau et al., 2014).

Bahdanau et al. (2014) addressed this limitation by proposing an encoder-decoder model enhanced by an attention mechanism. This mechanism allows the model to focus on different parts of the input sequence while translating, effectively aligning segments of the input text with their corresponding parts in the output text. This 'joint' learning aligns and translates input and output, which ensures that each word in the translation closely corresponds to the appropriate words in the input, enhancing both accuracy and fluency in MT.

The current State-of-the-art models for MT are transformers, which also utilize an attention mechanism. This model, introduced by Vaswani et al. (2017), represents a different approach, where it favors self-attention, positional encoding, and feed-forward layers instead of recurrent layers that are used in the traditional NMT models. Since

their development, transformers have become a foundation in many areas within NLP due to their effectiveness and efficiency with long sequences.

The advancements in NMT have substantially improved the quality of machine translation by enabling more accurate and contextually appropriate translations compared to earlier methods (Vaswani et al., 2017). This leap in quality was due to the models' ability to better grasp the complexities of language, including idiomatic expressions and nuanced grammatical structures.

2.2 Critical Translation Errors

Despite these advancements, the quality of automatic translation still falls short in many instances (Specia and Shah, 2018). State-of-the-art MT still lacks quality in many aspects. The translations still include errors that can differ in severity from minor to critical. These errors often carry risks. According to the taxonomy developed by Sharou and Specia (2022), a critical error in machine translation occurs when the meaning of the translated text is significantly different from the original text. This can potentially result in misunderstandings and harmful consequences related to health, safety, legal, reputation, religion, or finance for stakeholders. Therefore, mitigating these risks by identifying and evaluating these errors is important.

Identifying and categorizing translation errors in machine translation presents a significant challenge due to the subjective nature of language and its contextual nuances. Achieving consensus on what constitutes an error and its severity can be particularly difficult among linguists and translation experts, especially when texts include variability in languages, dialects, and cultural contexts because these can influence the perception of what is considered an error. Lommel et al. (2014) highlights these challenges and stresses the importance of developing robust frameworks for systematic error analysis and categorization.

Therefore, an essential resource is utilized to categorize the errors to align with an important metric, called the Multidimensional Quality Metrics (MQM) (Lommel et al., 2014). MQM was created as a response to the need for a more comprehensive and nuanced approach to evaluating translation quality beyond traditional metrics. It encompasses a broader range of quality dimensions such as fluency, accuracy, style, terminology, and more. MQM provides a structured framework for assessing and improving translation.

This thesis project mainly focuses on critical errors in accuracy as these contain the direst risks for misunderstandings, and identifying these errors is of utmost importance in MT. Sharou and Specia (2022) were one of the first to focus solely on critical errors and created a taxonomy listing common critical errors found in MT. This taxonomy will be used as a baseline for creating a challenge test set for the QE models applied during my project. The following subsections will present some of these common error categories that will be included in my challenge set with references to the MQM framework.

2.2.1 Accuracy

Accuracy mistranslations refer to the phenomenon where the target content does not accurately represent the source content. This can include errors in the misinterpretation of the meaning of a word, not translating a word/phrase from the source target, or

simply adding gibberish to the target sentences (Sharou and Specia, 2022).

Zeman (2008) showed that accuracy errors, specifically mistranslations, accounted for approximately 30-40% of all errors identified in MT outputs across various language pairs. This high percentage underscores the critical nature of accuracy in MT and the need for a refined QE model that is able to detect and mitigate these errors.

Below, I will introduce errors within this category that I have employed in my dataset. Including these error patterns in the dataset will help us understand whether QE models can detect these common MT errors.

Numbers

In this subcategory, the MT mistranslates a number, date, or time. For example, the source target refers to article 5(2) but gets mistranslated to 7(3). This error can cause misunderstandings that could lead to an unpleasant or major consequence such as missing an important appointment, making financial errors, and more. Depending on the end-user and stakeholders, this error can vary in criticality.Specia et al. (2019), reported on a shared task including several teams using different MTQE models to identify errors, and their results indicated that numerical errors occurred in about 15% of sentences.

Named Entities

Errors within this subcategory introduce distortions of the named entities (people, locations, and organizations) in the target sentence. For example, the source sentence referring to Amsterdam gets mistranslated as Lisbon in the target sentence. Critical errors introduce complete changes in the target sentence, making the named entity unrecognizable. This error can have severe consequences for stakeholders when, for instance, writing contracts. The same shared task reported by Specia et al. (2019) also measured around 20% of named entity errors in the translated outputs, indicating a need for a great sensitivity to identify and mark NER (named entitiy recorgnition) errors by QE models.

Hallucinations

Hallucinations have been a widely researched phenomenon within MT (Ji et al., 2022; Guerreiro et al., 2023). This subcategory refers to the random mistranslation of a word within the source sentence into a completely different word that has no particular relation with the source. See example:

- Target:"Conditions which may be attached to rights of use for **numbers**"
- Hallucination: "Conditions which may be attached to rights of use for **bananas**"

Usually, hallucinations have no similarity to the original meaning and are often out of context in relation to the sentence. Guerreiro et al. (2023) relate the occurrence of natural hallucinations to the lack of robustness in MT models, translation quality, and inherent biases or flaws in the training data. For instance, translating out of English tends to result in more hallucinations due to lower source contributions and potential toxic patterns in low-resource language pairs.

Müller et al. (2020) explicitly focused on the quantity of hallucinations appearances in MT and found out that approximately 5-10% of sentences in low-resource language pairs exhibited some form of hallucination, significantly impacting translation reliability.

Negation/Sentiment

Errors involving negation or sentiment changes can critically alter the meaning of a sentence. For instance, changing "The product is not safe for children" to "The product is safe for children" introduces a dangerous misinterpretation. Such errors can have significant implications, particularly in contexts involving safety, legal matters, or medical information.

(Sennrich, 2017) analyzed MT outputs for negation handling and found that around 25% of sentences involving negation had errors, either by omitting or incorrectly translating the negation, leading to potentially dangerous misinterpretations.

Omission

Omission errors occur when essential information from the source text is missing in the target text. This can lead to incomplete translations that fail to convey the full meaning or critical details intended by the original content. For example:

- Target: "The software update includes security patches, **performance improvements**, and new features."
- Omission:" The software update includes security patches and new features."

In the MQM framework, omission is considered a critical error because it impacts the completeness and accuracy of the translation, potentially leading to misunderstandings or misinformation.

Lommel and Burchardt (2014) experimented with a detailed error analysis across multiple MT systems and found that omission errors were particularly existent in complex sentences, occurring in approximately 12-18% of cases, depending on the language pair and MT model used.

Addition

Addition errors happen when extra information that was not present in the source text is included in the target text. This can lead to misleading or confusing translations. One specific type of addition error involves the introduction of non-existing words or gibberish into the target sentence. These errors can significantly impact the readability and comprehensibility of the translation, making it difficult for the reader to understand the intended message. According to the MQM framework, addition is a significant error as it introduces content that can distort the original message and affect the reliability of the translation. For example:

- Target: "Please read the user manual before operating the device."
- Addition: "Please read the user manual before operating the **flibber** device."

In this example, the addition of the non-existing word "flibber" makes the sentence confusing and potentially difficult to understand. Such errors can undermine the clarity and professionalism of the translated text, leading to user frustration or misinterpretation of important information. This highlights the critical nature of addition errors, particularly when they involve non-existent or nonsensical terms.

Studies such as those by Bentivogli et al. (2016) and Lommel and Burchardt (2014)often include addition errors as part of a broader error analysis. These studies show that addition errors, while less common than omission errors, still represent a significant challenge, particularly in neural machine translation systems that tend to generate fluent but occasionally overly verbose outputs.

2.3 Evaluation of Machine Translation

Evaluating the quality of MT is crucial for understanding and improving MT systems. This section explores traditional evaluation methods and their differences.

2.3.1 Traditional Evaluations

Traditional methods for evaluating MT primarily involve comparing the translated text to reference translations created by human experts. These methods have evolved over time and include both manual and automatic evaluation techniques.

Manual Post-Editing

This method involves human translators reviewing and correcting machine-generated translations. The changes made during this process provide insights into the types and frequencies of errors made by the MT system. Although manual post-editing offers high accuracy, it is time-consuming and expensive, making it impractical for large-scale evaluations.

BLEU Score

The Bilingual Evaluation Understudy (BLEU) score, introduced by Papineni et al. (2002), remains one of the most influential and widely used automatic evaluation metrics in machine translation. BLEU assesses the accuracy of machine-generated translations by comparing the overlap of n-grams—consecutive sequences of words—between the translated text and one or more human-generated reference translations. The metric calculates precision scores for n-grams of different lengths (usually up to 4-grams) and combines them using a geometric mean, then applies a brevity penalty to discourage overly short translations.

While BLEU is praised for its computational efficiency and its relatively good correlation with human judgment at the corpus level, it exhibits several notable limitations. Firstly, BLEU does not inherently assess the grammatical structure or correctness of the translation; instead, it focuses primarily on the lexical matching of n-grams. This can lead to high scores for translations that are lexically similar to the reference but grammatically incorrect (Callison-Burch et al., 2006). Additionally, BLEU often fails to evaluate the contextual appropriateness of translations, as it does not account for the conveyed meaning, which can be particularly problematic in translations involving idiomatic expressions or culturally specific content (Callison-Burch et al., 2006).

The effectiveness of BLEU also varies significantly across languages. It is notably less reliable for languages with rich morphology or flexible word order, such as Turkish, Hungarian, or Finnish (Turhan and Oflazer, 2008). In these languages, literal n-gram matches are less indicative of actual translation quality because such languages often involve inflectional changes and can reorder words without altering the meaning.

Despite these limitations, BLEU continues to be a benchmark in the field due to its ease of use and the ability to quickly compare different translation systems or track system improvements over time (Papineni et al., 2002). However, researchers and developers are encouraged to use BLEU in conjunction with other metrics that can compensate for its shortcomings, such as METEOR (Lavie and Denkowski, 2009), which considers synonymy and grammatical alignment, or newer neural-based evaluation metrics that can better capture semantic and syntactic translation qualities.

METEOR and TER

Building upon the work of BLEU, other automatic metrics like METEOR (Metric for Evaluation of Translation with Explicit ORdering) Lavie and Denkowski (2009) and Translation Edit Rate (TER) Snover et al. (2006) have been developed to address some of the limitations identified in BLEU. Unlike BLEU, which primarily relies on exact matches of n-grams, METEOR enhances evaluation by considering synonyms and stemming, allowing for a more flexible matching of words. Additionally, METEOR incorporates structural matching to reward translations that align well with the syntax of the reference translations, using alignments based on exact, stem, synonym, and paraphrase matches to produce a score (Denkowski and Lavie, 2012).

TER focuses on the edit distance, which is the minimum number of edits required to change a translation into one of the reference translations. This includes insertions, deletions, substitutions, and shifts of words in the translated text. TER is often used as a complementary metric to BLEU in professional translation workflows because it directly quantifies the effort required to post-edit machine-translated output into an acceptable final product. This makes TER particularly valuable in scenarios where post-editing efficiency is a critical performance indicator (Snover et al., 2006).

Despite the advancements these metrics represent, they still share a common limitation with BLEU: reliance on reference translations. While METEOR's use of synonyms and morphological variations allows it to capture meaning to a greater extent, and TER's edit-based approach offers a direct measure of translation edit effort, both metrics fundamentally depend on the quality and availability of reference translations (Denkowski and Lavie, 2012). This dependence can introduce biases, especially in cases where the reference translations do not fully reflect the target language's idiomatic usage or cultural nuances.

Furthermore, both METEOR and TER may still struggle to fully capture the nuances of language that go beyond the lexical or syntactic similarities, such as pragmatic appropriateness and stylistic conformity (Denkowski and Lavie, 2012). These aspects are often crucial in translations of literary texts or in localized marketing materials where the emotional or cultural resonance of the language is important. To address these nuanced requirements, newer evaluation frameworks and models that integrate advanced linguistic and semantic analyses are increasingly considered.

2.3.2 Quality Estimation

QE is an emerging field in machine translation that shifts away from traditional dependency on reference translations for evaluating translation quality. Unlike metrics such as BLEU, METEOR, and TER, which require high-quality human-translated texts for comparison, QE techniques aim to predict the quality of translated texts directly, without any reference Specia and Shah (2018). This approach is particularly valuable when reference translations are unavailable, such as in real-time translation scenarios or for languages with limited resources.

QE utilizes machine learning models that are trained on a dataset of source texts, their translations, and quality annotations (not necessarily reference translations) Specia and Shah (2018). These models learn to predict quality scores on word-, sentence-, and document-level based on features extracted from the source and translated texts, including lexical, syntactic, and semantic features. Advanced QE models also incorporate features from pre-trained neural networks that understand deeper linguistic contexts.

While QE is a promising approach, it faces several challenges. The accuracy of QE is highly dependent on the quality and representativeness of the training data. In cases where training data is biased or insufficient, QE models may not perform well. Additionally, developing robust QE models that can handle the variability and complexity of human languages across different contexts remains a significant challenge (Specia et al., 2018).

2.4 MTQE Models

The following section will provide a description of the four MTQE models utilized for the evaluation of the two challenge sets. The descriptions include a general overview, their architecture, training and a section defining comparisons and differences between the models.

