Universität Zürich[UZH]

Master's thesis
presented to the Faculty of Arts and Social Sciences
of the University of Zurich
for the degree of
Master of Arts

Evaluation of Pre-trained Metrics and ChatGPT as Document-level Machine Translation Metrics

Author: Naomi Bleiker

Student ID: 17-648-981

Supervisor: Prof. Dr. Rico Sennrich

Department of Computational Linguistics

Submission Date: June 1, 2023

Abstract

Automatic evaluation metrics play an important role in the development and optimization of machine translation (MT) systems as they are the main method used for evaluating and comparing the quality of MT systems. Thus, it is crucial to be aware of potential limitations of automatic metrics. Most standard MT metrics evaluate translations at the sentence-level. However, because sentences can be ambiguous, document-level context needs to be considered to accurately assess the quality of MT output. Therefore, document-level metrics that understand context cues and are susceptible to discourse-level errors are required. This thesis examines and evaluates two approaches to automatic, document-level MT evaluation. The first group of experiments explores to what extent inter-sentential context is utilized in pre-trained metrics which are extended to the document-level as proposed by Vernikos et al. [2022]. The aim is to gain a better understanding of these metrics – specifically with regards to their understanding of context cues and their sensitivity to discourse-level phenomena. In the second part of this thesis, I assess another novel approach to document-level MT evaluation. First, I extend GEMBA [Kocmi and Federmann, 2023] – a GPT-based MT metric – to the document-level. Then, I conduct a number of experiments to explore the potential as well as limitations of GPT as a document-level MT evaluator.

Zusammenfassung

Automatische Metriken sind entscheidend für die Entwicklung und Optimierung von Systemen zur maschinellen Übersetzung (MÜ). Denn oft sind sie die einzige Methode zur Bewertung und zum Vergleich der Qualität von MÜ-Systemen. Daher ist es wichtig, potenzielle Schwächen automatischer Metriken zu erkennen. Die meisten MÜ-Metriken bewerten die von MÜ-Systemen generierten Texte auf der Satzebene. Da alleinstehende Sätze jedoch oft ambig sein können, muss der Kontext des gesamten Dokuments berücksichtigt werden, um die Qualität des Textes akkurat zu beurteilen. Daher werden Metriken benötigt, die satzübergreifenden Kontext mitberücksichtigen und dadurch auch Fehler auf der Diskursebene erkennen. In dieser Arbeit werden zwei Ansätze zur automatischen Bewertung von MÜ auf der Dokumentebene untersucht und evaluiert. Die ersten Experimente versuchen zu durchleuchten, inwieweit satzübergreifender Kontext in vortrainierten Metriken verwendet wird, die auf die Dokumentebene erweitert wurden, wie von Vernikos et al. [2022] vorgeschlagen. Das Ziel der Experimente ist, ein besseres Verständnis dieser Metriken zu erlangen - insbesondere bezüglich ihrer Fähigkeit, Hinweise aus satzübergreifendem Kontext zu nutzen und Fehler auf der Diskursebene zu erkennen. Im zweiten Teil der Arbeit untersuche ich einen weiteren Ansatz zur Bewertung von MÜ auf der Dokumentebene. Zuerst erweitere ich GEMBA [Kocmi et al., 2021] – eine auf GPT basierende Metrik – auf die Dokumentebene. Danach führe ich eine Reihe von Experimenten durch, um das Potenzial sowie die Grenzen von GPT in der Evaluation von MÜ auf der Dokumentebene zu analysieren.

Acknowledgement

First of all, I would like to thank my supervisor Rico Sennrich who supported me during this project and always provided helpful advice. His inputs and guidance have been greatly appreciated. I also want to thank Samuel Läubli for the interesting insights and for kindly providing data which I could use for this thesis. Many thanks to Jannis Vamvas for helping with the server setup in the beginning of the project. Finally, I want to thank my parents for their support throughout my studies.

Contents

List of Figures

List of Tables

List of Acronyms

BLEU	Bilingual Evaluation Understudy
chrF	character n-gram F-score
DA	Direct Assessments
GEMBA	GPT Estimation Metric Based Assessment
HTER	Human-mediated Translation Edit Rate
LLM	Large Language Model
LM	Language Model
MT	Machine Translation
MQM	Multidimensional Quality Metrics
NLG	Natural Language Generation
NLP	Natural Language Processing
NMT	Neural Machine Translation
SQM	Scalar Quality Metrics
TER	Translation Error Rate
WER	Word Error Rate
WMT	Conference on Machine Translation

1 Introduction

1.1 Motivation

Automatic metrics are the most common and often only method used for evaluating, reporting and comparing the quality of machine translation (MT) systems [Marie et al., 2021; Kocmi et al., 2021]. Metrics that are effective, reliable and accurate are thus essential for the development and optimization of MT systems. If commonly used MT metrics do have weaknesses, it is important to be aware of them as not to draw wrong conclusions about the quality of MT systems or optimize systems in a way that does not lead to real improvements because of wrong interpretations of metric scores.

Reliable MT metrics become even more important as MT systems improve. The output of different high quality MT systems can be difficult to discriminate and rank – especially if it is only evaluated at the sentence-level. Thus, for MT to be evaluated accurately, inter-sentential context needs to be considered [Läubli et al., 2018]. However, most of the currently used automatic metrics do not consider document-level context. Some metrics that were proposed, target a specific discourse-level phenomenon such as lexical cohesion [Semenov and Bojar, 2022]. While such metrics are useful for the evaluation of specific linguistic aspects, they do not provide information about the overall quality of the MT output. More recently, Vernikos et al. [2022] suggested a new approach where pre-trained MT metrics are extended to the document-level by embedding additional context sentences. While these and other pre-trained metrics have achieved high levels of correlations with human judgments and outperformed traditional, string-based metrics, there are some downsides to them that should not be neglected. One of the main issues lies in the fact that neural metrics are often not explainable – i.e. as opposed to string-based metrics it is not as clear why a certain score is given – which in turn makes them challenging to interpret. Therefore, one of the aims of this thesis is to gain a better understanding of the document-level MT metrics proposed by Vernikos et al. [2022], specifically with regards to the metrics' ability of understanding and using inter-sentential context cues.

In the second part of the thesis, a new approach to MT evaluation is explored. Kocmi and Federmann [2023] recently proposed GEMBA, a metric which uses GPT [Brown et al., 2020] for MT evaluation. GEMBA achieved high pair-wise accuracy and as the authors suggest, GPT might also be interesting for document-level MT evaluation. Because large language models (LLM) have shown capabilities in document-level MT [Wang et al., 2023b], they could potentially be useful for document-level MT evaluation. As with other metrics, if GPT is to be used effectively for this task, it is imperative to have a good understanding of its strengths and weaknesses so that scores can be interpreted accordingly. Thus, the second part of this thesis explores the potential and limitations of ChatGPT[1] as a document-level MT metric.

1.2 Research Questions

This Master's thesis aims to answer the following research questions:

1. To what extent do the document-level metrics proposed by Vernikos et al. [2022] utilize inter-sentential context, i.e. do they have a true understanding of context cues?
2. Are said document-level metrics sensitive to discourse-level errors and does the added inter-sentential context improve awareness of such errors?
3. Is GPT – specifically ChatGPT – a useful document-level MT evaluator?
    a) Do scores by ChatGPT correlate well with human judgments?
    b) Is ChatGPT sensitive to discourse-level translation errors and does it understand and/or leverage context cues?
    c) Can ChatGPT discriminate output of sentence-level and document-level MT systems?

1.3 Thesis Structure

In this first chapter, I explained the motivation behind this thesis and gave a brief overview of the research questions that shall be answered. Chapter 2 introduces the theoretical background by first discussing document-level MT, then briefly touching upon MT evaluation methods and finally giving an overview of automatic MT

[1] `https://chat.openai.com`.

metrics. I then discuss previous work on (document-level) MT metrics in chapter 3, more specifically, the metrics used and evaluated in this thesis are explained in detail. In chapter 4, methods for evaluating MT metrics are introduced as well as the data used in this thesis. Chapter 5 covers the first group of experiments that examines context usage and sensitivity to discourse-level errors of the metrics proposed by Vernikos et al. [2022] – results are also reported for each experiment[2]. The second group of experiments is covered in chapter 6 where the capabilities of ChatGPT as a document-level metric are explored. In chapter 7, the results are synthesized and limitations are discussed with respect to each of the three research questions. Finally, chapter 8 presents a brief conclusion.

[2]Code and data created for this thesis are available on GitHub: `https://github.com/naomiblkr/doc-mt-metrics-eval`

2 Background

In this chapter, I introduce the theoretical background of my thesis. First, I briefly discuss document-level MT. Then, I introduce MT metrics, touch on challenges in developing reliable metrics and discuss previous approaches to document-level metrics.

2.1 Document-level MT

Over recent years, neural MT (NMT) [Kalchbrenner and Blunsom, 2013; Bahdanau et al., 2015; Vaswani et al., 2017] has made considerable progress and achieved high levels of performance. However, most MT systems still translate sentences separately without considering document-level context. Discourse-level phenomena such as lexical cohesion and disambiguation, coreference and coherence are thus neglected which leads to lower quality translations at the document-level [Bawden et al., 2018; Läubli et al., 2018]. As opposed to sentence-level NMT, document-level systems typically extend the translation unit from a single sentence to multiple sentences or in few cases the whole document [Maruf et al., 2019]. Document-level systems are thus capable of generating more consistent and coherent output and can better disambiguate words by using inter-sentential context [Bawden et al., 2018].

As the focus of this thesis are document-level MT metrics, I will not discuss the architecture of document-level MT systems. However, I will cover context usage because this is also relevant to the evaluation of neural metrics.

2.1.1 Context Usage in Document-level MT

In contrast to sentence-level MT, document-level MT takes into account inter-sentential context from both the source and target side. How many context sentences are incorporated and whether only preceding or also succeeding context is used varies based on the approach and can also depend on the language pair and domain. However, findings by Bawden et al. [2018] and Fernandes et al. [2021]

suggest that target-side context is more relevant than source-side context and that using two context sentences usually leads to the best performances. Additionally, Fernandes et al. [2021] show that most context is used from one context sentence, i.e. using just one context sentence can already lead to considerable improvements. This goes in hand with results from Kim et al. [2019] which suggest that incorporating very long context is not helpful in MT. Despite these findings, improvements of context-aware MT systems are often due to non-interpretable context usages, i.e. from non-linguistic factors [Kim et al., 2019]. It thus remains somewhat unclear how context is actually used as improved performances of document-level systems are largely due to factors such as more noise being generated by adding context which then acts as a regularizer.

2.2 MT Evaluation

MT evaluation is a difficult task as there are numerous possible correct translations of the same source text. Even for humans this can be challenging because judging the translation quality is not only somewhat subjective but also hard to quantify [Freitag et al., 2021a]. Moreover, the definition of what is considered good translation quality can also depend on the use case and target audience of the translation. For example, in a legal or medical context, accuracy is arguably the most important aspect, however, in other settings a more loose translation might be acceptable or even welcomed. Thus, it is difficult to adequately judge the translation quality or the quality of an MT system without knowing its use case.