2.4.1 CometKiwi

CometKiwi (Rei et al., 2022) combines the predictive capabilities of the COMET framework (Rei et al., 2020) with the architectural advancements of OpenKiwi (Kepler et al., 2019) to improve performance in MTQE. The model is designed to perform well in multilingual settings by utilizing comprehensive pretraining on diverse linguistic data to be able to generalize on unseen language pairs. Its main goal is to deliver efficient and reliable QE for translation tasks without the need for reference translations. This makes the model a great candidate for this project as I have a language pair that is often non-existent in training data and am applying a reference-free approach even though the original translations could be used as a 'reference.' I opted for a reference-free approach to identify whether the CometKiwi and the other models in this chapter are able to score the distorted translations independent of the original ones.

Architecture

CometKiwi employs a transformer-based encoder architecture that utilizes the pretrained XLM-Roberta model (See figure 2.1). This enables the encoding of source and target text into high-dimensional vector spaces. The model architecture is improved

by a scalar mix mechanism. This mechanism combines outputs from different levels of the transformer network. Since different layers of a transformer capture different types of information (some might understand the basic meaning of the words, while others might grasp the more complex relationships between parts of the sentence), this mixing allows CometKiwi to focus on the most relevant features for QE. By weighing the contribution of each layer, the model can better assess which parts of the translation need attention, leading to more accurate evaluations of translation quality.

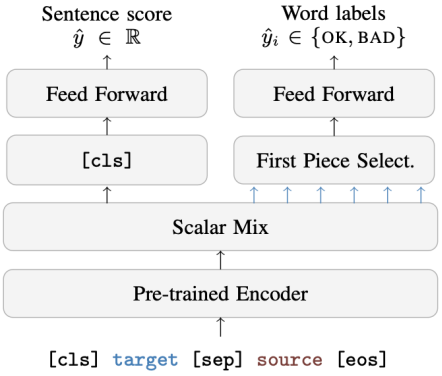

Figure 2.1: General Architecture of CometKiwi for Sentence Level (Left), Word Level (Right) (taken from (Rei et al., 2022))

Sentence-Level Quality Estimation

To estimate the quality of entire sentences, CometKiwi uses the embedding of the first token, usually the [CLS] token, from a combined output of different transformer layers. This single representation captures the essence of the entire sentence and is then processed through a feed-forward network. The output from this network provides a score that predicts the overall quality of the translated sentence.

2.4.2 TransQuest

TransQuest (Ranasinghe et al., 2020) is a framework designed for sentence-level MTQE, utilizing cross-lingual transformers to enhance its capabilities. By using a simpler architecture, it overcomes the usual computational and scalability issues seen in older neural-based quality estimation systems. This adaptability makes it particularly suitable for environments with limited resources and for languages that lack extensive annotated datasets.

Model Architecture

TransQuest has two main architectures: MonoTransQuest and SiameseTransQuest. For this project, I will focus on MonoTransQuest due to its simplicity and effectiveness in handling various languages. MTransQuest uses a single transformer model, the XLM-Roberta model, to encode both the source and target sentences separated by a [SEP] token (see Figure 2.2 for architecture). Its methodology revolves around various pooling strategies, which include:

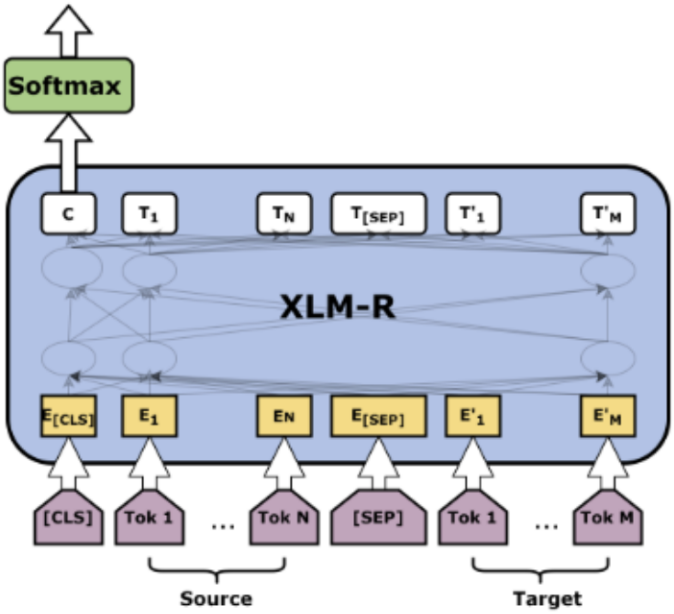

Figure 2.2: MTransQuest Architecture, taken from (Ranasinghe et al., 2020)

- CLS Strategy: This method uses the output from the [CLS] token that summarizes the entire input sequence.

- Mean Strategy: This strategy involves computing the mean (average) of all the output vectors generated by the transformer model for each token in the input sequence. This potentially smooths over outliers and emphasizes commonalities.

- Max Strategy: this strategy captures the most dominant or salient features in the input sequence. By focusing on the maximum values, it highlights the features that are most strongly expressed in the text, which are often crucial for understanding nuances such as specific errors or particularly well-translated segments.

The selected strategy's output is then fed into a softmax layer that predicts the quality score of the translation.

2.4.3 LASER

LASER (Language-Agnostic Sentence Representations) (Artetxe and Schwenk, 2019) is a framework designed to generate language-agnostic sentence embeddings. It can process text in over 140 languages using a single multilingual model. This approach allows LASER to efficiently handle cross-lingual tasks, such as textual similarity and retrieval, by transforming sentences into a high-dimensional space where similar sentences are positioned close to each other based on their meaning, regardless of the language.

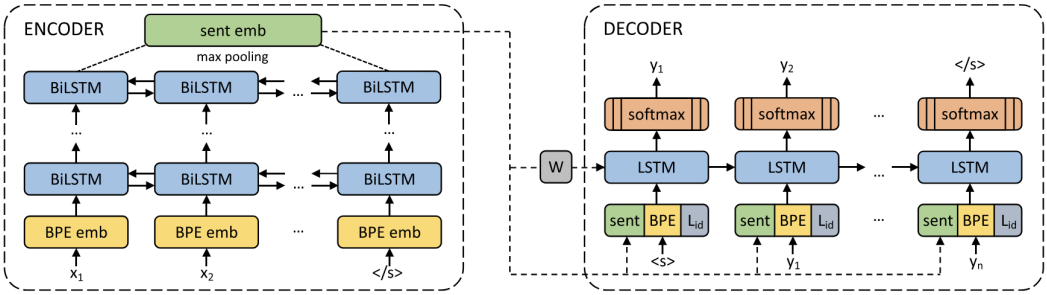

Figure 2.3: LASER Architecture, Taken From (Artetxe and Schwenk, 2019)

Model Architecture and Training

The training of LASER employs a sequence-to-sequence model using a multilingual corpus to create language-agnostic sentence embeddings. It involves a bidirectional LSTM encoder that generates dense vector representations of input sentences in various languages and a unidirectional LSTM decoder trained to reconstruct sentences in the target language, often using back-translation to enhance semantic accuracy (see Figure 2.3). As each word is processed and predicted, the softmax layer determines the most likely next word based on the encoder's context and the decoder's preceding words. This process aims to minimize the difference between original and reconstructed sentences, combining cross-entropy and cosine similarity losses to refine the embeddings.

Once trained, these embeddings are useful for cross-lingual tasks such as textual similarity and information retrieval, allowing comparisons across languages without direct translation. In the context of MTQE, LASER's embeddings enable the assessment of how well the semantic content of the original text is preserved in the translation, offering a robust tool for evaluating translation quality.

2.4.4 LabSE

LaBSE (Language-agnostic BERT Sentence Embedding) (Feng et al., 2020) is a similar model developed to generate multilingual and language-agnostic sentence embeddings. It integrates advanced methods from the field of machine learning and linguistics to address the challenges of cross-lingual semantic retrieval, making it effective for applications like translation ranking and sentence similarity assessments across different languages.

Architecture

The LaBSE model utilizes a dual encoder framework, each based on a 12-layer transformer architecture, similar to BERT (see figure 2.4). These encoders process the source and target text separately, generating embeddings for each. The model is notable for its utilization of a large, pretrained multilingual language model, which significantly enhances its performance by leveraging the learned representations from massive amounts of text data. After sentence embedding generation, the model uses an additive margin softmax to refine the separation between similar and non-matching sentence pairs. Intuitively, the goal is to ensure that the model assigns a higher score to the correct translation pair than to any other incorrect pairs within the same batch. By using a

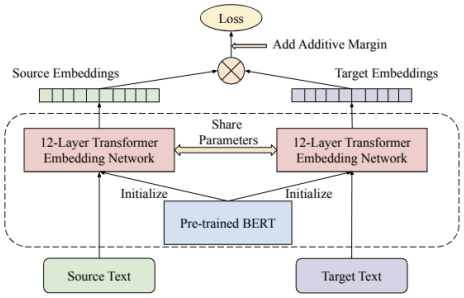

Figure 1: Dual encoder model with BERT based encoding modules.

Figure 2.4: LaBSE Architecture, Taken From (Feng et al., 2020)

softmax function, the model not only has to favor the correct answer but also learn to distinguish it significantly from a set of plausible but incorrect answers.

2.5 Challenge Sets

Challenge sets are essential tools for evaluating and benchmarking the performance of MT systems and QE models (Lehmann, 1996; Sennrich et al., 2017; Belinkov and Glass, 2019),. These sets are designed to test specific aspects of translation quality and to reveal weaknesses in QE systems by presenting translations with varying levels and types of difficulty. In this section, I will discuss the concept of challenge sets, their importance, how they are used in the context of QE evaluation, and their construction.

2.5.1 Concept of Challenge Sets

Challenge sets are carefully created datasets containing examples designed to push the boundaries of MT systems (Belinkov and Glass, 2019). Unlike regular test sets, which typically consist of a broad range of general examples, challenge sets focus on specific phenomena or error types that are known to be difficult for MT systems to handle. These can include idiomatic expressions, complex syntactic structures, rare vocabulary, and various linguistic errors.

The importance of challenge sets lies in their ability to provide a more nuanced and detailed understanding of an MT system's capability and limitations (Kocmi and Federmann, 2023). By focusing on specific challenges, researchers and developers can gain insights into how well their systems handle difficult cases and identify areas that require further improvement.

Challenge sets are particularly valuable in the context of QE, where the goal is to predict the quality of translations without reference translations. By using challenge sets, researchers can evaluate how well QE models identify and handle difficult cases and whether they can accurately predict the severity of different types of errors.

Kocmi and Federmann (2023) demonstrate the effectiveness of using LLMs for QE, showing that these models can achieve state-of-the-art performance even on challenging examples. They highlight the importance of well-designed prompts and diverse examples in training and evaluating QE models.

2.5.2 Construction of Challenge Sets

According to (Belinkov and Glass, 2019), the generation of challenge sets is a meticulous process that identifies specific linguistic phenomena that challenge machine translation systems, such as syntactic ambiguities or idiomatic expressions. Examples illustrating these phenomena are then either modified from existing resources or created by language experts to test the systems under controlled conditions. Each example is carefully annotated with expected translations and undergoes validation to ensure it accurately represents the intended linguistic challenge. Lastly, to ensure the challenge sets are broadly applicable, they are diverse and representative of various language pairs and contexts.

Chapter 3

Methodology

This chapter provides insights into the approaches applied during this project. Firstly, this section will describe the data used during the project. Secondly, I will discuss the creation of a synthetic dataset by prompting two GPT models. Thirdly, this chapter will summarize and compare the four QE models that have been applied during the experiment: CometKiwi (Rei et al., 2022), TransQuest (Ranasinghe et al., 2020), LASER (Artetxe and Schwenk, 2019), and LaBSE (Feng et al., 2020). Lastly, this section will introduce the evaluation of the project, including the Kullback-Leibler Divergence (Kullback and Leibler, 1951), a measure to evaluate the distributions of scores from the models.

3.1 Dataset Description

3.1.1 Introduction to the Dataset

The dataset utilized in this thesis was sourced during an internship at TAUS and comprises approximately 20,000 Dutch-to-English translation pairs from diverse domains such as computer software, financial services, professional and business services, and an undefined sector that includes miscellaneous domains that have not been categorized yet. These translation pairs provide varied linguistic ranges to provide a wide range for examining MTQE methodologies. Table 3.1 shows the domain distribution within the dataset.

Error category	Computer Software	Professional Business Services	Undefined Sector	Financials
Addition (Gibberish)	512	579	2849	130
Deletion	480	498	2423	117
Entity	681	733	6410	298
MT Hallucination	514	570	2836	126
Negation	510	577	2798	128
Number	557	961	5271	309
Sentiment	492	529	2702	122

Table 3.1: Error Category Distribution Across the Domains in the Base Dataset

The primary focus of this dataset is to act as a basis test dataset that can be altered for the exploration of quality estimation methodologies in machine translation systems. The original translations will be viewed as a baseline to be compared with the distorted sentences in the QE models.

3.1.2 Inherent Errors in the Dataset

After an initial inspection of the dataset, I detected inherent errors in the original dataset that included natural MT errors such as:

1. Target sentences not containing enough information given in the source sentence:
   - source: "Vraag nr. 36 van de heer Robles Piquer (H-169/87) Betreft: Verslag over Europese coördinatie bij het onderzoek van de oceanen"
   - target: "Question No 36, by Mr Robles Piquer (H-169/87)"
2. Target sentences containing more information than given in the source sentence:
   - source: "Brief van de Commissie aan de lidstaten van 4 maart 1991"
   - target: "Commission letter to Member States **SG(91) D/4577** of 4 March 1991"
3. Source sentences containing special characters that make the source sentence unclear:
   - source: "De stijging van de Belgische bevolking zou dan nog enkel berusten op de netto **\u2011instroom** van migranten."
   - target: "Further increases in the population of Belgium beyond that date are expected to consist only of net immigration."

However, these errors were not filtered or cleaned, as the translations appeared adequate for QE tasks. The existence of these intrinsic errors is not inherently detrimental to the QE process. In principle, having translations that are not entirely error-free reflects more realistic conditions under which QE models operate. Moreover, original errors in the target sentences are also present in the modified target sentences, where an additional artificial error is introduced. Therefore, the identification of the introduced artificial error should not be a problem for QE models.

3.1.3 Development of the Curated Subset

However, for the purpose of this research, where the focus is to determine how well QE models can identify the specific errors I prompted within the target sentences, it can provide insights to isolate these errors from pre-existing ones. Therefore, an additional controlled subset of the dataset was developed. This subset was carefully created to exclude original errors in the target sentences, ensuring it only contained high-quality translations. This curated dataset aims to clarify whether lower quality scores are indeed due to artificially introduced errors rather than pre-existing ones.

This carefully created subset consists of 660 sentences, providing a cleaner and more reliable basis for testing the QE models under controlled conditions (see 3.2, 3.3). This dataset does not contain the "Financials" domain and, in contrast to the original dataset, has most sentences in the "Professional and Business Services" domain. The average words per sentence per error category do not differ significantly from the original dataset.

Error category	Sentences	Words	Avg. Words/Sentence
Addition (Gibberish)	101	1,459	14.45
Deletion	77	1,153	14.97
Entity	97	1,514	15.61
MT Hallucination	108	1,450	13.43
Negation	92	1,333	14,49
Number	93	1,376	14.80
Sentiment	92	1,446	15.72

Table 3.2: Error Category Division in the Curated Dataset

Error category	Computer Software	Professional Business Services	Undefined Sector
Addition (Gibberish)	5	84	12
Deletion	4	65	8
Entity	0	7	90
MT Hallucination	7	85	16
Negation	7	77	8
Number	8	43	42
Sentiment	4	77	11

Table 3.3: Error Category Division of the Curated Dataset for each domain

3.1.4 Curated Dataset Selection

The selection and curation of the dataset were meticulous processes aimed at creating a highly controlled environment for testing QE models. The criteria for selecting translations into the curated subset included several key factors:

1. Absence of Pre-existing Errors: Only translations that did not contain obvious grammatical mistakes, misalignments, or mistranslations were included. This selection ensures that any detected discrepancies in the QE process are due to the newly introduced synthetic errors rather than inherent issues in the original text.

2. Linguistic Simplicity and Clarity: Sentences chosen for the curated dataset were required to have clear, straightforward linguistic structures. This was to minimize the risk of the QE models misinterpreting linguistic complexity as an error.

3. Representative Linguistic Content: The sentences needed to be representative of typical real-world translations but devoid of overly complex or domain-specific jargon unless it was directly relevant to the error being introduced. This ensures the dataset's applicability across different QE scenarios.

4. Balanced Error Representation: The dataset was designed to include a balanced representation of each error type to evenly test the QE models' ability to identify and score different kinds of errors.

While in the selection process for creating this subset, special attention was given to choosing sentences of medium length. This decision was made to avoid overly complex, very short, or abbreviated linguistic forms that might pose comprehension challenges for the QE models. For example, sentences like the following were not included in the curated subset:

- "Requirements to be met by analytical procedure for dioxins and dioxin-like PCBs"

- " write to data address %X failed:%s"

These decision ensure that the QE models are tested under conditions that are more useful for accurate quality estimation without being confused by unusual linguistic structures. The results from this dataset will provide insights into how QE models respond to specific, isolated errors, contributing to a more nuanced understanding of their capabilities and limitations. However, these sentences remained in the original dataset and were included during the experiments.

3.1.5 Dataset Expansion

To ease the process of prompting specific alterations in the sentences, I first extracted sentences containing Named Entities and sentences containing numbers. Because the prompting will be done by choosing random sentences for each error category, it is more efficient to filter out sentences with named entities and numbers and prompt these separately. The Natural Language Toolkit (nltk) was employed for Named Entity Recognition (NER), identifying proper nouns, organizations, and other entities that require precise translation. Regular expressions (regex) were utilized to detect and extract numerical information from the text. After, regex was also utilized to alter the numbers, so this error category was not altered with the use of GPT. This extraction led to a duplication of sentences. Therefore, a duplicate sentence can contain one error in Named Entities and the other one in hallucinations. This led to an expansion to a total of 34,712 segments.

The rest of the errors- negation/sentiment, hallucinations, and additions- were created only by GPT. The error categories in the original dataset are broadly distributed, with 'Entity' and 'Number' errors being the most common, as many of the filtered sentences contained entities and numbers. The rest of the errors are more or less evenly divided.

3.2 Synthetic Dataset Creation

This section outlines the methodology used to create a synthetic test dataset using GPT-3.5-turbo and GPT-4-turbo. The original dataset, obtained from TAUS, contains machine-translated sentence pairs from Dutch (source) to English (target). For this project, these translations are necessary to systematically introduce specific types of translation errors into the target sentences. These error categories include named entities, numbers, negations/sentiment, deletion, addition, and hallucinations. The implementation of these errors in the dataset aims to challenge the robustness of MTQE models across varied error types.

Throughout the project, I carried out experiments using both GPT-3.5-turbo and GPT-4-turbo to determine which model would be more efficient in generating the intended artificial errors within the dataset. My analysis revealed that GPT-4-turbo consistently and accurately outperformed GPT-3.5-turbo in terms of following the structured prompts and generating believable, error-specific alterations. As a result, the final synthetic dataset predominantly contains outputs from GPT-4-turbo, ensuring a higher

level of accuracy in simulating errors. The error analysis in chapter 4.4 will provide more information on the difference in outputs by GPT-3.5-turbo and GPT-4-turbo.

3.2.1 Selection of Base Dataset

The base dataset consists of sentence segments from MT output provided by my internship at TAUS. These sentences are not free from errors, exhibiting typical machine translation inaccuracies to varying degrees. However, this inaccuracy should not hinder the project's objectives; instead, it adds a layer of complexity. When comparing the source sentences to the original target sentences and the synthetically distorted sentences, the QE models should, in principle, be able to identify the original errors as they appear in both the original target and distorted sentences. As entirely correct translations are rare in natural conditions, it would be interesting to determine the detection of these implemented errors regardless of possible surrounding errors by the QE models.

However, by doing so, it might be hard to create a totally controlled environment where it is easy to identify whether the QE models' score is based on the original errors, the implemented errors, or a combination of both. To ensure a more reliable experimental environment, I created a curated subset of the base dataset that filtered out low-quality translations. With 'low quality,' I refer to sentences containing natural errors in the original translations, such as grammar errors, existing mistranslations, and wrong sentence alignment between source and target sentences. This subset serves as a cleaner baseline, where the introduced errors are the only variables. By isolating the distortions, the QE models can focus solely on the impact of the newly introduced errors rather than the previous noise in the original dataset. This setup is particularly valuable for evaluating the QE models' sensitivity and precision in detecting and quantifying specific error types introduced during the experiment, which is the main objective of this project.

This methodology attempts to apply both natural translations with realistic linguistic errors and a cleaned sample to recognize and evaluate errors in a more controlled form. This two-way approach can evaluate the performance of QE models in both typical and ideal scenarios. To avoid confusion, from here on, I will name the original dataset the 'base dataset' and the subset the 'curated dataset'.

3.2.2 GPT-3.5-turbo and GPT-4-turbo

In this project, the synthetic dataset modification utilized two advanced language models, GPT-3.5-turbo and GPT-4-turbo, developed by OpenAI. These models are highly advanced in natural language processing technology, leverage deep learning techniques, and are based on the transformer architecture (Vaswani et al., 2017), which allows for highly effective generation of human-like text.

Features of GPT-3.5-turbo and GPT-4-turbo

GPT-3.5-turbo and GPT-4-turbo are characterized by their large number of parameters, with GPT-4-turbo being more advanced with even greater parameter count and improved training algorithms. These models have been trained on diverse internet text, making them highly versatile for a range of applications including but not limited to translation, summarization, question-answering, and, in this case, synthetic dataset

creation (Brown et al., 2020). The 'turbo' versions of these models are optimized for faster inference, making them particularly suitable for applications requiring rapid text generation.

3.2.3 Error Introduction Strategy

For the purpose of this project, GPT-3.5-turbo and GPT-4-turbo are employed to introduce specific types of accuracy errors into existing translated sentences. To simulate common MT errors, a series of structured prompts are used to direct the language models in modifying the target sentences of the base dataset. These prompts are designed to induce specific error types, such as:

- Negation Errors: The model is prompted to alter the meaning of sentences by strategically adding or removing negations. This could be in the form of 'no' or 'n't' but also implementing prefixes and suffixes into the words to change their meaning. This manipulation aims to test the QE model's ability to detect subtle semantic shifts. The prompt used is:
  - ('negation', 'In the following sentence, reverse the meaning by using negation. either introduce or remove a negation while keeping the exact original words and structure of the sentence intact and consider using suffixes or prefixes like un-, im-, in-, -less, ir-, and dis- where appropriate.'),

All prompts are viewable in the Appendix A.

3.2.4 Post-generation Review

During generation, a small sample set was created to verify whether the GPT models acted according to the prompts, and after several attempts at prompt engineering, the models proved to be altering the sentences correctly. After each generation, the samples were manually reviewed for each error category to verify whether the sentence modification was done accordingly. After modifying the base dataset, a small sample was again manually reviewed for errors. This process is described in full detail in the error analysis in section 4.4.

3.3 Comparison of QE Models

This section explores the unique capabilities and expected performance of the four QE models used in this research: CometKiwi (Rei et al., 2022), TransQuest (Ranasinghe et al., 2020), LASER (Artetxe and Schwenk, 2019), and LaBSE (Feng et al., 2020). Each model possesses distinct attributes that potentially affect their performance in detecting and evaluating translation errors.

CometKiwi is designed to leverage deep contextual embeddings, which allow it to perform intricate analyses of linguistic subtleties. This model is expected to excel in identifying complex error types such as nuanced semantic shifts and contextual inaccuracies. Its depth in processing and analysis makes it suitable for comprehensive evaluation tasks where detailed linguistic feedback is crucial.

TransQuest focuses on a streamlined, efficient approach to quality estimation. It is expected to provide robust performance in rapid assessment scenarios, making it

ideal for real-time applications. Although it might not delve as deeply into linguistic complexities as CometKiwi, this model can be particularly effective in settings where computational resources are limited or when quick estimations are needed.

LASER stands out with its language-agnostic feature set, designed to create universal embeddings that capture semantic similarities across languages. This model is anticipated to be particularly effective in scenarios involving less common language pairs or translations where direct comparisons might be challenging. Its ability to generalize across languages could be critical for projects requiring broad linguistic coverage.

LaBSE also focuses on embedding generation but incorporates an additive margin softmax loss, which enhances its ability to distinguish between different levels of translation quality. This feature is expected to make LaBSE exceptionally good at grading subtle quality variations, providing a more granular insight into the quality levels of translations.

Overall, my expectation is that while all models will perform effectively, their individual specialties might lead them to excel in different aspects of the quality estimation tasks. The upcoming experimental results (4) will provide deeper insights into how these theoretical advantages translate into practical performance.

3.3.1 Multilingual Capabilities of the Models

While all four QE models are designed to handle multilingual inputs, there are distinct nuances in how they are suited for multilingual contexts. Each model uses different strategies to achieve language-agnostic capabilities.

CometKiwi and TransQuest both incorporate transformer-based architectures that have been trained on extensive multilingual corpora. CometKiwi, as part of the COMET framework, uses cross-lingual sentence embeddings to understand and evaluate texts across different languages, ensuring consistent performance. Similarly, TransQuest utilizes a Siamese network structure with RoBERTa models, focusing on semantic alignment between the source and target texts, which allows it to handle numerous language pairs efficiently.

However, LASER and LaBSE might be particularly better suited for certain multilingual applications due to their specific focus on creating truly language-agnostic embeddings. LASER employs a specific training method on a diverse set of languages, using a BiLSTM architecture with max-pooling to generate embeddings that capture deep semantic meanings regardless of the language. LaBSE extends this language-agnostic approach by incorporating an additive margin softmax loss in its training, which not only helps in generating robust embeddings but also sharpens its ability to distinguish between subtle differences in translation quality across languages.

3.4 Evaluation Techniques

This section outlines the methods used to evaluate the performance of MTQE models and the effectiveness of the synthetic test dataset. This evaluation includes manually checking the GPT prompt outputs, statistical measures, and lastly the Kullback-Leibler divergence.

3.4.1 Manual Error Checking

Throughout the creation and analysis of the synthetic dataset, manual error checking played a pivotal role. Initially, after generating synthetic errors using GPT-3.5-turbo and GPT-4-turbo, a sample of modified sentence underwent a thorough review to ensure that the errors accurately reflected the intended types, such as negations, additions, and deletions. This step verifies that the prompts are effectively guiding the language model to simulate realistic translation errors. Following the application of QE models to these translations, another round of manual reviews is conducted on a sample of the scored translations. This process is designed to identify any discrepancies or anomalies in the model assessments, ensuring that the QE models are accurately identifying and scoring the introduced errors.

3.4.2 Graphical Visualization of Results

To illustrate the outcomes of the QE models' assessments, graphical visualizations are utilized. These visualizations, which will be featured in the results section of the study (4), include predicted quality score distribution graphs by all the QE models. The quality score distribution graphs will highlight how each model scores the quality of translations, with a focus on revealing differences in sensitivity to various error types. These visual aids are essential for quickly understanding patterns in the models' performance and effectively communicating these findings.