There are four main approaches to MT evaluation: human evaluation, automatic evaluation metrics, task-based evaluation, and targeted evaluation of linguistic phenomena [Koehn, 2009]. While human evaluation - if carried out carefully - is arguably the most accurate, it is often too time consuming and can also be somewhat subjective. If accurate evaluations are to be elicited from human raters there are also numerous factors to consider. This will be discussed in more detail in chapter 4.1. In contrast to human judgments, automatic MT metrics allow for efficient and low cost comparison of different MT systems. They also have the advantage that they are tunable, i.e. systems can be directly optimized towards the metrics [Koehn, 2009]. However, there are many open issues and challenges which will be explained in the coming chapters. While human judgments and automatic metrics aim to evaluate the general translation quality, task-based evaluation is concerned with the MT quality given a specific task or application. For example, if the MT system is used to support human translators, this could mean measuring the required post-editing

effort [Koehn, 2009]. Finally, methods such as contrastive evaluation [Sennrich, 2017] or evaluation with adversial datasets target the evaluation of specific linguistic phenomena, i.e. they examine how well MT systems perform in translating them. Therefore, targeted evaluation can provide more detailed information about specific strengths and weaknesses of MT systems than a score assigned by an automatic metric.

2.3 Automatic MT Metrics

In order to develop and improve MT systems, there is a need for efficient and reliable evaluation methods. While human evaluations are still considered gold standard, they are too costly and not useful when testing several versions of an MT system in order to optimize it or when comparing multiple MT systems. As a consequence, automatic metrics are often the only method used for reporting MT quality Marie et al. [2021]; Kocmi et al. [2021]. Metric choice can therefore have a significant impact on MT developments [Kocmi et al., 2021]. Thus, numerous automatic metrics have been proposed over the years and it is still an ongoing research area. However, finding a reliable automatic metric has been proven to be a challenging task.

There are several approaches to MT metrics. Mainly, we can differentiate between traditional, string-based (or surface-level) metrics and learned or neural metrics. In their taxonomy of MT metrics Lee et al. [2023] classify the metrics into the following main categories: lexical metrics, embedding-based metrics, and supervised-metrics (see Figure 1).

We can further discriminate between reference-based and reference-free MT metrics. Reference-based metrics make use of one or multiple human reference translation(s) of the same source text to compare the hypothesis (i.e. MT output) against it. Especially string-based metrics usually compute the similarity between the hypothesis and a human reference translation. Some neural metrics additionally leverage the source text (e.g. COMET [Rei et al., 2020]). By contrast, reference-free metrics compare the MT output to the source text without using a reference translation. Thus, reference-free metrics are also referred to as quality estimation metrics because there is no gold standard that the hypothesis can be compared to, i.e. they predict the quality of the MT output.

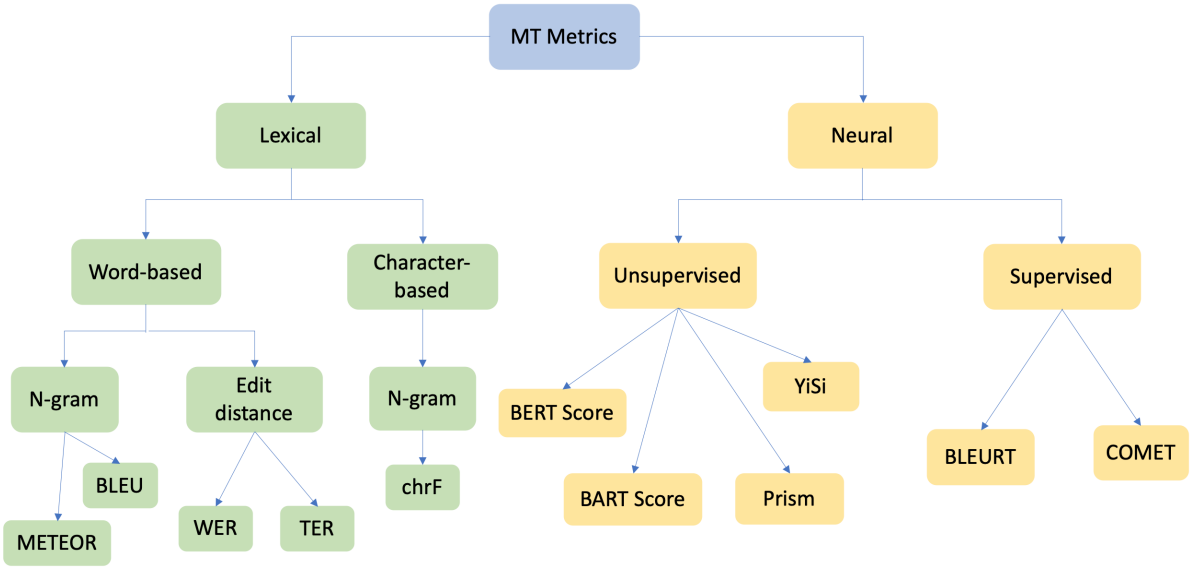

Figure 1: Non-exhaustive taxonomy of MT metrics. Adapted from Lee et al. [2023]

2.3.1 Challenges of Developing Reliable Metrics

As mentioned in chapter 2.2, there are many acceptable translations of the same source text. This lack of a gold standard is one of the biggest challenges in the development of MT metrics. While human references translation can be (and usually are) used as a gold standard, an MT output or even another reference translation can differ from the human reference but still be correct. Therefore, it has been suggested to make use of multiple references [Qin and Specia, 2015]. However, results by Freitag et al. [2020] indicate that using multiple references does not improve correlations with human judgements.

Moreover, MT metrics are required to not only be reliable, but also accurate, meaningful, consistent and low in cost [Koehn, 2009]. In order to allow for optimization of MT systems, the metric scores should be explainable, i.e. it should be clear why a metric assigns a certain score to a translation. Metrics also have to be sensitive to different types of errors and correctly weight the error severity. MT systems can make errors in multiple categories and on various levels from sentence-level aspects such as grammar mistakes, terminology / named entity mistranslations, incorrect syntax or punctuation errors to discourse-level errors such as global incoherence, lexical inconsistency and more. Chapter 4.1.1 will give a more comprehensive overview of translation error categories. An ideal metric would correctly weight mistakes of various categories. For example, an accuracy error – such as the mistranslation of a

named entity or numbers – heavily impacts the translation quality and should thus arguably be given more weight than a less severe mistake.

As Ma et al. [2019] point out, one of the main shortcomings of automatic metrics is the evaluation of high quality MT systems. Their findings show that system-level correlations of metrics' scores with human judgments are worse when only a few top-performing systems are considered. This can be a problem because in many cases, metrics are used to compare similar versions of MT systems and/or MT systems with similar performances. With further improvements of MT systems – and with the introduction of document-level MT systems – we thus also require accurate metrics to correctly evaluate these high quality systems. For a metric to be useful on a large scale, it should ideally also be language-independent, meaning it can be applied to various language-pairs.

2.3.2 String-based Metrics

Traditional, string-based metrics typically measure the surface-level distance between the MT hypothesis and a human reference translation. This comparison is either done on a word or character level. One of the older but still most widely used metrics is BLEU [Papineni et al., 2002], which is based on n-gram overlaps. BLEU counts the number of word n-gram overlaps between the reference and hypothesis, then calculates a precision score and is weighted by a brevity penalty. Banerjee and Lavie [2005] later introduced METEOR which addresses some of the weaknesses of BLEU and allows for more flexibility so that similar tokens with only morphological variations are also matched. Another popular surface-level metric is chrF [Popović, 2015] which is very similar to BLEU but uses character n-grams instead of word n-grams to compare the hypothesis with the reference. Metrics such as TER [Snover et al., 2006] (translation edit rate) take a different approach and measure the post-edit effort, i.e. the edit distance between the MT hypothesis and the reference translation.

BLEU and other surface-level metrics have been shown to correlate poorly with human judgments [Freitag et al., 2022]. Traditional metrics also assign low scores to human translations (compared to MT output) which is likely due to the fact that human translations contain more variability than MT output leading to fewer n-gram overlaps [Koehn, 2009]. While the many shortcomings of traditional metrics such as BLEU have been researched extensively [Mathur et al., 2020a], it still remains the most commonly used metric to evaluate and rank MT systems – often being the only evaluation method reported in publications [Marie et al., 2021]. A possible

reason for this is that because BLEU scores are reported so frequently, it allows for comparison with previous work [Marie et al., 2021]. An advantage of BLEU (and other string-based metrics) is that it can easily be applied to many language pairs. However, using low-quality metrics as the standard is an issue because it can lead to MT systems being optimized in ways that do not actually improve their quality [Kocmi et al., 2021].

2.3.3 Neural Metrics

During recent years, a number of neural MT metrics have been proposed. As opposed to traditional metrics, they do not utilize surface-level similarity but leverage neural networks to produce MT quality scores. We can distinguish between two main approaches. On the one hand, there are unsupervised metrics which exploit embeddings to compute the cosine similarity, usually, between the hypothesis and the reference translation (e.g. BERTScore [Zhang et al., 2020], YiSi [Lo, 2019]). On the other hand, supervised metrics are trained on human judgements of a translation (e.g. COMET [Rei et al., 2020], BLEURT [Sellam et al., 2020]). Some neural metrics also leverage the source text. Many of these metrics are pre-trained on LLMs.

Traditional metrics have recently been outperformed by pre-trained metrics [Mathur et al., 2020b; Freitag et al., 2022]. Nevertheless, there are also some downsides to neural metrics. Firstly, as with any neural model, it is often difficult to understand and explain the performance of neural MT metrics (i.e. knowing what a model actually "does" as it processes data). Thus, while neural metrics can achieve improved correlations with human judgments, we often do not truly understand why a translation is scored in a certain way. Secondly, due to this lack of explainability and because there often is no clear scoring range, scores of neural metrics can be difficult to interpret for humans. Thirdly, experiments have revealed that neural metrics might not be susceptible enough to certain mistakes or phenomena and misjudge the severity of errors [Freitag et al., 2021b]. For example, COMET – which is one of the best performing metrics – has been shown to not be sensitive to named entity mistranslations and number differences [Amrhein and Sennrich, 2022].

2.3.4 Document-level Metrics

A disadvantage of both string-based and neural metrics is that most of them evaluate translations on the sentence-level. However, a single sentence out of context could

be regarded as correct and thus rated highly – even by human raters – although it is not the correct translation when considered in context. For example, it might contain mistranslations of ambiguous words or incorrect coreferences. Overall fluency, coherence and cohesion between sentences or paragraphs, lexical consistency, formality level (i.e. register) and style are other aspects that can mainly be evaluated at the document-level. Thus, for accurate evaluation inter-sentential context needs to be considered [Läubli et al., 2018].