3.4.3 Statistical Analysis

In addition to visual tools, quantitative analysis is conducted using standard statistical measures. This includes calculating the median scores to determine the central tendency of the quality scores provided by each model, as well as computing the standard deviation to assess the variability in these scores. Such variability indicates the consistency of the model evaluations.

3.4.4 KL-Divergence

To assess and compare the performance of MTQE models on both the original and synthetically distorted sentences, the Kullback-Leibler (KL) divergence is utilized (Kullback and Leibler, 1951). This statistical measure is employed to quantify how one probability distribution diverges from a second, expected probability distribution. In this context, KL divergence will help in understanding how the error predictions by QE models differ from the original distribution of errors introduced into the dataset and their difference from the curated distribution without errors.

Background

KL divergence, also known as relative entropy, is a measure from the field of information theory that quantifies the difference between two probability distributions (Kullback and Leibler, 1951). For discrete probability distributions P and Q defined on the same probability space X the KL divergence from Q to P is given by:

$$D_{KL}(P\|Q) = \sum_{x\in\mathcal{X}} P(x) \log\left(\frac{P(x)}{Q(x)}\right)$$

Where $P(x)$ is the distribution of original translations, and $Q(x)$ is the predicted distribution of distorted translations by the QE model.

Application in QE Model Evaluation

In this research, the KL divergence is applied in two key scenarios:

1. Comparison between distorted and original sentences: By evaluating how the QE model's predictions for undistorted/natural sentences diverge from those for explicitly distorted sentences, insights into the model's sensitivity and robustness to different types of errors can be gained.

2. Cross-Dataset Comparisons: KL divergence provides a method to compare different datasets in terms of how their probability distributions of predicted errors or predicted 'good translations' deviate from the base to curated datasets (and vice versa). This analysis not only helps validate the robustness and completeness of each dataset but also understands how the size of a dataset might influence its ability to test QE models that are sensitive to different translation errors.

Chapter 4

Experiments

This chapter details the results of the experiments on the QE models tested on machine-translated Dutch to English sentences, and the error analysis, that gives further findings of the results. The experiments are designed to assess the capability of QE models to identify and quantify translation quality in both original and intentionally distorted sentences. Various QE models are evaluated, including CometKiwi (Rei et al., 2022), TransQuest (Ranasinghe et al., 2020), LASER (Artetxe and Schwenk, 2019), and LaBSE (Feng et al., 2020), to understand their performance across different types of errors and translation domains.

4.1 Results

This section presents the results of the QE experiments conducted on both the original and curated datasets. These experiments were designed to assess the performance of four QE models: CometKiwi (Rei et al., 2022), TransQuest (Ranasinghe et al., 2020), LASER (Artetxe and Schwenk, 2019), and LaBSE (Feng et al., 2020). The performance of these models was evaluated based on their ability to distinguish between original translations and those that had been intentionally distorted.

The primary goal of the QE models in this study is to effectively differentiate between original translations and distorted translations. Distorted translations include artificially introduced errors, which are expected to be scored lower by the QE models than the original sentences.

I am expecting the curated dataset to provide a clearer distinction between original and distorted translations compared to the original dataset. This is because the curated dataset has been meticulously created to remove any pre-existing errors in the original translations, thereby providing a more controlled environment for evaluating the QE models. By isolating the artificially introduced errors, the curated dataset should enhance our understanding of each model's sensitivity to specific error types and its overall effectiveness in quality estimation tasks.

4.1.1 Results of Base Dataset

Figure 4.1 shows the overall distribution of QE scores for both original and distorted sentences across all four models. A clear separation between the score distributions would indicate the models' effectiveness in detecting translation errors.

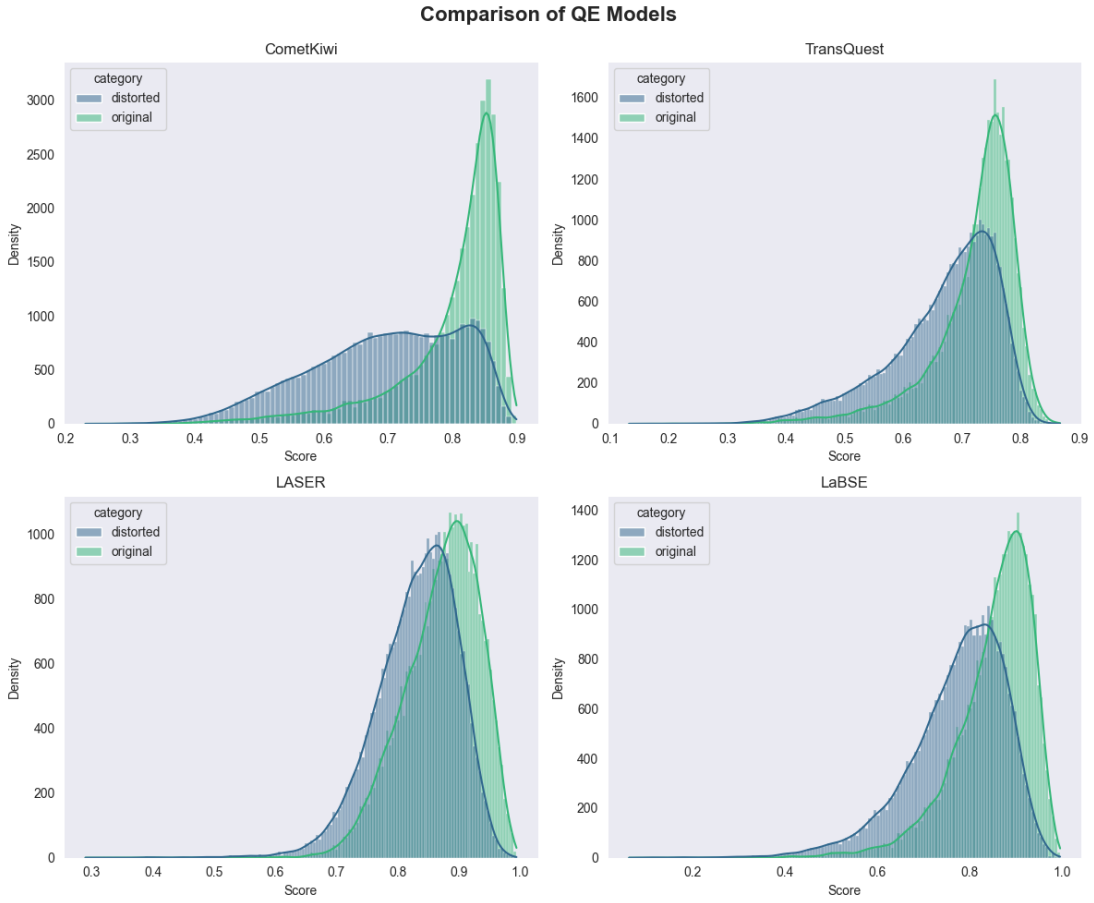

Figure 4.1: Distribution of Original and Distorted Sentence Scores by all QE Models on the Base Dataset

CometKiwi shows a noticeable separation between original and distorted translations, indicating high sensitivity to translation errors. TransQuest demonstrates a moderate distinction while LaBSE shows a little les distinction and finally LASER shows the least distinct seperation, indicating challenges in distinguishing between original and distorted translations.

The second Table (Table 4.1) displays the KL divergence for original to distorted translations and vice versa for each QE model using the base dataset. All the models exhibit a relatively symmetrical behavior, indicating a similar level of divergence in both directions. CometKiwi shows a slightly higher level of divergence, suggesting that it may be more sensitive to differences between distorted and original sentences compared to the other models. This implies that the remaining models might treat original and distorted translations more similarly than CometKiwi. It's important to note that the "original" translations in the base dataset are not entirely error-free and may contain errors, which is why the subset was carefully selected.

QE model	Divergence	
	original to distorted	distorted to original
CometKiwi	0.564	0.570
TransQuest	0.233	0.233
LaBSE	0.406	0.409
LASER	0.250	0.211

Table 4.1: KL Divergence - Base Dataset

CometKiwi

This Figure shows the distribution of CometKiwi scores for each error category (4.2). CometKiwi shows varied sensitivity across different error categories with the highest sensitivity to hallucination and lowest sensitivity to numbers. It also shows a great distinction between original and distorted sentences for gibberish, sentiment, and negation.

The statistical measures for CometKiwi as shown in Table 4.2 highlight its performance across different error categories. The median and standard deviation values provide insight into the central tendency and variability of the scores, respectively. A higher standard deviation in the distorted sentences suggests greater variability in scoring, which may be attributed to the model's response to the range of errors introduced in the synthetic dataset. It is noTable that the mean and median of the distorted sentences are considerable lower than the original sentences, which indicates that CometKiwi has identified the additional artificial errors to some extent even though the minimum and maximum scores do not differ extensively.

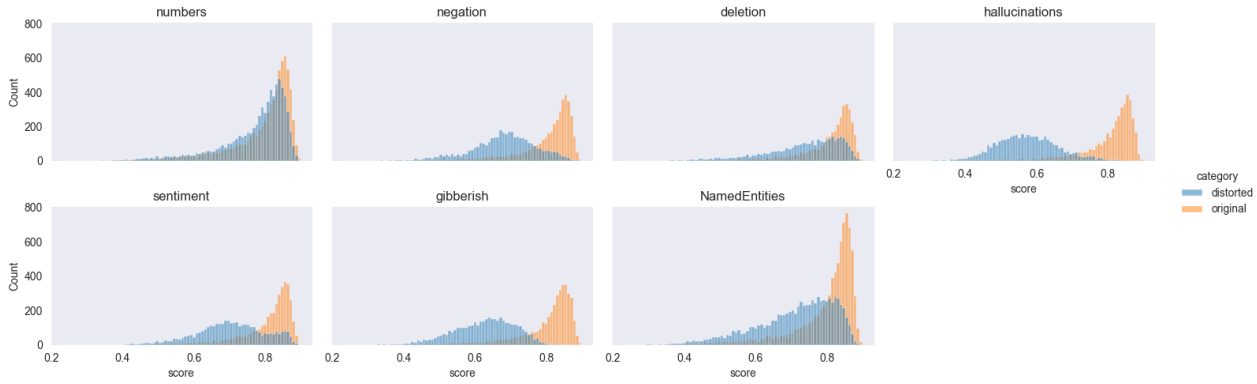

Figure 4.2: Distribution of CometKiwi Scores per Error Category

Statistic	Original	Distorted
Mean	0.7991	0.6955
Median	0.8293	0.7075
Std Dev	0.0849	0.1145
Min	0.2933	0.2307
Max	0.8988	0.8952

Table 4.2: CometKiwi QE Model Scores - Base Dataset

TransQuest

Similar to the previous Figure, Figure 4.3 presents the TransQuest score distributions per error category. TransQuest performs consistently across most error categories with most sensitivity towards hallucinations, and gibberish and less sensitivity towards deletion, named entities, and numbers.

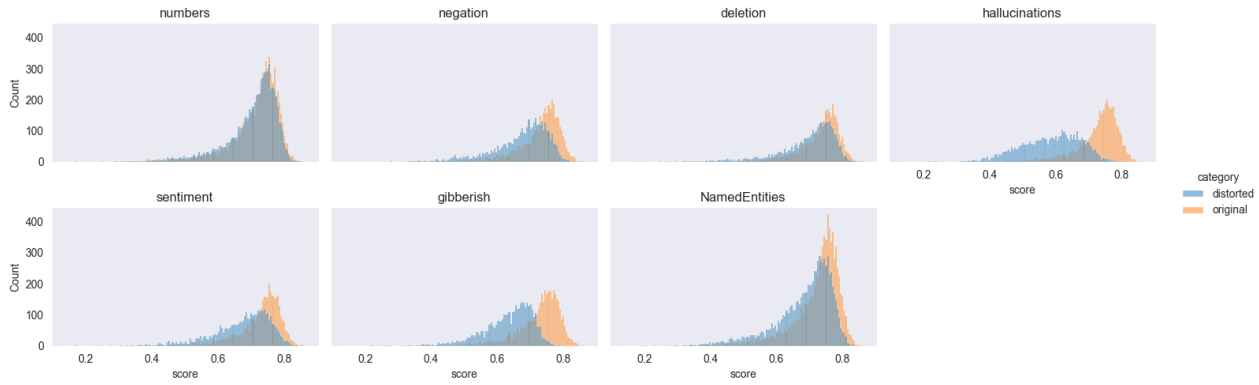

Figure 4.3: Distribution of TransQuest Scores per Error Category

In Table 4.3, the TransQuest scores show moderate variability as indicated by the standard deviation. The median scores between original and distorted translations differ less significantly compared to CometKiwi, suggesting that TransQuest may have a more uniform scoring range, which affects its sensitivity to certain error types. Also noTable is that the maximum score is higher for the distorted sentence than the original one, which indicates that adding an additional error created a higher score in one or more of the sentences.

Statistic	Original	Distorted
Mean	0.7216	0.6690
Median	0.7412	0.6897
Std Dev	0.0744	0.0919
Min	0.1949	0.1319
Max	0.8649	0.8672

Table 4.3: TransQuest QE Model Scores - Base Dataset

LASER

Contrasting to the previous models, LASER appears more consistently insensitive to all error types (see Figure 4.4). Hallucinations and named entities have a slightly more distinct difference between original and distorted translations.

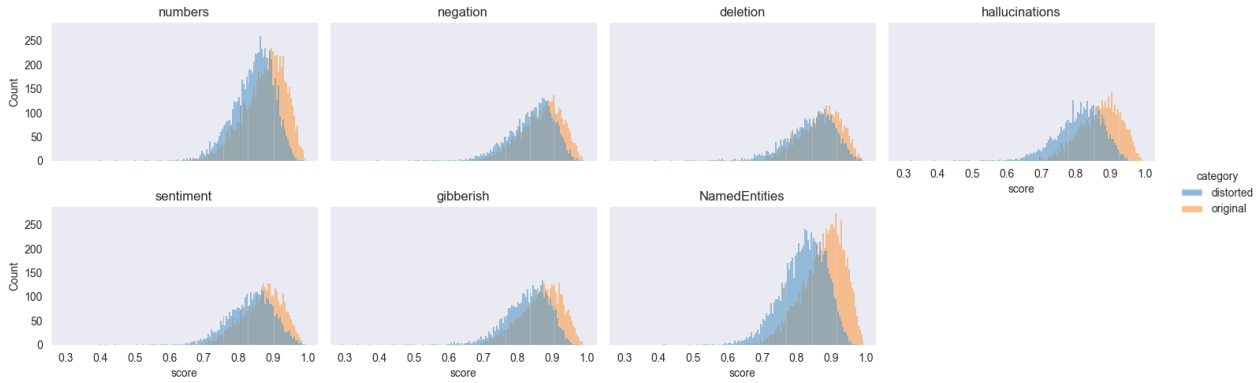

Figure 4.4: Distribution of LASER Scores per Error Category

Statistic	Original	Distorted
Mean	0.8712	0.8307
Median	0.8795	0.8387
Std Dev	0.0618	0.0668
Min	0.2917	0.2898
Max	0.9944	0.9929

Table 4.4: LASER QE Model Scores - Base Dataset

The LASER model, as shown in Table 4.4, demonstrates a lower standard deviation compared to CometKiwi and TransQuest, indicating a more consistent scoring behavior across the range of translations. Furthermore, the differences in mean and median scores are closer between the distorted and original translations. This could suggest that LASER is less sensitive to the variability of errors within the distorted translations.

LaBSE

LaBSE seems to perform well on named entities, numbers, and hallucinations (see Figure 4.5). All previous models showed a relatively high insensitivity towards errors containing numbers, so LaBSE seems to perform best on numbers among the four QE models.

Table 4.5 illustrates that LaBSE has a higher standard deviation for distorted translations, which indicates greater score dispersion. This could reflect LaBSE's sensitivity to the nuanced differences in translation quality introduced by the errors. The median scores also show a notable drop from original to distorted translations, which is indicative of the model's effective differentiation between the two sets.

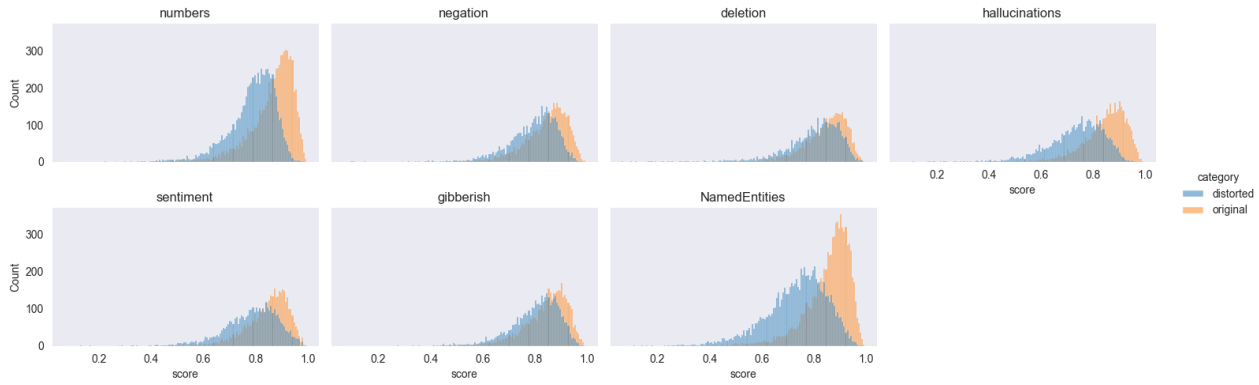

Figure 4.5: Distribution of LaBSE Scores per Error Category

Statistic	Original	Distorted
Mean	0.8516	0.7735
Median	0.8697	0.7912
Std Dev	0.0848	0.1055
Min	0.2627	0.0627
Max	0.9957	0.9946

Table 4.5: LaBSE QE Model Scores - Base Dataset

4.1.2 Results Curated Dataset

Figure 4.6 shows again the overall distribution of QE scores for both original and distorted sentences. The curated dataset results show similar trends to the base dataset, with CometKiwi having the highest sensitivity to errors followed by TransQuest and LaBSE with LASER having the lowest sensitivity to errors.

Moving on to Table 4.6, it presents a similar divergence but for the curated dataset. The KL divergence values are generally higher in the curated dataset, especially evident in CometKiwi from distorted to original translations (2.011). The higher KL divergence in the curated dataset suggests that when the data is well-created, the QE models, especially CometKiwi, are capable of effectively distinguishing quality. This indicates that the model responds well to high-quality, well-defined data.

QE model	Divergence	
	original to distorted	distorted to original
CometKiwi	1.013	2.011
TransQuest	0.347	0.427
LaBSE	0.487	0.581
LASER	0.279	0.257

Table 4.6: KL Divergence - Curated Dataset

While the higher divergence in the curated dataset is promising, it also emphasizes the significance of data quality in training and evaluating QE models. Models may perform differently when trained or tested on datasets of varying quality, impacting

their reliability.

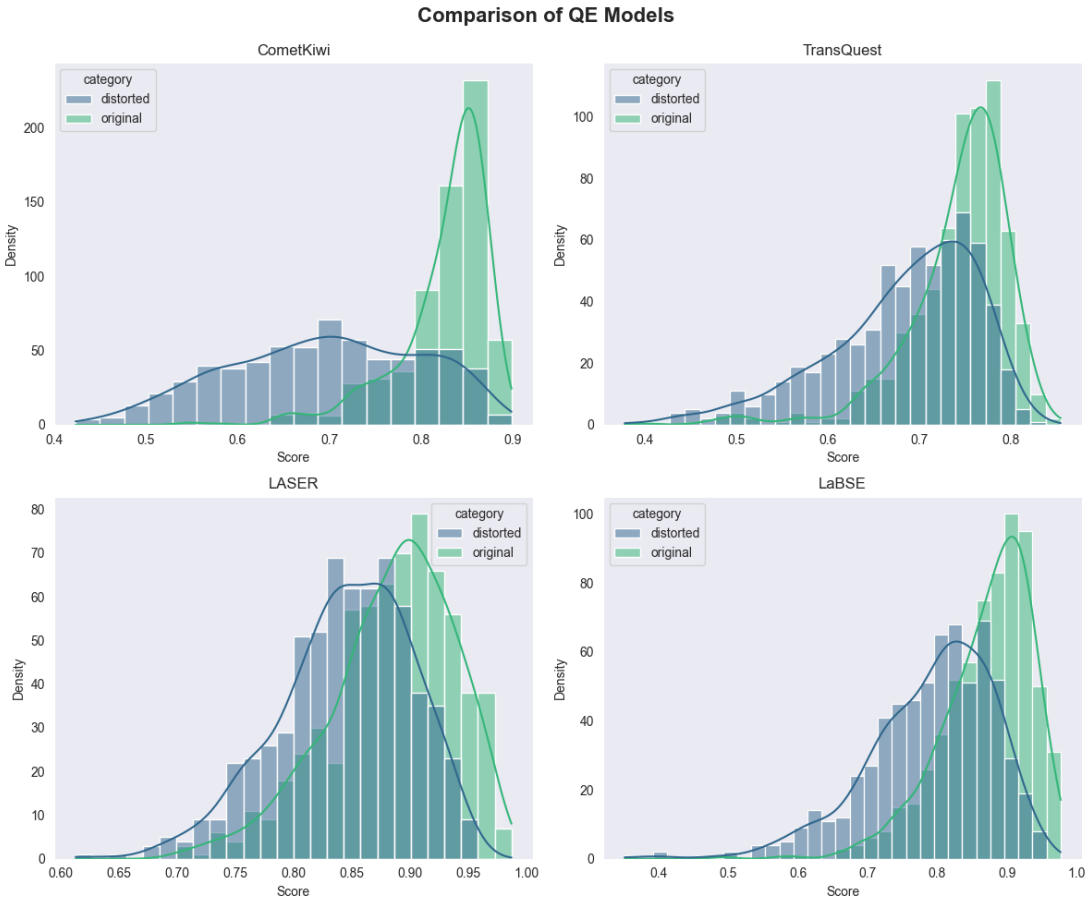

Figure 4.6: Distribution of Original and Distorted Scores by all the QE Models

CometKiwi

Similar to the base dataset, CometKiwi seems to be able to differentiate between original and distorted translations for all the error categories except numbers (see Figure 4.7).

Statistic	Original	Distorted
Mean	0.8260	0.6956
Median	0.8412	0.7018
Std Dev	0.0506	0.1043
Min	0.5432	0.4231
Max	0.8988	0.8802

Table 4.7: CometKiwi QE Model Scores - Curated Dataset

In the curated dataset, CometKiwi's performance metrics (Table 4.7) continue to show a high median and a noTable increase in the standard deviation for distorted sentences. This indicates that CometKiwi responds sensitively to the range of intro-

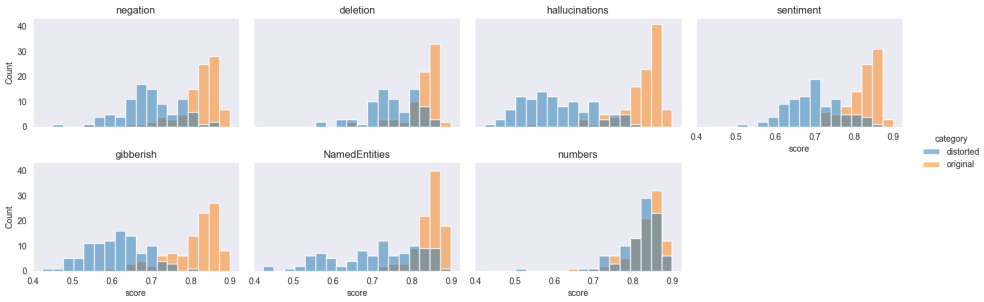

Figure 4.7: Distribution of CometKiwi Scores per Error Category

duced errors, which aligns with the observed higher KL divergence, suggesting a strong discriminative ability.

TransQuest

TransQuest shows high sensitivity for hallucinations, and gibberish and lowest sensitivity for numbers (see Figure 4.8).

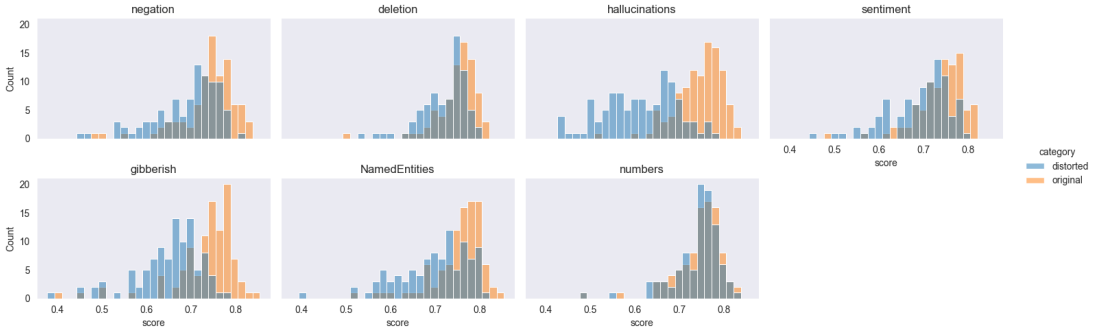

Figure 4.8: Distribution of TransQuest Scores per Error Category

Statistic	Original	Distorted
Mean	0.7430	0.6833
Median	0.7555	0.7006
Std Dev	0.0579	0.0811
Min	0.4012	0.3781
Max	0.8545	0.8222

Table 4.8: TransQuest QE Model Scores - Curated Dataset

TransQuest's results in the curated dataset (Table 4.8) demonstrate a consistent detection capability across error types, with a moderate increase in standard deviation for distorted sentences, indicating its response to the synthetic errors is sTable but less pronounced than CometKiwi.

LASER

Similar to the base dataset, Figure 4.9 illustrates that LASER has the highest sensitivity for named entities, while the rest of the errors show relatively low sensitivity.

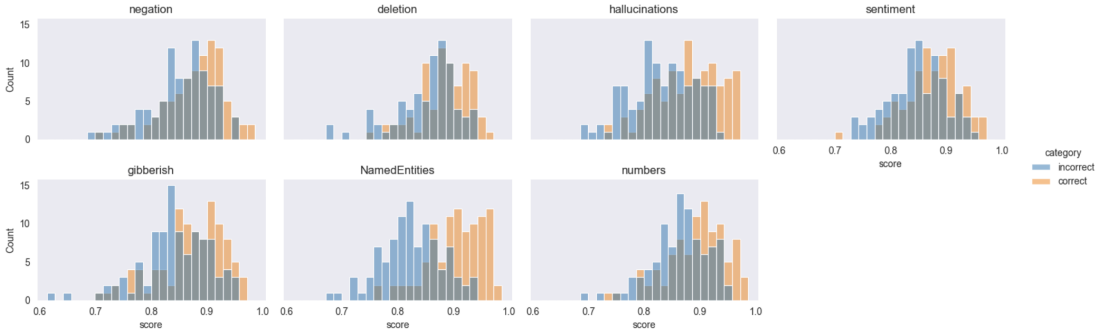

Figure 4.9: Distribution of LASER Scores per Error Category

Statistic	Original	Distorted
Mean	0.8842	0.8444
Median	0.8915	0.8479
Std Dev	0.0538	0.0573
Min	0.6999	0.6131
Max	0.9869	0.9547

Table 4.9: LASER QE Model Scores - Curated Dataset

LASER maintains consistent performance with minimal variation in the scores' standard deviation, as shown in Table 4.9. This model's scores suggest a relatively sTable but less sensitive detection of errors compared to the other models, as also reflected by the lower KL divergence values.