Especially with improvements in MT and the development of document-level MT systems, there is an increasing need to then also evaluate these systems at the document-level. When comparing high quality MT systems, evaluation at the sentence-level is often not sufficient because errors which cannot be spotted without considering context sentences are often decisive for discriminating different MT hypotheses [Läubli et al., 2018]. In order to improve correlations with human evaluation scores and to accurately assess MT, metrics therefore need to be expanded to the document-level.

So far context-aware MT evaluation has mainly been carried out by utilizing testsets (challenge sets) that target specific discourse-level phenomena [Bawden et al., 2018; Isabelle et al., 2017; Voita et al., 2019]. However, only few automatic document-level metrics have been developed as of yet. For example, metrics were proposed that target a specific discourse-level phenomenon such as lexical consistency or coherence [Gong et al., 2015; Semenov and Bojar, 2022].

A different approach was recently proposed by Jiang et al. [2022]. The document-level metric BlonDe incorporates a number of discourse phenomena in order to capture and measure aspects of translation quality beyond the sentence-level. BlonDe calculates the similarity of text spans from the following categories: named entity, tense, pronouns and discourse markers. It additionally accounts for sentence-level accuracy by adding n-grams to the set of categories.

By contrast, Vernikos et al. [2022] suggest a method for converting three pre-trained metrics – i.e. COMET, BERTScore and Prism [Thompson and Post, 2020] – into document-level metrics by incorporating additional context sentences. They do this by concatenating each sentence with two preceding sentences of context which is then used by the model when embedding each reference and hypothesis (and source) sentence. Before computing the scores, the context sentences are discarded as to only score the current sentence. The idea is to improve the embedding representation while only computing scores for one sentence at a time. For the most part, this led to improved correlations of metric scores with human judgments compared to the original versions of the metrics. The authors also evaluated the reference-free version

of sentence- and document-level COMET on a contrastive testset (which targets pronoun translation), reporting improved accuracy for Doc-COMET. The accuracy in this case is the percentage where the metric scored the correct translation higher than the incorrect translation. Vernikos et al. [2022] mention that the contrastive evaluation was only carried out for the reference-free version because access to a reference would allow metrics to easily disambiguate words without needing context. The metrics will be introduced in more detail in the following chapter.

3 Previous Work

In this chapter, I discuss the metrics that are examined in this thesis in more detail. I first introduce each of the three metrics that Vernikos et al. [2022] converted to the document-level and I explain their approach. In chapter 3.2.1, I then introduce GEMBA [Kocmi and Federmann, 2023] – a newly proposed metric which prompts GPT to evaluate translation quality. Finally, I briefly discuss related work which utilizes GPT for document-level MT.

3.1 Extending Pre-trained Metrics to the Document-level

As briefly introduced in chapter 2.3.4, Vernikos et al. [2022] proposed converting the pre-trained metrics COMET, BERTScore and Prism into document-level metrics by incorporating context sentences. In the following chapters, each of the three metrics is explained in more detail.

3.1.1 Doc-COMET

COMET [Rei et al., 2020] is a neural machine translation metric that has achieved high levels of correlation with human judgments. It is a supervised metric that leverages a multilingual pre-trained language model (LM). There are several COMET models which are all trained on different types of human judgments and generate prediction estimates of said judgments. In contrast to most other MT metrics, COMET does not solely rely on the human reference but also makes use of the source text. The models are trained on the following types of human judgments: Direct Assessments (DA) [Graham et al., 2013], Human-mediated Translation Edit Rate (HTER) [Snover et al., 2006] and Multidimensional Quality Metrics (MQM) [Lommel et al., 2014; Freitag et al., 2021a]. There are also variations in the models' architectures. However, most models use a regression/estimation architecture as opposed to the ranking architecture. As document-level COMET (Doc-COMET)

[Vernikos et al., 2022] utilizes the model MQM-2021 [Rei et al., 2021], I will focus on the method used for this model. More information on the different types of human judgments is provided in chapter 4.1.

COMET MQM-2021 is an estimator model and is built on top of XLM-RoBERTa-large [Conneau et al., 2020]. The COMET model is pre-trained on DA evaluations from WMT 2015 to 2020 (from 2015-2019 for English to German) and fine-tuned on z-normalized MQM scores [Freitag et al., 2021a]. COMET encodes the source, hypothesis and reference using a multilingual encoder. Then, a sentence embedding is created which is the average of the sentence's output token embeddings. Finally, the encoded sentence representations are further combined and concatenated and are then fed to a regressor that predicts a score for each hypothesis sentence.

Vernikos et al. [2022] extend COMET to the document-level by integrating source context c_s and reference context c_r in the encoder. They concatenate each source s, hypothesis h and reference r with two previous context sentences $\langle c_s; s, c_r; h, c_r; r \rangle$. The output embeddings of the tokens of each sentence are then averaged before being passed to the regressor to obtain sentence representations. In this way, the authors aim to achieve improved embedding representations. As the underlying LM is inherently context-aware, Vernikos et al. [2022] did not make any changes. However, the COMET model is fine-tuned on sentence-level human judgments – which the authors argue should not impact the model's capabilities of handling longer text sequences because the amount of fine-tuning is limited.

It should be mentioned that for the hypothesis sentences they do not add hypothesis context but reference context. The reasoning behind this decision is that adding the reference context to the hypothesis achieved higher correlations with human judgments than adding the hypothesis context. The authors hypothesize that this is due to possible propagation of errors if context from the hypothesis is used. I.e. if there is a mistake in a hypothesis sentence, this could possibly not only lead to said sentence receiving a worse score but also the next two sentences (where the hypothesis sentence with an error is added as context). However, using the reference context seems slightly counterintuitive as the reason for using context is to evaluate the hypothesis as a whole document. If the hypothesis differs from the reference – which is usually the case – then the hypothesis sentence is potentially a worse translation in the reference context than in its original context. Thus, this also raises the question, to what extent the context is used and what is learned from the context.

3.1.2 Doc-BERTScore

BERTScore [Zhang et al., 2020] is a general text generation metric. It makes use of a reference text and computes a similarity score between the reference and candidate. Thus, it is also an effective reference-based MT metric. BERTScore leverages contextual embeddings to represent the tokens of the reference and hypothesis sentence. It then computes the cosine similarity between the reference and hypothesis representations. In a next step, precision, recall and F1 scores of the hypothesis compared to the reference are calculated.

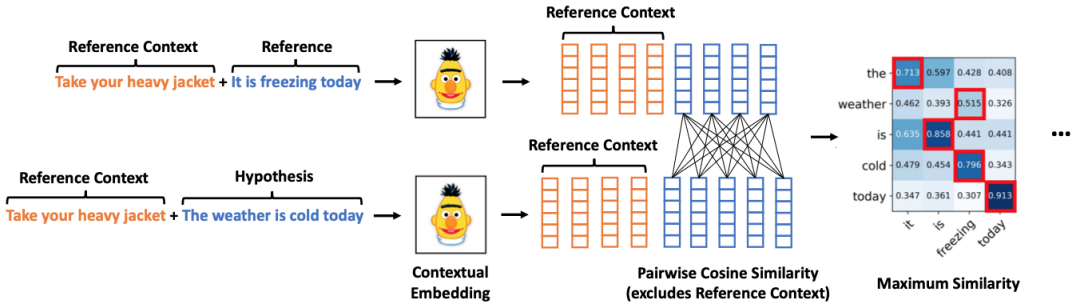

Figure 2: Document-level BERTScore. Image from Vernikos et al. [2022]

In order to extend BERTscore to the document-level, Vernikos et al. [2022] concatenate two preceding context sentences with the reference and hypothesis $\langle c_r; h, c_r; r \rangle$. As with Doc-COMET, reference context is added to both the hypothesis and reference sentences. As Figure 2 shows, the context is added before embedding the reference and hypothesis to improve their contextual representation. However, the additional context is discarded when computing the cosine similarity so that only the current sentence is scored. For document-level BERTScore (Doc-BERTScore), the authors use the multilingual BERT-base model [Devlin et al., 2019] for English to German (other models are used for other language pairs). BERT is inherently document-level in the way that the model is trained on up to 512 tokens at a time, which corresponds to roughly 30 sentences, assuming an average sentence length of 17 tokens.

3.1.3 Doc-Prism

Thompson and Post [2020] took a different approach and proposed using a sequence-to-sequence paraphraser to evaluate MT hypotheses. This approach has the advantage that no human judgments are required for training which are often not available

for lower-resource language pairs or for certain domains. Essentially, Prism (Probability is the metric) measures how well a hypothesis paraphrases a human reference translation. This is done by querying the paraphrasing model – which implicitly represents the entire set of potential paraphrases of a sentence – with the hypothesis. More concretely, the reference is fed to the encoder and the MT output is force-decoded. The resulting score – i.e. the average token-level log probability – is then used to determine how well the hypothesis paraphrases the reference. In order to equally penalize the presence of information in one sentence which is dropped in the other sentence, the reference is also scored on the hypothesis, i.e. the hypothesis is fed to the encoder and the reference is force-decoded. Let hyp denote the MT hypothesis and ref the human reference, scoring in both directions results in the following metric:

$$\text{Prism} = \frac{1}{2}\text{H}(\text{hyp}|\text{ref}) + \frac{1}{2}\text{H}(\text{ref}|\text{hyp})$$

Thompson and Post [2020] propose using a multilingual MT system as a paraphraser which allows for language-agnostic semantic representation. Paraphrasing is then treated as an intralingual translation task (e.g. German to German). However, the model released by Thompson and Post [2020] is trained on the sentence-level – but to extend Prism to the document-level, a document-level model is required as a paraphraser.

As document-level data is not available for as many languages as sentence-level data, Vernikos et al. [2022] use a multilingual encoder-decoder LM as a paraphraser: mBART-50 [Tang et al., 2020] which is trained on document segments of up to 512 tokens in 50 languages. As with the previous two metrics, each hypothesis and reference sentence is concatenated with two preceding context sentences and the reference context is added to both the reference and hypothesis $\langle c_r; h, c_r; r \rangle$.

3.2 GPT as an Evaluator of MT Quality

Recent research has applied GPT [Brown et al., 2020] to various NLP tasks – among others translation as well as MT evaluation. This chapter first introduces GEMBA [Kocmi and Federmann, 2023], a GPT-based MT metric, then discusses other related work that applied ChatGPT[1] to natural language generation (NLG) evaluation and to document-level MT.

[1] https://chat.openai.com

3.2.1 GEMBA

A novel approach to automatic MT evaluation was recently suggested by Kocmi and Federmann [2023]. They propose GEMBA (GPT Estimation Metric Based Assessment) which makes use of GPT for the task of translation quality evaluation. Leveraging the fact that large language models (LLMs) such as GPT support multilingual Q&A, they design several prompts to score translation quality. GPT has recently been shown to achieve high performances in MT in high-resource settings [Hendy et al., 2023], thus the authors examined whether GPT can also be used as an MT metric.

GEMBA scores one segment per prompt and averages all segment-level scores to obtain the system-level score. Although recent versions of GPT can use large context windows, the approach by Kocmi and Federmann [2023] is not inherently document-level as it does not incorporate inter-sentential context but instead scores each sentence separately. However, as the authors suggest, the approach has potential to be extended to the document-level which will be explored in chapter 6.