LaBSE

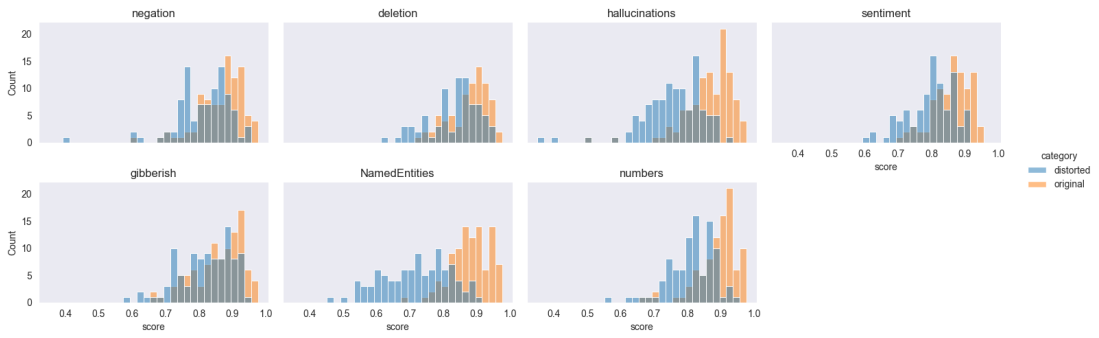

Figure 4.10: Distribution of LaBSE Scores per Error Category

Lastly, Figure 4.10 shows the distribution of scores for each error types and LaBSE seems to perform best on numbers again compared to the other QE models even on

Statistic	Original	Distorted
Mean	0.8710	0.7912
Median	0.8846	0.8071
Std Dev	0.0655	0.0901
Min	0.4927	0.3512
Max	0.9771	0.9498

Table 4.10: LaBSE QE Model Scores - Curated Dataset

the curated dataset. Furthermore, LaBSE shows a higher distinction of errors for hallucinations and named entities. The rest of the errors are relatively similar in their low sensitivity levels.

LaBSE's results (Table 4.10) show a noticeable difference in the median and an increase in standard deviation for distorted translations, highlighting its capability to effectively differentiate between the varying quality of translations, especially in a cleaner dataset setting. The higher divergence and variability suggest that LaBSE is responsive to the nuanced differences introduced in the curated dataset.

4.1.3 Domains

The Figures below demonstrate the overall QE score distributions of original and distorted sentences per domain. These scores are accumulated per domain to evaluate the models' performance accross different contexts. For both datasets the 'undefined sector' and 'Professional and Business Services' domains show highest distinctions between original and distorted translations.

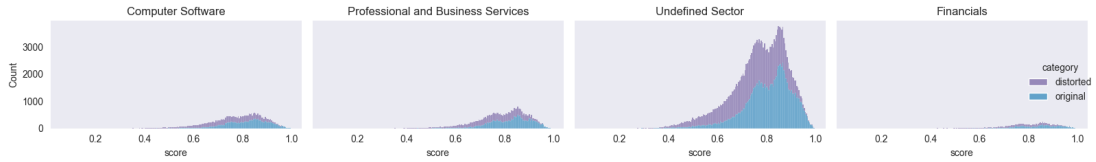

Figure 4.11: Distributions of Original and Distorted QE Scores per Domain for the Original Dataset

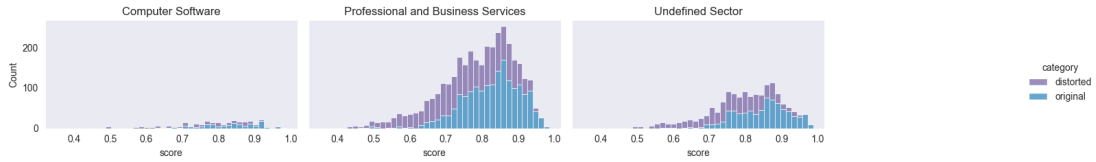

Figure 4.12: Distributions of Original and Distorted QE Scores per Domain for the Curated Dataset

4.2 KL-Divergence between Datasets

Table 4.11 compares the divergence between the curated and base datasets based on distorted sentence evaluations. All models exhibit very low divergence values, with

QE model	Divergence	
	Curated to Base	Base to Curated
CometKiwi	0.020	0.032
TransQuest	0.021	0.038
LaBSE	0.019	0.050
LASER	0.026	0.118

Table 4.11: KL Divergence - Comparison of Curated and Base Dataset on Distorted Sentences

slightly higher values when comparing the original base to curated datasets. The low divergence suggests that the models are relatively consistent in evaluating distorted sentences across dataset sizes, although there is slightly more variation when moving from the base to the curated dataset. LaBSE shows the lowest divergence from curated to base, suggesting high consistency in scoring the curated dataset. LASER exhibits the highest divergence in both directions, particularly from base to curated. This suggests that LASER is highly sensitive to the inherent errors in the base dataset, resulting in significant changes in its score distribution when those errors are present.

QE model	Divergence	
	Curated to Base	Base to Curated
CometKiwi	0.072	0.436
TransQuest	0.060	0.093
LaBSE	0.034	0.067
LASER	0.027	0.137

Table 4.12: KL Divergence - Comparison of the Curated and Base Dataset on Original Sentences

Lastly, in Table 4.12, the divergence in evaluations of original sentences between the two dataset sizes is measured. Divergence values are generally low, similar to the distorted sentence evaluations, but with slightly higher divergence when transitioning from base to curated dataset. There is generally consistency across dataset sizes for original sentences, but there are indications that moving to a curated dataset increases divergence, possibly due to the accuracy improvements of original sentences in the curated dataset.

4.3 Impact of Curated Dataset

The statistical data, including KL divergence, mean, and median statistics, provides valuable insights into the performance consistency of QE models across different datasets.

The results indicate that all four MTQE models are affected by the inherent errors in the base dataset to varying degrees. LASER appears to be the most sensitive, followed by LaBSE, TransQuest, and CometKiwi.

Moreover, it shows slightly higher divergence for original translations. This suggests that the models are potentially more sensitive to the cleaner, more accurately translated sentences in the curated dataset, reflecting their capability to appreciate higher quality

translations. The results suggest that CometKiwi is the most sensitive to inherent errors in the base dataset, showing the highest divergence in the base to curated direction. TransQuest, LaBSE, and LASER are less sensitive, with LaBSE being the most stable across both datasets.

The curated dataset, which filters out inherent errors and only includes artificial distortions, results in a more consistent score distribution across the models. This suggests that curating a smaller subset of high-quality sentences without inherent errors provides a clearer challenge set for evaluating MTQE models.

Additionaly, the difference in mean and median scores between original and distorted sentences increase for the curated dataset, especially with CometKiwi. This increase suggests that models are more sensitive and score either the distorted sentences lower or the original sentences higher when working with high-quality data.

4.4 Error Analysis

In this section, I will analyze various errors produced by GPT-3.5-turbo and GPT-4-turbo during our prompting experiments. These analyses help in understanding the nature of unwanted outputs and how they might influence the QE models.

4.4.1 Analysis during Prompting of GPT-3.5-turbo

Named Entities

A common issue when altering named entities was that GPT-3.5-turbo often replaced more than just the entities, substituting many words with synonyms. For example:

- MT: “The commission provides support and arranges contacts between firms working in similar areas of research.”
- Distorted: “The alliance offers assistance and facilitates connections between companies operating in related fields of study.”

Although the semantics remain the same, these changes between original translations and distorted sentences can influence QE scores, impacting our controlled environment.

Similarly, GPT-3.5-turbo sometimes changed words within the same phrase as the named entities, particularly in possessive noun phrases:

- MT: “Firstly, the motion does not condemn clearly enough the vagueness of the priorities and commitments indicated in the **Commission’s programme.**”, ’
- Distorted: “Firstly, the motion does not condemn clearly enough the vagueness of the priorities and commitments indicated in the **Council’s agenda.**”,

Another frequent issue was uncreative replacements for newly named entities, often resulting in nonsensical abbreviations like ‘CBFA’ becoming ‘XYZ’ and ‘Frankfurt’ becoming ‘ABC.’ Additionally, numbers were sometimes randomly removed, such as ‘amended)1’ becoming ‘amended’.

Sentiment/Negation

When altering sentiment, GPT-3.5-turbo occasionally produced outputs with the same sentiment as the original sentence but with different causation:

- MT:"**Poorly** managed coastal resorts can also **cause** serious air and sea pollution."
- Distorted: "**Well** managed coastal resorts can also **prevent** serious air and sea pollution."

To address this, an additional prompt line was added: 'Change the sentiment so the meaning of the sentence changes.'

In cases where sentences had no clear sentiment to change, negations were included instead. However, if neither sentiment change nor negation was possible, I prompted GPT to leave the sentence unchanged and then such sentences were filtered out.

GPT-3.5-turbo also tended to finish incomplete sentences, adding phrases like:

- MT: "Surviving veterans of World War I"
- Distorted: "Surviving veterans of World War I are becoming increasingly rare."

MT Hallucinations and Gibberish

These two error categories seemed to operate fairly well. Their only lack is that the chosen words were repetitive. For hallucinations, words like 'refrigerator', 'giraffe', and 'sandwich' were frequently used. For non-existing gibberish 'flibber' was commonly used.

Deletion

For deletion errors, GPT-3.5-turbo often removed function words like 'the' and 'a', instead of content words, despite multiple prompt adjustments. Also, often times it failed to remove any words from sentences.

Grammar

When comparing the full dataset with naturally occurring errors to the more distorted sentences, GPT-3.5-turbo often corrected grammar issues in the original target sentences along with the prompted changes. This included fixing punctuation errors or concatenated words like 'she's' to 'she is' or correcting spacing errors like 'regarding' to 'regarding.'

4.4.2 Analysis during Prompting of GPT-4-turbo

Named Entities

GPT-4-turbo performed better with named entities, although it occasionally replaced all words with synonyms, albeit less frequently than GPT-3.5-turbo. Repetitive simple abbreviations were used, but geographical locations were handled better:

- MT: "The original CUI, MNU, or MNS file is not modified."
- Distorted: "The original ABC, XYZ, or DEF file is not modified."

Sentiment/Negation

GPT-4-turbo improved on sentiment changes and avoided finishing incomplete sentences with creative phrases. However, it occasionally created non-existing negations:

- MT: “Surviving veterans of World War I”
- Distorted: “Non-surviving veterans of World War I”

Or with short sentences:

- MT: “short rest periods [...]”
- Distorted: “short unrest periods [...]

MT Hallucinations and Gibberish

These categories continued to perform well, with repetitive word use being the main issue. Moreover, especially in this version, I changed the output to directly implement sentences into a JSON file. Therefore, most of the distorted sentences concerning hallucinations contained:

- MT: “joint group of experts on the scientific aspects of marine pollution”
- Distorted: “joint group of experts on the scientific aspects of marine **ballet**”

For gibberish, some of the non-existing words would concatenate with existing words, which can cause confusion when categorizing this error.

- MT: “[...], and it will be **adopted** in the near future.”
- Distorted: “[...], and it will be **flibberadopted** in the near future.”

This additional error leads to not only an addition of gibberish but also a deletion of an existing word in the original translation. Therefore, the error is not fully isolated in these cases. The score representations become more blurry because it may be unclear whether a score represents the gibberish or the deletion/concatenation of a word.

Deletion

With deletion, GPT-4-turbo continued to remove function words as well as critical words. As I prompted it to delete important words in the sentence, it sometimes removed negations, creating a negation error instead of deletion error. Of course, one can argue that a negation error could also essentially be a deletion error as it removes negations from the sentence to reverse the sentiment.

This version still struggled with actually removing words, and similar to hallucinations, often included these examples:

- MT: “”Error writing stream to output. Output stream may be corrupted or truncated.”
- Distorted: “”Error writing stream to output. Output stream may be corrupted or **truncated**.”

General

GPT-4-turbo struggled with implementing sentences into a JSON structured file as desired. Instead of directly implementing the distorted sentence, it often included verbose descriptions of the changes:

1. "The modified sentence after deleting a critical word would be: (c1) be fitted with axles with:"
2. "relationships between society, economy, biodiversity and habitats; Removing the word and changes the sentence to: relationships between society, economy, biodiversity habitats; This alteration changes the meaning by"
3. "The sentence It shall report, in particular, on: does not contain a sentiment to reverse, as it is a neutral statement indicating a requirement or obligation to report specific information. Therefore, the sentence"

   While the modifications were correct, their structure proved unhelpful and was therefore filtered out of the final output.

4.4.3 Analyzing Low Scores in Curated Dataset

Despite the careful creation of a smaller dataset that does not include pre-existing errors, the results indicate lower scores exist in the original translations in the curated dataset. To analyze this, I examined the lowest-scoring original translations in the curated dataset. The lowest scores given to the original translations are in the 0.4 range. This includes seven sentences that include all QE models. However, it is noTable that six out of seven sentences scored this low on TransQuest. These sentences include:

- Source: "Wanneer de varkens ouder zijn dan drie tot vier maanden, moet bij een deel van de varkens ook de temperatuur worden gemeten."
- Target: "In case of pigs older than **three to four months**, this examination must include the taking of temperature of a proportion of pigs."
- Distorted target: "In case of pigs older than **four to five months**, this examination must include the taking of temperature of a proportion of pigs."