Kocmi and Federmann [2023] designed four prompt types, two of which are scoring and two are classification tasks. The two prompts for the scoring task are inspired by the design of DAs (GEMBA-DA) and scalar quality metrics (GEMBA-SQM, see Figure 3). In the scoring task, GPT is prompted to score a translation on a scale between 0 and 100. Thus, the expected score lies in the range [0,100]. In the classification tasks, on the other hand, GPT is asked to rank the translation. The first variant is a one-to-five stars ranking task (GEMBA-stars) while the second prompts GPT to assign one of five discrete quality class labels (GEMBA-classes). The score for the classification task therefore ranges between [1,5]. For each prompt type, there are two variations, a reference-based and a reference-free version. The source text is also provided in all prompts.

Kocmi and Federmann [2023] test seven GPT models from GPT 2 up to ChatGPT (GPT 3.5 model). They use the MQM 2022 testset from the WMT22 Metrics shared task [Freitag et al., 2022] which contains human judgments for English into German, English into Russian and Chinese into English. They evaluate the system-level performance of GEMBA by computing pair-wise accuracy (see chapter 4.2). GEMBA-DA achieved a higher system-level accuracy than all other best-performing metrics from the WMT 2022 Metrics shared task. The GPT models that achieved the best performances were Davinci-002, Davinci-003, and ChatGPT – all of which are GPT 3.5 or larger models. Davinci-002 and Davinci-003 are both text completion models while ChatGPT is a conversational model.

```
Score the following translation from {source_lang} to
{target_lang} with respect to the human reference on a continuous
scale from 0 to 100 that starts with "No meaning preserved", goes
through "Some meaning preserved", then "Most meaning preserved and
few grammar mistakes", up to "Perfect meaning and grammar".

{source_lang} source: "{source_seg}"
{target_lang} human reference: "{reference_seg}"
{target_lang} translation: "{target_seg}"
Score (0-100):
```

Figure 3: Example prompt based on scalar quality metrics. For the reference-free version, the text in bold is dropped. Image from Kocmi and Federmann [2023]

Kocmi and Federmann [2023] also calculate the distribution of all segment-level scores for the model Davinci-003 to determine whether the scores are actually in the desired range. The results show that GEMBA-DA scores are only multiples of five. Moreover, the scores 80, 90, or 95 make up over three-quarters of all scores. The authors point out that this is likely due to the underlying MT systems being of high quality and – as the pair-wise accuracy scores show – the metric manages to discriminate the systems from each other. Nevertheless, this does raise the question of how well very similar translations can actually be differentiated and whether more fine-grained scores are required. The distribution also shows that SQM-based prompts produce more fine-grained scores. Whether this leads to a better or worse performance is not entirely clear. With Davinci-003, GEMBA-DA achieved the highest performance, however, with Davinic-002, GEMBA-DA and GEMBA-SQM obtain the same accuracy score while with ChatGPT, GEMBA-SQM outperformed GEMBA-DA. The distribution of scores could be relevant when extending GEMBA to the document-level, as an accurate document-level metric might require more fine-grained scores that can reflect minor differences between translations of different systems.

3.2.2 ChatGPT as an NLG Evaluator

Recently, a similar application of ChatGPT was proposed by Wang et al. [2023a] who explore the capabilities of ChatGPT as an automatic NLG metric. They prompt ChatGPT to score the generated text with one to five stars. The authors carried out a meta-evaluation of ChatGPT as an NLG metric for the following three tasks: text summarization, story generation and data-to-text. For each of the tasks, Wang et al.

[2023a] designed four prompts each of which evaluates one specific aspect of a text, i.e. coherence, relevance, consistency and fluency. They report sample-level and dataset-level correlations with human judgments (Spearman correlation, Pearson's correlation and Kendall's tau). Their results show mostly improved correlations for the summarization and story generation evaluation tasks compared to other widely used metrics.

3.2.3 Document-level MT with ChatGPT

Wang et al. [2023b] test the capabilities of GPT models (ChatGPT and GPT-4) in document-level MT. Therefore, their findings might also be relevant for the application of ChatGPT as a document-level MT metric. The motivation behind their approach is the fact that LLMs have shown the ability of generating coherent and consistent text. The authors thus examine whether GPT models are capable of handling discourse phenomena in MT such as named entity consistency and pronoun translation.

They prompt ChatGPT to translate multiple sentences at a time – as opposed to just prompting it to translate one sentence per conversational turn – with the aim that this improves discourse awareness and thus document-level translation quality. The authors test several prompt variations, in one of which they insert a boundary tag between the sentences. To evaluate the translation quality, Wang et al. [2023b] use a range of evaluation metrics - including metrics that target specific discourse phenomena – as well as human evaluation.

Their findings suggest that combining multiple sentences in one prompt can marginally improve discourse awareness and thus overall translation quality. According to human evaluations, ChatGPT and GPT-4 outperform commercial MT systems. As the authors mention, results using automatic MT evaluation metrics are less clear, indicating that they might be unsuitable for document-level MT evaluation. Finally, Wang et al. [2023b] performed targeted evaluation of specific discourse phenomena (deixis, lexical consistency and ellipsis) by prompting GPT to predict which of two contrasting translations is the correct one. The results show that ChatGPT performs worse than document-level MT systems, suggesting that ChatGPT might not be as capable of utilizing context to resolve discourse-level ambiguities and inconsistencies. However, the authors report significantly better performance for GPT-4 over the contrastive testsets.

4 Methods and Data for the Evaluation of MT Metrics

In this chapter, I first give a general overview of methods for evaluating MT metrics. I then introduce different methods of collecting human judgments, explain how correlations are calculated and finally, I briefly discuss challenge sets.

The standard approach to evaluating MT metrics is to compute correlations of the metrics' scores with human quality judgments. Human evaluations of MT output are thus considered the gold standard to which metric scores are compared. However, correlations are not always reliable and their stability depends on aspects such as the sample size (i.e. number of MT systems) [Mathur et al., 2020a]. The reliability of correlations also depends on the quality of the underlying human judgments [Freitag et al., 2021b,a]. Additionally, while correlations are useful for comparing different metrics to each other, they do not provide any further insight into specific strengths and weaknesses of metrics. Therefore, other approaches attempt to gain a better understanding of MT metrics. This is especially the case for neural metrics, which while achieving high correlations with human judgments are often not well understood, i.e. their performance is often not explainable. To this end, challenge sets can be useful to measure the performance of a metric on specific linguistic phenomena [Freitag et al., 2022]. They can thus help to determine whether metrics are sensitive to specific translation errors.

Further experiments can provide additional insight into a metrics performance. As this thesis focuses on document-level metrics, it is especially of interest whether these metrics are capable of accurately discriminating output from sentence-level and document-level MT systems. To test this, metric scores for the output of a sentence-level version and document-level version can be compared – the underlying assumption being that the output of the document-level system is better and should thus receive a higher score.

4.1 Collection of Human Judgments

As briefly mentioned, when computing correlations of metric scores with human judgments, high quality human evaluation is crucial [Freitag et al., 2021b]. Numerous methods for collecting human judgments of MT have been proposed over the years. Judging MT quality is a challenging task not only for automatic metrics but also for humans. As the performance of MT systems improves, discriminating high quality systems can be difficult even for humans, especially when only judging translations sentence by sentence without considering context [Läubli et al., 2018]. Aspects such as access to the source text while rating the translations and whether the ratings are collected from professional translators or crowd workers also impact the quality of human judgments [Freitag et al., 2021a]. Finding the right approach is thus crucial as inadequate evaluation methods can lead to misleading conclusions about the quality of MT and MT metrics [Läubli et al., 2018].

We can mainly differentiate between four methods for collecting human judgments of MT quality. The first approach is to simply assign scalar ratings to the translation [Koehn and Monz, 2006; Graham et al., 2013] – the score is usually assigned per sentence. The second method can be described as a ranking task, i.e. raters are asked to rank different translations of the same source text [Vilar et al., 2007]. The third method measures editing rates, i.e. how much an MT has to be edited to be correct [Snover et al., 2006]. The last approach derives the score from annotating mistakes [Lommel et al., 2014; Freitag et al., 2021a]. Human judgments are usually either collected from crowd workers, researchers or professional translators. While annotations from crowd workers are usually cheaper, judgments collected from professional translators have been shown to be of higher quality [Freitag et al., 2021a].

The methods that were most commonly used in previous years are Direct Assessments (DA) as proposed by Graham et al. [2013] and Multidimensional Quality Metrics (MQM) [Lommel et al., 2014]. In DAs, humans rate the MT output from each system with an absolute score or label (on a Likert-type scale). For example, the raters are asked to rate the fluency of a sentence on a scale of 5. The system-level rating is then calculated by averaging the sentence-level ratings. By contrast, in MQM-based evaluations, the score is derived from errors which raters annotate sentence by sentence.

4.1.1 MQM Framework and Ontology

I will now discuss MQM in more detail, as MQM-annotated data is used in this thesis for computing correlations with human judgments. MQM is a framework that was developed by Lommel et al. [2014] to better describe and declare translation quality metrics. Because the evaluation of translation quality can be a subjective task and is often dependent on the use case, the MQM framework provides guidelines for evaluating both human and machine translations. Specifically, annotation guidelines and error categories were created that aim to make translation quality judgments more transparent and comprehensible. The current version (MQM 2.0[1]) contains the following high-level error categories: terminology, accuracy, linguistic conventions, style, locale conventions, audience appropriateness and design and markup. These are then further divided into subcategories.

Freitag et al. [2021a] adapted the MQM framework for their large-scale human evaluation corpus and proposed a standard MQM scoring scheme. An overview of the categories can be found in table 1. Errors were additionally annotated according to severity levels, consisting of major, minor and neutral levels. The error annotation was carried out by professional translators. Unlike was the case in previous collections of human judgments, the annotators had access to the full document context when annotating each sentence and were instructed to consider said context. However, it should be mentioned that the errors were annotated only at the sentence-level, i.e. although context was considered, there is no true document-level or overall quality score (besides the overall error count). How well document-level quality is reflected in these scores is thus not entirely clear.

Error annotations allow for predefined weighting of certain categories, i.e. the raters are not asked to assign a score that they synthesize from all the information [Freitag et al., 2021a]. The idea behind this approach is that this should lead to less variation in how different annotators assign scores – although, there can still be differences in the error categories the annotators assign.

Finally, I will briefly discuss the Scalar Quality Metric (SQM) [Freitag et al., 2021a] which was used as inspiration for some of the prompts in GEMBA. In this method, scalar ratings are collected at the segment-level. For each segment, raters are asked to select a rating from 0 through 6. The raters have access to all source and target text context while judging each segment. In contrast to DAs, a more narrow, non-continuous scale is used. SQM-annotations achieved high correlation with MQM on the system-level.

[1] `https://themqm.org/error-types-2/typology/`

Error Category	Subcategory
Accuracy	Addition
	Omission
	Mistranslation
	Untranslated text
Fluency	Punctuation
	Spelling
	Grammar
	Register
	Inconsistency
	Character encoding
Terminology	Inappropriate for context
	Inconsistent use
Style	Awkward
Locale convention	Address format
	Currency format
	Date format
	Name format
	Telephone format
	Time format
Other	
Source error	
Non-translation	

Table 1: MQM error categories (adapted from Freitag et al. [2021a])

4.2 Correlation Between Metric Scores and Human Judgments

Correlations between metric scores and human judgment scores (i.e. gold standard) of a translation are computed either on the sentence-, document-, or system-level. On the system-level, Pearson's r is most commonly used [Mathur et al., 2020a]. Pearson's correlation coefficient ranges between -1 and 1, where 0 means "no correlation", -1 means "perfect negative correlation" and 1 "perfect positive correlation". It measures the strength and direction of the relationship between two variables, i.e. in our case, between the metric score and the human judgment. On the sentence-level (also referred to as segment-level), Kendall's Tau is commonly used as a correlation

measure. As opposed to Pearson's correlation, Kendall's tau calculates correlations based on ranks of the variables. This means it does not assess the correlation between the absolute values but measures similarities in the ranking of the scores.

However, in recent years the reliability of Pearson's r has been questioned. The main issue with Pearson's correlation is its sensitivity to outliers which can heavily impact the correlation coefficient; i.e. it can in some cases seem like there is a strong correlation where there is not [Mathur et al., 2020a]. Pearson's r can also be unstable for small sample sizes (i.e. few MT systems) [Mathur et al., 2020a]. Additionally, Kocmi et al. [2021] point out that Pearson's correlation is not representative of actual use cases of MT metrics where scores are generally used to compare a pair of MT systems. For measuring system-level performance, Kocmi et al. [2021] thus suggest calculating pair-wise accuracy. They define pair-wise accuracy as follows:

$$\text{Accuracy} = \frac{|\text{sign}(\text{metric}\Delta) = \text{sign}(\text{human}\Delta)|}{|\text{all system pairs}|}$$

where:

$$\Delta = \text{score}(\text{System A}) - \text{score}(\text{System B})$$

Δ denotes the difference in scores between system A and and system B. Thus, for each system pair, the difference of metric scores (metricΔ) and the difference in average human judgments (humanΔ) is calculated.

4.3 Contrastive Evaluation: Fine-grained Evaluation of Metrics

Challenge sets or contrastive testsets can be used for targeted evaluation in various NLP tasks. They are testsets that measure the performance of systems on specific linguistic phenomena [Freitag et al., 2022; Sennrich, 2017]. Challenge sets are therefore useful for gaining a better understanding of systems or metrics and detecting their strengths and weaknesses. When evaluating MT metrics, they can thus provide insight into the metrics' capabilities of identifying certain errors. While there are a number of challenge sets for MT evaluation, for the task of evaluating MT metrics, challenge sets previously appeared in Freitag et al. [2021b]. To my awareness, there also exist four recently created challenge sets: (1) ACES [Amrhein et al., 2022] which focuses on translation accuracy errors. (2) The challenge set by Avramidis and Macketanz [2022] which is based on a manually devised MT test suite

and covers over 100 linguistically-motivated phenomena. (3) SMAUG [Alves et al., 2022] which uses sentence-level multilingual data augmentation to create translations with critical errors to evaluate the robustness of MT metrics to these errors. (4) The challenge set by Chen et al. [2022] which examines MT metrics' capability to relate synonyms and to identify critical errors at the word- and sentence-level.

A challenge set for the evaluation of MT metrics requires contrasting pairs of correct and incorrect translations. They typically contain a source sentence, a correct translation, an incorrect translation and a second correct translation (i.e. a human reference) [Amrhein et al., 2022]. The incorrect translations are usually constructed by inserting an error into the correct translation which leads to two versions of the same translation that differ only by one translation mistake. Below is an example from the ACES challenge set [Amrhein et al., 2022]:

(4.1) Source (de): Auf einem 100 Fuß langen Teilabschnitt läuft Wasser über den Damm.
Reference (en): Water is spilling over the levee in a section 100 feet wide.
Correct: On a 30.5 metres long section, water flows over the dam.
Incorrect: On a 100 metres long section, water flows over the dam.

Creating contrasting translations makes it possible to evaluate whether metrics can detect a specific translation error. Both the correct and incorrect versions are scored against the reference. A metric that is susceptible to the particular translation error is expected to score the correct version higher than the incorrect one. I.e. the challenge sets are used to test whether metrics assign the higher score to the correct translation.

4.4 Data and MT Metrics Evaluation Toolkit

As my experiments are based on existing metrics and in order to be able reproduce and compare results, I make use of utilities and data provided by the MT Metrics Eval Toolkit[2]. The toolkit was developed for the evaluation of MT metrics on testsets from the WMT Metrics Shared Tasks. It provides MT system outputs along with the respective source and reference translations. Additionally, it provides human scores for the system outputs and contains utilities for computing correlations between human and metric scores. Depending on the year and the language pair, different methods were used for obtaining human scores. However, the testsets utilized in this thesis are MQM-annotated by professionals (see 4.1.1).

[2] `https://github.com/google-research/mt-metrics-eval`

Testset	Sentences	Documents	Sentences per Doc	System Outputs	References
Newstest 2021	1'002	68	14.7	25	4
TED talks 2021	606	6	101	25	1
General 2022	2'037	181	11.3	17	2

Table 2: Overview of en→de WMT 2021 and 2022 testsets

For my experiments, I use English to German (en→de) testsets from the WMT 2021 and 2022 campaigns [Freitag et al., 2021b, 2022]. The WMT 2021 data includes testsets from the following two domains: news and (transcribed) TED talks. For the en→de Newstest 2021 set, 4 human reference translations are provided along with 25 MT system outputs. It contains 1'002 segments and 68 documents, i.e. on average there are around 14.7 sentences per document. The TED talks 2021 testset contains 1 human reference and 25 system outputs, there are 606 segments and 6 documents, i.e. on average 101 sentences per document. The general MT 2022 testset – which covers the domains news, social, conversational, and e-commerce – contains 2'037 segments and 181 documents with an average of roughly 11.3 sentences per document. So far, 2 human references and 17 system outputs are available.

5 Experiments: Context Awareness of Pre-trained Document-level MT Metrics

This chapter introduces the experimental setup and results for the first group of experiments. The experiments explore to what extent context is used in the pre-trained document-level metrics proposed by Vernikos et al. [2022]. Additionally, the experiments also aim to examine the metrics' sensitivity to discourse-level translation errors.

As mentioned in chapter 2.3.4, Vernikos et al. [2022] proposed to extend pre-trained MT metrics to the document-level by concatenating each sentence with two preceding context sentences to improve their embedding representation. Extending, BERTScore, Prism and COMET to the document-level leads to improved correlations over most of the testsets they used. The document-level metrics outperformed their sentence-level counterparts in 18 out of 24 cases with 6 out of these 18 cases being statistically significant ($p < 0.05$). The authors also evaluated the reference-free version of both COMET and Doc-COMET on a contrastive testset, reporting improved accuracy scores for Doc-COMET. However, it remains unclear to what extent these metrics actually make use of context, i.e. if the improved performance is due to a true understanding of context cues or whether these improvements are simply a consequence of other factors such as additional noise.

In order to examine the context usage in Doc-Prism, Doc-BERTScore and Doc-COMET, I thus carried out two experiments. In the first, preliminary experiment the added context sentences are corrupted in order to measure the impact on the resulting correlations with human judgments. In the second experiment, contrasting translations with discourse-level pronoun mistranslations are created and then scored with the metrics. The aim of the second experiment is to detect whether the metrics are sensitive to discourse-level mistakes, and thus, whether the additional context helps in detecting these errors.

5.1 Preliminary Experiment: Corruption of Context Sentences

I first conducted a preliminary experiment to measure how corrupting the added context sentences impacts correlations of metric scores with human judgments. Using corrupted (i.e. nonsensical) context sentences should provide insight into whether the document-level metrics make use of context cues. To begin with, I reproduced the results by Vernikos et al. [2022] for the en→de data using their publicly available code[1]. The resulting system-level correlations match the authors' reported results and are used as a baseline for the experiment (see table 3). As I work with en→de data, I only reproduced the results for the WMT en→de Newstest 2021 set and the TED talks 2021 testset. In a next step, I then corrupted the context sentences before concatenating them with each corresponding sentence, computed system-level scores and calculated Pearson correlations with human judgments (MQM annotations).

In order to be able to compare the results against the sentence-level and document-level baselines - i.e. the original, sentence-level version and the document-level version of the metrics – I used the same data, models and setup as Vernikos et al. [2022]. As discussed in chapter 2.3.4, this means that two previous context sentences were utilized. Specifically, each hypothesis sentence h was concatenated with two preceding reference sentences c_r and each reference sentence r was also concatenated with two previous reference sentences c_r. Additionally, for Doc-COMET, each source sentence s was concatenated with two preceding source context sentences c_s. Between each of the two context sentences and between the current sentence a sentence separator token is added. No context is used for the first sentence and only one context sentence for the second sentence. I.e. the sentences of each document are embedded in the following pattern: [**1**], [1,**2**], [1,2,**3**], [2,3,**4**], [3,4,**5**], and so on, with only the last sentence of each embedding input being scored by the metric. Thus, by corrupting the context sentences, the sentences being scored remain unchanged and only the embedding input is altered. For obtaining the baseline correlations, document boundaries were considered when adding the context sentences, however, when corrupting the context sentences they have no impact.

The following underlying, pre-trained models were used for each of the document-level metrics:

- Doc-Prism: mBART-50 [Tang et al., 2021]
- Doc-BERTScore: BERT-base model [Devlin et al., 2019]

[1]https://github.com/amazon-science/doc-mt-metrics

- Doc-COMET (WMT21-COMET-MQM): XLM-RoBERTa-large [Conneau et al., 2020]

5.1.1 Corruption of Context Sentences

The context sentences were corrupted in two simple ways:

1. Removing all added context sentences: context sentences are replaced with an empty string
2. Random permutation of the context sentences: any two random context sentences from the same testset are used

The first corruption was carried out by simply replacing all context sentences with an empty string. I.e. the resulting embedding input is essentially just each sentence (as is the case for the sentence-level version of the metrics). For the second corruption, the respective context sentences (i.e. the sentences from the corresponding testset) were shuffled to create a random permutation of the sentences. This results in each sentence being concatenated with two random context sentences which are from the same testset but not necessarily from the same document. For all three metrics, I corrupted the reference context which is used for both the reference and hypothesis. For Doc-COMET, the source context was also corrupted.

The underlying assumption of this experiment is that if the document-level metrics truly use context clues, then corrupting the context sentences, i.e. using random context sentences or removing context, should lead to correlations that are similar to or potentially even lower than the sentence-level baseline and lower than the document-level baseline. Otherwise, this would indicate that the improved performance of document-level metrics is due to other factors than a true understanding of context as no meaningful context is used when the context is corrupted.