Or:

- Source: "BEPALINGEN IN VERBAND MET ELEKTRONISCHE GEGEVENSVERWERKING;"
- Target: "PROVISIONS GOVERNING ELECTRONIC DATA PROCESSING‘;"
- Distorted Target: "PROVISIONS **NOT** GOVERNING ELECTRONIC DATA PROCESSING;"

Most of these sentences are actually accurate in translation and would normally not acquire a low score such as 0.4. The fact that TransQuest gives low scores to these original translations, despite their accuracy, might suggest that the model may have specific sensitivity or bias towards certain linguistic features or constructions in the

text. This could be due to the training data used or the model's inherent architectural biases.

Furthermore, for these two example sentences, the distorted target acquired a higher QE score than the original translations. This might indicate that TransQuest is less adept at recognizing subtler nuances in translation quality. It could also imply that the model weights certain errors or changes less severely than the absence of those features.

4.4.4 Quantifying the Errors

The fact that TransQuest gives low scores to these original translations, despite their accuracy, suggests that the model might have specific sensitivity or bias towards certain linguistic features or constructions in the text. This could be due to the training data used for TransQuest or the model's inherent architectural biases. Moreover, the discrepancy in scoring between original and distorted sentences by TransQuest, where distorted sentences sometimes receive higher scores, might indicate that the model is less adept at recognizing subtler nuances in translation quality. It could also imply that the model weights certain errors or changes less severely than the absence of those features.

In order to objectively measure the impact of artificially implemented errors and their frequency, I conducted a quantitative analysis using a random sample of 100 sentences. This sample was specifically chosen to compare the original target sentences with the distorted versions produced by the GPT models. The analysis aimed to identify and quantify the types of errors introduced during the prompting process. See Table 4.13 for a distribution of errors in the sample.

Error Category	Occurances in Sample	Error Frequency
Named Entity	24	20,8%
Numbers	21	0%
Addition	14	35.7%
Deletion	12	25%
Negation	14	14.3%
Sentiment	16	12.5%
MT Hallucination	18	0%

Table 4.13: Distribution of Error Categories in the Sample

Analysis of Additions

The most common error observed in this sample was the deletion of words, which occurred in conjunction with the addition of gibberish in five out of 14 instances. This particular combination was the most frequent, highlighting a tendency of the models to remove essential words while adding nonsensical ones, potentially complicating the interpretation of the output. This deletion adds an unintentional error and might make the isolation of gibberish errors less controlled.

Analysis of Named Entities

In the context of named entities, the results were mixed. In two instances, the sentences remained unchanged despite the prompts, indicating a failure in the model's response

to the instruction. However, in three cases, the named entities were replaced with synonyms. These replacements did not significantly alter the meaning of the sentences from the source, they still represented a deviation from the target translation, however still remained similar to the source sentence, which could affect quality estimation.

Deletion Errors

Regarding the deletion prompts, the models frequently removed function words instead of content words, which occurred three times in this sample. Such deletions are not considered critical enough for this project and might influence the intended decrease of QE scores when comparing them to the original translations.

Sentiment and Negation Errors

Errors related to sentiment and negation were also noted. In two cases, the sentiment of the sentence remained unchanged because of multiple attempts to reverse it, as the models inadvertently reversed the sentiment twice within the prompting sequence. Additionally, two instances of non-existent negations were identified, introducing inaccuracies in the conveyed information.

General Observations

A general structure error was observed once, where the output included prompt notes, altering the structure of the sentence to include meta-information like "the original sentence was altered to [...]." This type of error, while only observed once in this sample, underscores potential issues in how models handle structured prompts.

Limitations

It is crucial to acknowledge that this sample of 100 sentences, while informative, represents only a fraction of the base dataset. Therefore, the observed frequencies and patterns may not fully capture the prevalence or distribution of errors across the base set of distorted sentences. This analysis serves as an indicative snapshot, useful for identifying trends and common issues but not definitive in scope.

4.4.5 Relating Errors to Literature Background

The error patterns observed in the prompting experiments with GPT models show considerable alignment with those discussed in the literature, particularly the impact of named entity errors, sentiment/negation inaccuracies, and hallucinations. These findings validate the relevance of the selected error categories for our challenge test sets.

However, there are discrepancies in the frequency and severity of some errors compared to those reported in the background literature. For instance, the higher occurrence of named entity and sentiment errors in our dataset suggests that these error types may be more challenging for the GPT models than previously reported in general MT contexts. However, It must be noted again that the sample size for my error analysis is small and might be inconclusive when regarding the entire dataset.

4.5 Summary of Main Results

- **QE Model Performance:**
  - CometKiwi showed the highest sensitivity to errors, especially to hallucinations, clearly distinguishing between original and distorted translations.
  - TransQuest demonstrated moderate sensitivity, with notable performance towards hallucinations but less effectiveness with deletion, named entities, and numbers.
  - LaBSE performed consistently well, particularly in handling named entity and number errors.
  - LASER had the least sensitivity, showing minimal distinction between original and distorted translations across error categories.
- **Error Analysis Insights:**
  - Errors involving inappropriate replacements, particularly with named entities and synonyms, were common and sometimes led to nuanced semantic shifts possibly affecting the controlled environment for predicting the scores.
  - Sentiment and negation errors underscored the complexity of translating sentiment accurately, with occasional model failures in altering or maintaining the correct sentiment.
  - Repetitive and contextually inappropriate choices were noted in hallucinations and gibberish errors, pointing to ongoing challenges prompt engineering.

Chapter 5

Discussion and Conclusion

5.1 Main Findings

This study aimed to answer the main question: "How effectively can current Machine Translation Quality Estimation models identify and quantify different types of translation errors introduced by advanced large language models in a Dutch-English dataset?" To address this, the study explored several sub-questions, each focusing on various aspects of the research.

5.1.1 Utilization of Generative Models for Creating Test Sets

One sub-question investigated whether generative models like GPT-3.5 and GPT-4 could be utilized to create challenge test sets by intentionally modifying sentences to include specific error patterns.

The study found that generative models such as GPT-3.5-turbo and GPT-4-turbo were indeed capable of generating synthetic datasets with specific error patterns. Using structured prompts, these models successfully introduced various types of translation errors, including named entities, numbers, negations, deletions, additions, and hallucinations. GPT-4-turbo outperformed its previous version by following the prompts slightly better and its output was therefore applied during the project. However, the study also highlighted certain limitations, such as the variability and potential inconsistency of the generative models' outputs. Furthermore, a human should be in the loop to identify whether these models accurately executed the prompts. My study showed flaws in the production of some of these prompts, but perhaps this can be improved by prompt engineering. Despite these challenges, the synthetic datasets provided a strong foundation for testing the sensitivity and performance of MTQE models.

5.1.2 Performance of MTQE Models

Another sub-question examined which MTQE model—CometKiwi, TransQuest, LASER, or LaBSE—performed superior when faced with these synthetic test sets.

Among the four models evaluated, it was found that CometKiwi exhibited the highest sensitivity to the introduced translation errors, particularly in identifying hallucinations and named entity distortions. TransQuest and LaBSE also performed well but were less sensitive to number errors. LASER, on the other hand, showed the least distinction between good and bad translations, indicating that it is least effective in detecting specific types of errors. The performance of these models varied significantly

across different error types, underscoring the importance of selecting MTQE models based on the specific error patterns relevant to the application context.

Table 5.1 summarizes the best to least performing QE models based on the error categories for both the base and curated dataset. CometKiwi outperformed the other models in all categories except numbers. In some cases where models performed similarly, the models are presented with a backslash.

Error Category	Best to Least
Named Entity	CometKiwi - LaBSE - LASER/TransQuest
Numbers	LaBSE - LASER - CometKiwi - TransQuest
Addition	CometKiwi - TransQuest - LASER/LaBSE
Deletion	CometKiwi - TransQuest/LASER/LaBSE
Negation	CometKiwi - TransQuest/LASER/LaBSE
Sentiment	CometKiwi - TransQuest/LASER/LaBSE
MT Hallucination	CometKiwi - TransQuest/LaBSE - LASER

Table 5.1: Summary of Best Performing QE Model per Error Category

5.1.3 Challenges Posed by Specific Error Patterns

The study, while identifying several error patterns that posed consistent challenges for the QE models, also highlighted the potential for improvement. Except for CometKiwi, most models struggled with almost all error categories. Mainly errors involving numbers, omissions, and negations/sentiments were particularly problematic, with models showing lower sensitivity to these types of distortions. However, the variability in performance across different error types and domains also presents opportunities for enhancing QE models in the future.

5.2 Discussion and Limitations

5.2.1 Model Sensitivity to Translation Errors

As mentioned, the study found that different MTQE models respond differently to various translation errors. The varying sensitivity of can mainly be attributed to their underlying structures. CometKiwi, for instance, exhibited increased sensitivity to errors such as hallucinations and distortions in named entities because of its design, which takes into account a wide range of linguistic aspects, including tone and fluency. This approach enables it to identify subtle flaws in text quality that models like LaBSE and LASER, which primarily evaluate semantic similarity through sentence embeddings, might miss. These differences in model architecture directly impact their capability to distinguish between high-quality translations and mistakes, underscoring the importance of selecting a QE model that aligns with the anticipated error types in the specific use case.

5.2.2 Impact of Data Quality

Furthermore, the comparison between the base dataset and the manually curated subset highlighted the importance of dataset quality. Higher KL divergence values in the

curated dataset suggest that well-created data may improve the ability of QE models to distinguish between correct and incorrect translations. This finding underscores the need for careful dataset creation and validation in developing robust QE systems. The relatively low divergence values in comparing distorted and original sentences across dataset sizes indicate that QE models are relatively consistent in evaluating bad translations, although slight variations exist when transitioning from large to small datasets. This consistency suggests that while data quality is crucial, QE models can still provide reliable assessments even in less controlled environments.

5.2.3 Domain-Specific Performance Variability

The study also analyzed domain-specific performance and found that models performed best in the 'Professional and Business Services' and 'Undefined Sector' domains, while the 'Financials' domain posed challenges, likely due to technical jargon and numerical information that is typically included in this domain. This variability highlights the need for either tailored approaches in QE or an improvement in training data, including several domains based on target users.

5.2.4 Study Limitations

Generalizability

It is important to note several limitations of this study. Firstly, the research focused solely on Dutch-English translation pairs, limiting the generalizability of the findings. The results may not be directly applicable to other language pairs, particularly those with different linguistic structures and translation challenges. Previous research has explored additional language pairs to assess the broader applicability of these findings, and this research attempted to add a new language pair to this particular field of study.

Furthermore, one might ask whether these insensitivities are significant enough for MT or if they occur commonly in real situations. However, research has shown that the errors examined in my report are frequent MT occurrences (Bentivogli et al., 2016; Lommel et al., 2014; Sennrich et al., 2017; Müller et al., 2020; Specia et al., 2019) and may have alarming consequences when left unidentified. Therefore, it is important that the QE models are effectively and accurately identifying and quantifying these errors.

Reproducibility

Reproducibility is another significant challenge in this study. The use of generative models like GPT-3.5-turbo and GPT-4-turbo introduces variability in the synthetic data creation process. As these models may produce different outputs with each run, ensuring the exact reproducibility of the datasets can be challenging. This inherent variability can affect the consistency of the results and poses a limitation in replicating the study.

Circularity Concerns

The issue of circularity is also a concern. A potential circularity exists in using language models to generate errors that are subsequently evaluated by QE models trained on similar data. This could lead to biased evaluations if the models are overly familiar with the types of distortions introduced. CometKiwi and TransQuest are known to be

trained on previous WMT annotated texts in multiple languages. Recent WMT tasks included critical error challenge tests, which pose similar error categories evident in this study. However, one main difference is that these models have no direct Dutch training data from WMT, as WMT has never experimented with this language. Therefore, it is unlikely the models will have seen Dutch source sentences. However, it could be likely that the distorted English target sentences could have been familiar to the QE models. Investigating alternative approaches to synthetic data creation that do not rely on the same models used for evaluation can help reduce potential biases and circularity in the research.

Dataset Complexity

Finally, the complexity of the dataset, including incomplete sentences, domain-specific jargon, and sentences consisting solely of numbers, may have impacted the performance of QE models. These grammatical complexities can confuse the proper performance of QE models and may need careful consideration in dataset preparation and model evaluation.

5.3 Conclusion and Future Work

5.3.1 Study Summary and Key Findings

This study aimed to assess the effectiveness of current Machine Translation Quality Estimation (MTQE) models in identifying and quantifying different types of translation errors within a Dutch-English dataset. To achieve this, advanced generative models, including GPT-3.5-turbo and GPT-4-turbo, were used to create synthetic datasets with deliberate error patterns. This approach provided a controlled environment to evaluate the robustness of MTQE models such as CometKiwi, TransQuest, LASER, and LaBSE.

The findings reveal significant variations in the performance of these models. CometKiwi appeared to be the most sensitive to various translation errors, especially hallucinations and named entity distortions, while TransQuest and LaBSE also showed commendable performance, though with less sensitivity. LASER demonstrated the least effectiveness in detecting the specific types of errors introduced in this study, suggesting that its intrinsic training limitations restrict its applicability to these error patterns. Surprisingly, LaBSE appeared to be the most sensitive to number error distortions, while the rest of the models remained relatively insensitive towards this error type. These results underscore the necessity of selecting appropriate QE models based on the specific types of errors most relevant to the context in which they will be applied.