5.1.2 Results

In table 3, system-level Pearson correlations of the metrics' scores with human judgments are reported for the en→de 2021 news and TED talks testsets. Correlations are reported for Prism, BERTScore and COMET as well as their document-level versions – which are used as a baseline. For Doc-Prism, Doc-BERTScore and Doc-COMET, correlations with human judgments after removing and shuffling context sentences are reported.

Model	Input / Context Corruption	TED talks	News
Prism (mBART-50)	$\langle h, r \rangle$	0.486	0.661
Doc-Prism (mBART-50)	$\langle c_r; h, c_r; r \rangle$	**0.692**	0.825
Doc-Prism (mBART-50)	$\langle c_r; h, c_r; r \rangle$, removed c	0.604	0.779
Doc-Prism (mBART-50)	$\langle c_r; h, c_r; r \rangle$, shuffled c	0.665	**0.856**
BERTScore	$\langle h, r \rangle$	0.506	0.930
Doc-BERTScore	$\langle c_r; h, c_r; r \rangle$	**0.613**	**0.948**
Doc-BERTScore	$\langle c_r; h, c_r; r \rangle$, removed c	0.509	0.936
Doc-BERTScore	$\langle c_r; h, c_r; r \rangle$, shuffled c	0.554	0.947
COMET (21-MQM)	$\langle s, h, r \rangle$	0.818	0.772
Doc-COMET (21-MQM)	$\langle c_s; s, c_r; h, c_r; r \rangle$	0.816	0.802
Doc-COMET (21-MQM)	$\langle c_s; s, c_r; h, c_r; r \rangle$, removed c	**0.827**	0.780
Doc-COMET (21-MQM)	$\langle c_s; s, c_r; h, c_r; r \rangle$, shuffled c	0.821	**0.804**

Table 3: System-level Pearson correlations with WMT 2021 MQM annotations of Doc-Prism, Doc-BERTScore and Doc-COMET with removed and shuffled context sentences. Correlations for the original version of the metrics are also reported.

In total there are 6 scenarios consisting of 3 metrics (each of which have multiple versions) and 2 testsets. In 3 out of these 6 scenarios, the document-level metric (without context corruptions) achieves the best correlation, in the other 3 cases the best correlation is reported for the document-level metric with corrupted context. In none of the instances does corrupting the context sentences – whether removing or shuffling the context – lead to correlations at or below the sentence-level baseline. Thus, even with corrupted context that is not meaningful, higher correlations are achieved compared to the sentence-level version of the metrics.

For Doc-Prism, corrupting the context mostly leads to correlations that are clearly better than the sentence-level baseline (no significance testing was performed) but lower than document-level baseline. Over the news testset, shuffling the context sentences results in an even stronger correlation than the document-level baseline. For Doc-BERTScore, the results when removing the context sentences are closer to what could be expected with correlations that are only slightly better than the sentence-level baseline. However, when context sentences are shuffled, the correlation over the news testset is almost at the document-level baseline with a difference of only 0.001. Finally, with Doc-COMET the best correlations are achieved with corrupted context sentences over both testsets. Both shuffling and removing context sentences leads

to better correlations over the TED talks testset than the baselines. Over the news testset, shuffling context sentences leads to a marginally higher correlation than the document-level baseline while removing context results in a correlation between the sentence- and document-level baselines. Overall, shuffling context sentences, i.e. selecting random sentences as context, generally produces stronger correlations than removing context sentences and results in similar correlations as the document-level baseline.

These results suggest that while by extending pre-trained metrics to the document-level improved correlations with human judgments are achieved, these improvements are likely due to non-linguistic factors such as added noise. The fact that in most instances adding random context sentences leads to correlations similar to the document-level baseline supports this assumption. The metrics' correlation improves regardless of whether preceding or random context sentences are used. The fact that completely removing context sentences mostly results in worse correlations than shuffling context sentences, suggests that the models still potentially learn something from the context. However, it seems that embedding additional context sentences does not lead to a true understanding of document-level contextual clues and that the context usage is non-interpretable. However, this does not necessarily mean that the metrics are unable to detect document-level mistakes, which will be evaluated in a second experiment.

5.2 Follow-up Experiment: Sensitivity to Discourse-level Errors

In order to examine the document-level metrics' sensitivity to discourse-level mistakes, I conducted a second experiment. In this experiment, I only evaluate Doc-COMET as well as COMET to also determine whether the document-level version reacts differently to the errors than the original version. As in the previous experiment, I used the WMT21-COMET-MQM model. To evaluate susceptibility to discourse-level errors, I created adversarial examples by changing a human reference translation and adding discourse-level errors[2]. The perturbed as well as the original reference translation are then scored against a second reference translation – i.e. the first reference (with and without the added error) is used as the hypothesis. A metric that is susceptible to the error would be expected to assign a higher score to the correct translation than to the incorrect version.

[2]The data is available on GitHub: `https://github.com/naomiblkr/doc-mt-metrics-eval/tree/main/pro_mistranslations`

In theory, it would also be possible to score the correct and incorrect reference translation against the same reference – however, then the version without the mistake would be the exact same as the reference. Thus, this task would be trivial. Therefore, to create these adversarial examples, no MT output is needed, however, in addition to the source text at least two reference translations are required. As in the previous experiment, I worked with WMT en→de data for which multiple human reference translations are available. For the following experiment, I used the 2021 news testset and the 2022 general testset. The created data contains: source sentences, a correct translation (i.e. a human reference), an incorrect translation (perturbed human reference) and a second human reference – the number of corresponding context sentences to be used can be adapted.

In the following sections, I first elaborate how the translation errors were created, then explain how the scores were obtained and finally, the results are reported.

5.2.1 Inspecting Document-level Errors

In order to determine which discourse-level mistakes could be added, I first inspected the datasets as well as the MQM error annotations to find possible errors where inter-sentential context might be useful to detect said errors. Possible phenomena where document-level context can help include: ambiguous words, pronouns, coreference, discourse connectives, lexical consistency / variability, coherence, and formality level. Inspecting the data quickly revealed that discourse-level errors can often not be as clearly categorized as sentence-level errors. Additionally, they usually do not occur as frequently – i.e. the number of clearly categorizable document-level errors is inherently small. Thus, translations with discourse-level mistakes cannot always be created by simply changing a single word or a small number of words. Another aspect to consider is that metrics have access to a reference translation, so in many instances where it is essential for an MT system to have access to inter-sentential context in order to disambiguate a phenomenon, this is not necessarily the case for MT metrics because the reference translation can provide enough information. It should also be mentioned that for other language pairs, different phenomena are more prevalent which were not discussed here as I only inspected the en→de datasets. For example, when evaluating translations from pro-drop languages (e.g. Chinese, Japanese) to non-pro-drop languages (e.g. English), the correct translation of zero pronouns is essential to the overall discourse coherence.

5.2.2 Creating Pronoun Translation Errors

I decided to focus on pronoun translation errors, as these can be added by changing single words, occur relatively frequently and because these mistakes are important to identify due to their impact on translation accuracy. Specifically, I focused on the translation of anaphoric occurrences of the pronoun "it" as the translation of said pronoun into German can be ambiguous (possible German translations are "es", "sie" and "er"). Additionally, if the antecedent which "it" refers to is not within the same sentence, inter-sentential context is required. Thus, only said instances of "it" were considered when creating the contrasting translations as to focus on discourse-level errors. Consider the following adversarial example which was created by changing the translation of "it" in a segment from reference A of the 2021 news testset:

(5.1) Source context (en): These neighbors and I share a fence.
Source (en): It separates our back yards.
Reference A context (de): Zwischen diesen Nachbarn und mir liegt ein Zaun.
Reference A: Er trennt unsere Hinterhöfe.
Perturbed reference A (de): Sie trennt unsere Hinterhöfe.
Reference B context (de): Diese Nachbarn und ich haben einen gemeinsamen Zaun.
Reference B (de): Er trennt unsere Gärten.

We need to know which word "it" refers to in the previous sentence in order to find the correct translation. Without considering the context sentence, "sie" could also be a correct translation of "it". In the example, reference A as well as the perturbed reference A are used as hypotheses – i.e. the original reference is used as the correct translation. Reference B is used as the actual reference translation. While the human reference can help with disambiguation in many instances, this is not always the case: if the antecedent is translated differently than in the reference, then the correct translation of "it" might also differ from the reference.

The translation errors were created semi-automatically with all four reference translations (A, B, C and D) of the en→de 2021 News testset as well as with reference A of the en→de 2022 general testset. I first tokenized the respective source and reference texts using Moses Tokenizer. In a second step, I filtered out sentences where the source sentence contains the pronoun "it", and the respective reference sentence contains one of the German pronouns "es", "sie" or "er". I then performed word alignment using SimAlign [Jalili Sabet et al., 2020] to get only the instances where "it" is aligned with one of the mentioned German pronouns. The next part

was done manually: I checked for anaphoric instances of "it" where the antecedent is not within the same sentence as "it" itself but in a preceding context sentence. If this was the case, I replaced the German pronoun by another one, ensuring that the sentence itself remains grammatically correct if no context sentences are considered. Per 1'000 segments (length of 2021 testsets) around 15 mistakes could be created on average. In some cases the same pronoun occurred twice in the same sentence (referring to the same antecedent), thus two mistakes were added in those segments which results in some sentences containing more severe mistakes. Some, although few, segments consist of multiple sentences which could impact the error severity as well (i.e. in longer segments, a singular error might be less severe).

5.2.3 Computing Scores for the Translations

In order to score the sentences where I added translation errors, I extracted each of the perturbed reference sentences along with the corresponding correct reference sentence, alternative reference sentence and source sentence. The sentences where concatenated with the corresponding two preceding context sentences $\langle c_s; s, c_h; h, c_r; r \rangle$. Unlike in the previous experiment, I concatenated the hypothesis sentences with the hypothesis context (and not with the reference context). The reasoning behind this is that the antecedent to which the pronoun refers to might have been translated differently in the two references, therefore, using the reference for the hypothesis context could lead to inconsistencies. I only scored the segments were errors were added (i.e. not the whole testsets). For the 2021 news testset, I scored the translations the following way:

- reference A (with and without added errors) against reference B: 11 sentences with a total of 13 added errors
- reference B (with and without added errors) against reference A: 15 sentences with a total of 17 added errors
- reference C (with and without added errors) against reference D: 15 sentences with a total of 17 added errors
- reference D (with and without added errors) against reference C: 11 sentences with a total of 12 added errors

And additionally, for the 2022 general testset:

- reference A (with and without added errors) against reference B: 24 sentences with a total of 26 added errors

For all of these scenarios, I computed scores with both COMET as well as Doc-COMET to determine whether the two metrics react differently to the discourse-level errors. Of course, for COMET no context sentences where added. For each scenario, the average over all segment-level scores was computed.

5.2.4 Results

	COMET		Doc-COMET	
	with errors	no errors	with errors	no errors
Ref A on Ref B (news 2021)	0.0425	0.0439	0.0165	0.0176
Ref B on Ref A (news 2021)	0.0416	0.0439	0.0159	0.0175
Ref C on Ref D (news 2021)	0.0438	0.0462	0.0162	0.0180
Ref D on Ref C (news 2021)	0.0445	0.0474	0.0141	0.0158
Ref A on Ref B (general 2022)	0.0407	0.0420	0.0123	0.0143
Overall average	0.0426	0.0447	0.0150	0.0166

Table 4: COMET and Doc-COMET scores for contrasting translations with and without pronoun mistranslations

Both COMET and Doc-COMET score the correct translations higher than the versions with the errors in all instances. This suggests that there is at least some sensitivity to this specific type of discourse-level error. Whether the metrics are susceptible enough to the mistakes is challenging to interpret due to the scoring ranges. Doc-COMET seems to generally assign lower scores than COMET which indicates that the two models have a slightly different scoring range. While the latest models of COMET such as WMT22-COMET-DA [Rei et al., 2022] have a scoring range between 0 and 1, previous versions such as the model used for this experiment do not have a clearly defined scoring range which makes scores difficult to interpret. For sentence-level COMET, the scores for the correct translations are on average roughly 0.0021 lower while for Doc-COMET they are around 0.0016 lower. Although the scoring range is unclear, we would expect bigger scoring differences between the translations with and without errors if the metrics are susceptible enough to the mistakes due to the fact that every sentence contains a mistake which clearly impacts translation accuracy. As a consequence of the different scoring ranges of sentence-level and document-level COMET, while it seems that Doc-COMET could be slightly more sensitive to the discourse-level errors (taking into consideration the generally lower scores), we cannot draw any real conclusions from these results. For the 2021 general testset – where more sentences were scored than for the other

scenarios – the scores are overall slightly lower for both the correct and incorrect translations.

6 Experiments: ChatGPT as a Document-level MT Metric

This chapter introduces the methodology applied in the second group of experiments which explores the capabilities of ChatGPT as a document-level MT evaluation metric. I first discuss how GEMBA [Kocmi and Federmann, 2023] is extended to the document-level (Doc-GEMBA) and then examine whether ChatGPT is an effective document-level MT metric by adopting a number of evaluation methods. First, pair-wise accuracy is computed for Doc-GEMBA, then its sensitivity to discourse-level errors is examined and finally, I score the output of a document-level and sentence-level MT system with Doc-GEMBA. Additionally, I conduct some qualitative analyses for further insights.

As mentioned in chapter 3.2.1 Kocmi and Federmann [2023] proposed using GPT Brown et al. [2020] as an MT evaluation metric (GEMBA). They designed prompts so that GPT scores a single hypothesis sentence at a time. However, as discussed in chapter 3.2.3, LLMs are capable of modeling coherence and consistency over longer sequences of text. This thus raises the question whether GPT models would also be capable of document-level MT evaluation and whether they have the ability to utilize context cues to spot discourse-level MT errors. While there are numerous GPT models, I only use ChatGPT for my experiments. GPT-4 would potentially achieve better performances, however, access is currently restricted. Similarly, the text completion model Davinci - for which Kocmi and Federmann [2023] reported the best results out of all the GPT models – has higher costs for access than ChaptGPT.

6.1 Extending GEMBA to the Document-level

In order to use ChatGPT for document-level MT evaluation, I selected two of the prompts suggested by Kocmi and Federmann [2023] as a template: GEMBA-DA and GEMBA-SQM. Both of these prompts ask GPT to score the MT hypothesis on a scale from 0 to 100 with respect to the human reference translation based

on whether the meaning of the text has been preserved. I then adapted them by (1) incorporating two preceding context sentences, (2) adding to the prompt, that only the last sentence of the translation segment should be scored and (3) explicitly mentioning that "fluency" should also be evaluated and not solely the preservation of meaning. The aim was that three sentences would be provided per prompt, asking ChatGPT to only score the final sentence – i.e. the current sentence is scored while the two preceding sentences are used for context clues only. When adding the context sentences, document-boundaries are considered. The sentences per document are thus scored as follows for a document containing 6 sentences: [**1**], [1,**2**], [1,2,**3**], [2,3,**4**], [3,4,**5**], [4,5,**6**]. The number in bold is the sentence that should be scored. In this way, all sentence transitions are being captured as opposed to say scoring sentences 1-3, 4-6, etc. per prompt. The approach is similar to methods for eliciting human judgments where raters have access to context sentences but only annotate (i.e. score) one sentence at a time.

```
DA_doc: Score the last sentence of the following translation from {source_lang} to
{target_lang} with respect to the human reference on a continuous scale 0 to 100
where score of zero means "no meaning preserved, not fluent" and score of one hundred
means "perfect fluency, meaning and grammar".

{source_lang} source: "{src_seg}"
{target_lang} human reference: {ref_seg}
{target_lang} machine translation: "{trg_seg}"
Score:
```

Figure 4: DA-based document-level prompt. "Src seg", "ref seg" and trg seg" are replaced with the respective segment, i.e. the sentence being scored and the two previous context sentences. Prompt adapted from Kocmi and Federmann [2023]

```
SQM_doc: Score the last sentence of the following machine translation from {source_lang} to
{target_lang} with respect to the human reference on a continuous scale from 0 to 100
that starts with "No meaning preserved, not fluent", goes through "Some meaning preserved,
lacking fluency", then "Most meaning preserved, mostly fluent", up to "Perfect fluency,
meaning and grammar".

{source_lang} source: "{src_seg}"
{target_lang} human reference: "{ref_seg}"
{target_lang} machine translation: "{trg_seg}"
Score (0-100):
```

Figure 5: SQM-based document-level prompt. Prompt adapted from Kocmi and Federmann [2023]

One possible disadvantage of this method is that ChatGPT might not actually score only the last sentence but all three sentences provided in the prompt. If that is the case, then the final two sentences of each document would also not be weighted equally because they are only prompted once or twice respectively whereas the

other sentences are prompted three times. The latter problem could be avoided by scoring the last two sentences again, thus allowing for equal weighting. However, I decided against this approach because this would make the document longer, i.e. it would lead to more segment-level scores than there are segments which would make further analyses difficult. Another possible solution could be to ignore document boundaries but this would then impact the relevance of the context in instances where cross-document context is used.

Each source, hypothesis and reference sentence was concatenated with two preceding context sentences $\langle c_s; s, c_h; h, c_r; r \rangle$. I.e. the source sentences were concatenated with source context, hypothesis sentences with hypothesis context and reference sentences with reference context. No sentence boundary tags were inserted between sentences. This decision was based on the fact that when prompting ChatGPT to answer which sentence of a text sequence is the last one, it seemed capable of recognizing sentence boundaries. However, this does have the downside that in instances where no end of sentence punctuation marks are present (e.g. after titles, lists, etc.), there is no clear sentence boundary. Due to API request limits, the prompts in Figure 4 and Figure 5 – which will be referred to as Doc-GEMBA-DA and Doc-GEMBA-SQM – are the only two prompts used in the following experiments as well as their sentence-level counterparts, i.e. GEMBA-DA and GEMBA-SQM.

6.2 Correlations of Doc-GEMBA Scores with Human Judgments

In order to evaluate the performance of ChatGPT as a document-level metric, the first experiment determines how well Doc-GEMBA correlates with human quality judgments by measuring system-level, pair-wise accuracy. This is the same method as Kocmi and Federmann [2023] used for evaluating GEMBA and should thus allow for comparison of performances of the proposed document-level prompts with their sentence-level version.

I used MQM data from the WMT22 Metrics shared task, i.e. the en→de 2022 general testset for computing scores and calculating pair-wise accuracy. I.e. the human gold standard scores are MQM ratings. As mentioned in chapter 4.2, system-level, pair-wise accuracy measures the number of MT systems ranked correctly by the metric. Kocmi et al. [2021] argue that this method is more suitable than e.g. Pearson correlation because we are usually interested in comparing two MT systems.

I computed pair-wise accuracy using publicly availabe code[1] by Kocmi and Federmann [2023] which was extended to include the suggested, document-level prompts[2]. It should be mentioned that I utilized reference B as the human reference translation as opposed to reference A, which – to my knowledge – was used for obtaining the results reported by Kocmi and Federmann [2023]. As a consequence, results might differ slightly. The temperature – which controls the randomness of the generated text – is set to 0 and is only increased if no valid answer is returned by ChatGPT, i.e. the randomness of the response increases if no valid answer is given. I first computed results using the original sentence-level GEMBA as a baseline and then computed results for Doc-GEMBA.

6.2.1 Results

Metric	Pair-Wise Accuracy
GEMBA-DA	82.1%
GEMBA-SQM	88.5%
Doc-GEMBA-DA	89.7%
Doc-GEMBA-SQM	89.7%

Table 5: System-level, pair-wise accuracy for GEMBA and Doc-GEMBA for DA- and SQM-based prompts using ChatGPT.

Metric	Pair-Wise Accuracy
GEMBA-DA	81.0%
GEMBA-SQM	85.0%

Table 6: System-level, pair-wise accuracy for GEMBA using ChatGPT as reported by Kocmi and Federmann [2023]

For both document-level prompt designs, improved pair-wise accuracy is achieved with both prompts obtaining the same accuracy score of 89.7%. While the document-level version of the SQM-based prompt only leads to an improvement of 1.2%, extending the DA-based prompt produces a more significant improvement of 7.6%. As expected, the results for GEMBA-DA and GEMBA-SQM differ from the results reported by Kocmi and Federmann [2023]. For GEMBA-DA the results differ by 1.1% while for GEMBA-SQM they differ by 3.5%. This is possibly due to the usage

[1] `https://github.com/MicrosoftTranslator/GEMBA`

[2] Document-level version of GEMBA: `https://github.com/naomiblkr/Doc-GEMBA`

of a different human reference. However, scores might also be unstable due to model updates of ChatGPT as well as some degree of randomness in the provided answers.

Overall, extending the prompts to include inter-sentential context leads to better pair-wise accuracy scores, indicating that ChatGPT could be useful for document-level MT evaluation. However, it is unclear whether these improvements are due to usage of context cues, i.e. whether ChatGPT can detect and correctly score discourse-level MT errors by using context. This will be examined in chapter 6.5. Additionally, for the SQM-based prompt, the improvement is not as notable. Thus, more experiments need to be conducted to draw any conclusions.

6.3 Analysis of Score Distribution

To further examine the scores assigned by Doc-GEMBA and to inspect whether the scores are in the desired range, I computed absolute and relative frequencies for segment-level scores. As Kocmi and Federmann [2023] put forward, GEMBA returns a discrete value between 0–100 as opposed to other automatic evaluation metrics. They argue that, as a result, there is a high probability of translations receiving the same score – which could potentially be an issue when comparing high quality systems, especially at the segment-level. The authors therefore calculated the frequency of each distinct score that was assigned at the segment-level. The resulting distribution showed that for the DA-based prompt the scores 80, 90, or 95 were in total assigned for more than three-quarters of all scores. They also found that with the DA-based prompt, only scores that are multiples of five are produced (e.g. 70, 75, 80, etc.).