5.3.2 Limitations of the Study

Despite these valuable insights, the study faced several limitations, including the focus on a single language pair (Dutch-English), reproducibility issues inherent in the use of generative models, and the complexity of the dataset. Addressing these limitations in future research will be crucial for advancing the field of MTQE. Exploring additional language pairs, developing more consistent synthetic data generation methods, and tailoring QE models to handle domain-specific challenges will enhance the reliability and applicability of these systems.

5.3.3 Future Research Directions

To improve the effectiveness and dependability of MTQE models, future work can explore various routes based on the findings and insights from this research. Following studies should contain a wider range of translation errors, encompassing those not only associated with accuracy but semantics, pragmatics, and stylistic variations. Assessing how QE models handle these complex error types can offer a more comprehensive evaluation of their stability and effectiveness in various linguistic circumstances.

Moreover, the results section indicated that while introducing distortions, the synthetic dataset created using GPT models also corrected grammar errors present in the original translations. This unintentional correction poses an interesting route for future research, namely, assessing whether these grammatical improvements alone could lead to higher QE scores. Investigating this aspect would offer insights into the impact of syntactic correctness on the overall performance of QE models.

The results section also revealed an intriguing deviation: some high-quality original translations received surprisingly low QE scores, while certain distorted sentences achieved higher scores than their original counterparts. This discrepancy highlights a critical need for a detailed analysis of the factors influencing QE model evaluations. Future studies should delve into why these inconsistencies occur, exploring potential biases or limitations within the QE models that may misjudge translation quality. A thorough understanding of these phenomena could lead to significant improvements in the accuracy and reliability of QE assessments.

Furthermore, considering the variability in QE model performance across different domains, as identified in this study, upcoming research should concentrate on training and fine-tuning QE models using domain-specific data. This can increase the models' sensitivity to specialized terminology and context-dependent errors, making them more effective in professional and technical translation settings. Adapting QE models to specific domains will heighten their accuracy and reliability in real-world applications.

Additionally, the unpredictability and inconsistency observed in the outputs of generative models such as GPT-3.5-turbo and GPT-4-turbo underscore the necessity for more advanced methods of synthetic data creation. Future work should focus on refining prompt engineering strategies to ensure that produced translations are consistent and error-free. Developing robust prompt templates and exploring alternative generative approaches can amplify the reliability and reproducibility of synthetic datasets, establishing a stronger foundation for QE model evaluation.

Appendix A

Appendix Title

Domain	Sentences	Words	Avg. Words/Sentence
Computer Software	3,746	47,833	12.77
Financials	1,230	18,680	15.19
Professional and Business Services	4,447	67,169	15.10
Undefined Sector	25,289	383,806	15.18

Table A.1: Domain Division of the Full Dataset

All the final prompts used for GPT-3.5-turbo and GPT-4-turbo:

- 'NamedEntities', 'In the following sentence change some named entities (people, organizations, locations) so they are unrecognizable but still existing named entities. Leave the rest of the sentence as is.'
- ''gibberish', 'In the following sentence add a nonsense word that has no meaning somewhere in the sentence. Leave the rest of the sentence as is. Ensure that the chosen replacement is random and not limited to a narrow set of words.'
- 'sentiment', 'In the following sentence, reverse the sentiment to convey the opposite meaning while keeping as much of the original sentence intact as possible. If sentiment reversal does not make sense or is not possible, do not just add negation, but leave the sentence unchanged.'
- 'hallucinations', 'In the following sentence, replace one important word with random, grammatically correct but semantically incorrect words that change the meaning of the sentence. Ensure that the chosen replacement is random and not limited to a narrow set of words. Do not add any additional words to the sentence and leave the rest of the sentence as is. '
- 'deletion', 'In the following sentence, delete a single critical word that changes the meaning of the sentence while ensuring the sentence remains grammatically correct. The rest of the sentence should be left unchanged. '
- 'negation', 'In the following sentence, reverse the meaning by using negation. either introduce or remove a negation while keeping the exact original words and structure of the sentence intact and consider using suffixes or prefixes like un-, im-, in-, il-, ir-, and dis- where appropriate. '

Bibliography

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. doi: 10.1162/tacl_a_00288. URL `https://doi.org/10.1162/tacl_a_00288`.

D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. *ArXiv*, 1409, 09 2014.

Y. Belinkov and J. Glass. Analysis methods in neural language processing: A survey. In *Transactions of the Association for Computational Linguistics*, 2019.

L. Bentivogli, A. Bisazza, M. Cettolo, and M. Federico. Quality and challenges of neural machine translation. *Journal of Machine Translation*, 30(3):233–249, 2016.

O. Bojar, R. Chatterjee, C. Federmann, Y. Graham, B. Haddow, M. Huck, A. Jimeno Yepes, P. Koehn, V. Logacheva, C. Monz, et al. Findings of the 2016 workshop on statistical machine translation. In *Proceedings of the First Conference on Machine Translation*, pages 131–198. Association for Computational Linguistics, 2016.

T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language models are few-shot learners. *NeurIPS 2020*, 2020.

C. Callison-Burch, M. Osborne, and P. Koehn. Re-evaluating the role of Bleu in machine translation research. In D. McCarthy and S. Wintner, editors, *11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 249–256, Trento, Italy, Apr. 2006. Association for Computational Linguistics. URL `https://aclanthology.org/E06-1032`.

K. Cho, B. van Merrienboer, Ç. Gülçehre, F. Bougares, H. Schwenk, and Y. Bengio. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *CoRR*, abs/1406.1078, 2014. URL `http://arxiv.org/abs/1406.1078`.

M. Denkowski and A. Lavie. Challenges in predicting machine translation utility for human post-editors. In *Proceedings of the 10th Conference of the Association for Machine Translation in the Americas: Research Papers*, San Diego, California, USA, Oct. 28-Nov. 1 2012. Association for Machine Translation in the Americas.

G. Doddington. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the second international conference on Human Language Technology Research*, pages 138–145. Morgan Kaufmann Publishers Inc., 2002.

F. Feng, Y. Yang, D. Cer, N. Arivazhagan, and W. Wang. Language-Agnostic BERT Sentence Embedding. 2020. URL https://arxiv.org/abs/2007.01852.

M. Fomicheva, R. Rei, P. Lertvittayakumjorn, J. G. de Souza, S. Eger, D. Kanojia, D. Alves, C. Orăsan, A. F. Martins, and L. Specia. Findings of the wmt 2022 shared task on quality estimation. In *Proceedings of the Seventh Conference on Machine Translation*. Association for Computational Linguistics, 2022.

E. Fonseca, K. Ram, J. Baldridge, and D. Johnson. Findings of the 2019 workshop on statistical machine translation. In *Proceedings of the Fourth Conference on Machine Translation*, volume 2, pages 1–61. Association for Computational Linguistics, 2019.

N. M. Guerreiro, D. Alves, J. Waldendorf, B. Haddow, A. Birch, P. Colombo, and A. F. T. Martins. Hallucinations in large multilingual translation models, 2023.

F. Guzmán, P.-J. Chen, M. Ott, J. Pino, G. Lample, P. Koehn, V. Chaudhary, and M. Ranzato. Flores: Evaluation datasets for low-resource machine translation: Nepali–english and sinhala–english. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, 2019.

W. J. Hutchins and H. Somers. An introduction to machine translation. *Academic Press, London*, 1992.

Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. Bang, A. Madotto, and P. Fung. Survey of hallucination in natural language generation. *CoRR*, abs/2202.03629, 2022. URL https://arxiv.org/abs/2202.03629.

N. Kalchbrenner and P. Blunsom. Recurrent continuous translation models. In D. Yarowsky, T. Baldwin, A. Korhonen, K. Livescu, and S. Bethard, editors, *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1700–1709, Seattle, Washington, USA, Oct. 2013. Association for Computational Linguistics. URL https://aclanthology.org/D13-1176.

F. Kepler, M. Vera, R. Rei, T. Luís, and L. Coheur. 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, 2019. doi: 10.18653/v1/P19-3020. URL https://aclanthology.org/P19-3020.

T. Kocmi and C. Federmann. Large language models are state-of-the-art evaluators of translation quality. *arXiv preprint arXiv:2302.14520*, 2023. URL https://arxiv.org/abs/2302.14520.

P. Koehn. *Statistical Machine Translation*. Cambridge University Press, 2010.

P. Koehn. *Neural machine translation*. Cambridge University Press, 2020.

S. Kullback and R. A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951.

A. Lavie and M. Denkowski. The meteor metric for automatic evaluation of machine translation. In *Machine translation summit*, volume 14, pages 86–95. Asia-Pacific Association for Machine Translation, 2009.

Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. Backpropagation applied to handwritten zip code recognition. *Neural Computation*, 1(4):541–551, 1989. doi: 10.1162/neco.1989.1.4.541.

Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. doi: 10.1109/5.726791.

E. M. M. J. Lehmann, Sabine. Machine translation evaluation: An analysis of methods. In *Proceedings of the First International Conference on Language Resources and Evaluation*, 1996.

A. Lommel and A. Burchardt. Error analysis of machine translation output. In *LREC 2014 Workshop*, 2014.

A. Lommel, A. Burchardt, and H. Uszkoreit. Multidimensional quality metrics (mqm): A framework for declaring and describing translation quality metrics. *Tradumàtica: tecnologies de la traducció*, 0:455–463, 12 2014. doi: 10.5565/rev/tradumatica.77.

M. Müller, A. Rios, and R. Sennrich. Quantifying the hallucination problem in neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3855–3865, 2020.

K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. Bleu: a method for automatic evaluation of machine translation. In P. Isabelle, E. Charniak, and D. Lin, editors, *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL `https://aclanthology.org/P02-1040`.

T. Ranasinghe, C. Orasan, and R. Mitkov. Transquest: Translation quality estimation with cross-lingual transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 139–149, 2020. doi: 10.18653/v1/2020.emnlp-main.11. URL `https://www.aclweb.org/anthology/2020.emnlp-main.11`.

R. Rei, C. Stewart, A. C. Farinha, and A. Søgaard. Comet: A neural framework for mt evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, 2020. doi: 10.18653/v1/2020.emnlp-main.213. URL `https://aclanthology.org/2020.emnlp-main.213`.

R. Rei, M. Treviso, N. M. Guerreiro, C. Zerva, A. C. Farinha, C. Maroti, J. G. C. de Souza, T. Glushkova, D. Alves, L. Coheur, A. Lavie, and A. F. T. Martins. CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task. In P. Koehn, L. Barrault, O. Bojar, F. Bougares, R. Chatterjee, M. R. Costa-jussà, C. Federmann, M. Fishel, A. Fraser, M. Freitag, Y. Graham, R. Grundkiewicz, P. Guzman, B. Haddow, M. Huck, A. Jimeno Yepes, T. Kocmi, A. Martins, M. Morishita, C. Monz, M. Nagata, T. Nakazawa, M. Negri, A. Névéol, M. Neves, M. Popel, M. Turchi, and M. Zampieri, editors, *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid), Dec. 2022. Association for Computational Linguistics. URL `https://aclanthology.org/2022.wmt-1.60`.

D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning representations by back-propagating errors. *Nature*, 323:533–536, 1986. URL `https://api.semanticscholar.org/CorpusID:205001834`.

R. Sennrich. Negation in neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 21–27, 2017.

R. Sennrich, B. Haddow, and A. Birch. Grammatical error correction as a foreign language teaching tool. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 721–731, 2017.

K. A. Sharou and L. Specia. A taxonomy and study of critical errors in machine translation. In H. Moniz, L. Macken, A. Rufener, L. Barrault, M. R. Costa-jussà, C. Declercq, M. Koponen, E. Kemp, S. Pilos, M. L. Forcada, C. Scarton, J. Van den Bogaert, J. Daems, A. Tezcan, B. Vanroy, and M. Fonteyne, editors, *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, pages 171–180, Ghent, Belgium, June 2022. European Association for Machine Translation. URL `https://aclanthology.org/2022.eamt-1.20`.

M. Snover, B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul. A study of translation edit rate with targeted human annotation. In *Proceedings of association for machine translation in the Americas*, volume 200, pages 223–231, 2006.

L. Specia and K. Shah. *Machine Translation Quality Estimation: Applications and Future Perspectives: From Principles to Practice*, pages 201–235. 07 2018. ISBN 978-3-319-91240-0. doi: 10.1007/978-3-319-91241-7_10.

L. Specia, F. Blain, V. Logacheva, R. Astudillo, and A. Martins. Findings of the wmt 2018 shared task on quality estimation. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 689–709. Association for Computational Linguistics, 2018.

L. Specia et al. Findings of the 2019 conference on machine translation (wmt19) shared task on quality estimation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 1–12, 2019.

I. Sutskever, O. Vinyals, and Q. V. Le. Sequence to sequence learning with neural networks. *CoRR*, abs/1409.3215, 2014. URL `http://arxiv.org/abs/1409.3215`.

J. Tiedemann and Y. Scherrer. Neural machine translation with extended context. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 82–91. Association for Computational Linguistics, 2018.

B. Turhan and K. Oflazer. Evaluating the evaluation metrics: A case study on machine translation of agglutinative languages. In *Proceedings of the Language Resources and Evaluation Conference*, pages 345–351, 2008.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. *CoRR*, abs/1706.03762, 2017. URL `http://arxiv.org/abs/1706.03762`.

D. Zeman. The acl anthology reference corpus: A reference dataset for bibliographic research in computational linguistics. *6th Language Resources and Evaluation Conference*, 2008.

B. Zoph, D. Yuret, J. May, and K. Knight. Transfer learning for low-resource neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2016.