The distribution of scores for Doc-GEMBA can be seen in table 7. For both document-level prompts, 90 is by far the most frequently assigned score, making up 58.51% of all scores for the DA-based prompt and 62.36% of all scores for the SQM-based prompt. For the DA-based metric, the scores 90 and 95 alone make up over 87% of scores, for the SQM-based metric this number is lower at over 73.2%. Doc-GEMBA-SQM also gave more scores of 80 (10.69%) than Doc-GEMBA-DA (3.2%).

This distribution does raise a few questions with regards to how suitable ChatGPT is as a document-level MT metric. The MT system outputs used in this experiment are arguably of high quality and thus explain the large number of very high scores to some extent. However, such scores leave little room for improved MT systems receiving better scores. As a consequence, it is somewhat questionable whether Doc-

	Doc-GEMBA-DA		Doc-GEMBA-SQM	
Score	Abs. Freq.	Rel. Freq.	Abs. Freq.	Rel. Freq.
0	3	0.01%	3	0.01%
5	-	-	-	-
10	11	0.03%	13	0.04%
15	-	-	-	-
20	16	0.05%	22	0.06%
25	-	-	-	-
30	9	0.03%	25	0.07%
35	-	-	-	-
40	16	0.05%	38	0.11%
45	-	-	-	-
50	60	0.17%	68	0.2%
55	-	-	-	-
60	122	0.35%	362	1.05%
65	-	-	-	-
70	706	2.04%	1457	4.21%
75	-	-	-	-
80	1’109	3.20%	3’702	10.69%
85	12	0.03%	1’292	3.73%
90	20’262	58.51%	21’595	62.36%
95	10’030	28.96%	3’747	10.82%
100	2’273	6.56%	2’305	6.66%

Table 7: Distribution of segment-level Doc-GEMBA scores

GEMBA has the ability to accurately evaluate and discriminate (document-level) MT systems that only demonstrate minor improvements. Moreover, it is also unclear whether Doc-GEMBA can successfully spot and score errors at the segment-level.

6.4 Analysis of Large Score Discrepancies

In order to further explore the scores assigned by Doc-GEMBA in comparison to scores given by GEMBA and to potentially analyze whether Doc-GEMBA improves at scoring segments where context awareness is required, I filtered out sentences where there is a large discrepancy in the scores assigned by the two metrics. I analyzed 30 sentences each, where the scoring discrepancy between the sentence-

level and document-level version of GEMBA was either exactly 25 or above 40. In both instances, the first 30 sentences with the respective score difference were selected for the analysis. Only scores computed with the DA-based prompt were utilized. This analysis is by no means systematic or comprehensive – it simply aims to further explore the scoring behaviour of ChatGPT. If there is a large score discrepancy this could also indicate that the metrics struggled to score the respective sentences, thus leading to inconsistent scoring. Because of this and the few segments selected, the analysis might be somewhat biased.

6.4.1 Translations with a Score Discrepancy of 25

For segments with a score discrepancy of 25, Doc-GEMBA gave a higher score than GEMBA in most cases. Only two of the 30 sentences that I analyzed, received the higher score by GEMBA. To test this further, I computed the total number of times Doc-GEMBA assigned the higher score when there was a score discrepancy of 25 between the two metrics. The result shows that in 144 out of 151 cases Doc-GEMBA gave the higher score. For the selected 30 sentences, Doc-GEMBA most frequently gave a score of 95 and thus GEMBA most frequently gave a score of 70. One possible explanation for this could be that Doc-GEMBA generally awards higher scores (as can also be seen in the distribution of scores). In order to find out whether this is the case because e.g. Doc-GEMBA generally awards higher scores or can better evaluate high quality translations, I analyzed if the scores for the selected segments are correct, i.e. which of the two scores is more accurate – which is of course somewhat subjective.

I found that the score by Doc-GEMBA is incorrect for around half of the sentences. I.e. the score is too high and the score by GEMBA would be more accurate. In many of these cases, the hypothesis contains poor phrasing or wording and/or unknown tokens. For translations where Doc-GEMBA assigned the correct score (compared to GEMBA), this does not seem to be due to improved context awareness as for most of these segments no context is required to detect errors. However, there are two to three translations that contain at least somewhat ambiguous words which Doc-GEMBA scored more accurately compared to GEMBA. Whether this is because of context awareness is unclear because the reference translation is likely sufficient for disambiguation in these cases. By contrast, the example below illustrates an instance where context usage should help to recognize a pronoun mistranslation, however, Doc-GEMBA gave a score of 95 (despite the hypothesis also being worded badly). Thus, the score by GEMBA is arguably more accurate.

(6.1) Score GEMBA: 70
Score Doc-GEMBA: 95
Source context (en): Q. I have been waiting a while for my order and it hasn't arrived.
Source (en): Is **it** possibly lost?
Hypothesis context (de): F. Ich habe eine Weile auf meine Bestellung gewartet und sie ist nicht angekommen.
Hypothesis (de): Ist es vielleicht verloren?

6.4.2 Translations with a Score Discrepancy of 40 or More

For segments with a score difference of 40 or larger, similar observations can be made. While the higher score is also more frequently assigned by Doc-GEMBA than GEMBA, there are a few more cases where GEMBA gave the higher score. In total, Doc-GEMBA returned the higher score for 120 out of 163 sentences. According to my judgment, the score by Doc-GEMBA is not accurate for over half of the 30 translations. Generally, a large number of the analyzed MT hypotheses seem to be of comparatively low quality. In some cases the source text is also lacking coherence. This might suggest that low quality MT output is scored either differently by the two metrics or that GEMBA and Doc-GEMBA generally score low quality translations inconsistently. The sentences were Doc-GEMBA incorrectly awarded high scores frequently contain poor phrasing or accuracy errors.

Interestingly, in some of the instances where the score by Doc-GEMBA is the lower one, the hypothesis sentence is a relatively good translation, however, the preceding context sentence contains mistakes or is a generally poor translation. This indicates that errors or low translation quality in the context sentences might impair the score of the current sentence, i.e. propagation of errors could be a potential issue. In the example given below, not only the hypothesis context but also the source context are of poor quality.

(6.2) Score GEMBA: 100
Score Doc-GEMBA: 40
Source context (en): Pleaae check your mail moment please I wait here for you
Source (en): How did it go?
Hypothesis context (de): Bitte überprüfen Sie Ihren E-Mail-Moment Ich warte hier auf dich
Hypothesis (de): Wie ist es gelaufen?

While propagation of errors can negatively impact segment-level performance of metrics, system-level performance should not be impacted as much. It could also be argued that a poor translation of a sentence not only impacts the quality of the respective sentence but also the overall coherence of the document. However, we cannot conclude whether errors are in fact propagated due to the low number of segments that I analyzed.

Hypotheses which Doc-GEMBA scored more accurately compared to GEMBA mostly did not contain any discourse-level phenomena where context would be required for accurate scoring. By contrast, the example below shows a translation where context awareness could potentially help with the disambiguation of the word "close" and thus help to score the MT output accordingly. However, Doc-GEMBA assigned a score of 90 although the hypothesis is clearly inaccurate. The segment is challenging to both translate and score because it contains an ellipsis of "the restaurant" (i.e. "they had me close the restaurant") and the reference might be more useful for accurate scoring than context in such cases.

(6.3) Score GEMBA: 30
Score Doc-GEMBA: 90
Source context (en): Yeah, when I was 16 I applied and got offered a job at a restaurant. Dishwasher.
Source (en): First shift they had me close.
Hypothesis context (de): Ja, als ich 16 war, bewarb ich mich und bekam einen Job in einem Restaurant angeboten. Geschirrspülmaschine.
Hypothesis (de): In der ersten Schicht hatten sie mich in der Nähe.

6.5 Sensitivity to Discourse-level Errors

I conducted a further experiment to evaluate Doc-GEMBA's capabilities with regards to sensitivity discourse-level translation errors. The aim is to determine whether the metric can correctly score a hypothesis sentence that contains a discourse-level translation mistake. This should also provide insights into ChatGPT's abilities to use context clues. For this experiment, I used the same data and approach as in chapter 5.2, i.e. the discourse-level phenomenon I examined are anaphoric instances of the pronoun "it". I evaluated Doc-GEMBA's sensitivity to pronoun mistranslations by scoring a correct and incorrect translation (i.e. unperturbed and perturbed reference) against a second reference. If Doc-GEMBA is susceptible to the pronoun errors, we would expect it to score the translation without the error higher than the

translation with the error.

Scores between 0 and 100 were computed using both document-level prompts (SQM and DA) and the temperature was set to 0 to avoid randomness in the generated answers. For each reference, only sentences with the added errors were scored and the average score is reported. The source, hypothesis and reference sentences were concatenated with the corresponding preceding two context sentences. In total, 76 segments were scored for each of the four scenarios (i.e. once the correct translation, once the incorrect translation for each of the two prompts). Finally, the overall average is also reported.

6.5.1 Results

	Doc-GEMBA-DA		Doc-GEMBA-SQM	
	with errors	no errors	with errors	no errors
Ref A against Ref B (news 2021)	90.00	90.00	86.36	88.18
Ref B against Ref A (news 2021)	90.00	90.33	89.33	88.67
Ref C against Ref D (news 2021)	90.00	90.00	88.67	89.67
Ref D against Ref C (news 2021)	88.64	89.09	85.45	87.27
Ref A against Ref B (general 2022)	89.17	89.79	87.29	87.71
Overall average	89.56	89.84	87.42	88.30

Table 8: Average Doc-GEMBA (DA and SQM prompts) scores for translations with and without pronoun translation errors

In all but three instances the version with the pronoun errors has a lower score than the correct translation. However, the scores are only marginally lower. Because every sentence being scored (except for the context sentences) contains an error, scores of the correct and incorrect translations would have been expected to be more distinguishable. In two instances the scores are exactly the same and in one case the reference without an error has a lower score.

Thus, while the metrics are potentially able to recognize the pronoun translation errors they are arguably not susceptible enough to them, i.e. do not weight them accurately. However, because the differences in scores are minor, it is questionable whether the metrics actually detect the errors. Considering that the scores given by Doc-GEMBA are always multiple of fives, most segments seem to have received the same score, whether they contain an error or not. On the one hand, the results might suggest that Doc-GEMBA does not use context clues to resolve ambiguities, on the other hand, it could also indicate that Doc-GEMBA generally performs worse

at the segment-level. Finally, except for the DA-based prompt generally leading to higher scores, the DA-based metric and SQM-based metric behave similarly, i.e. one of the two is not more sensitive to pronoun mistranslations.

6.5.2 Analysis of Prompt Responses

In most cases, ChatGPT does not only return a score for the translation but also gives an explanation, i.e. text is generated that gives feedback concerning the translation quality. The responses explicitly mention translation errors and possible improvements that could be made to the translation. In order to further explore how Doc-GEMBA reacts to the pronoun translation errors and whether these errors are actually spotted, I analyzed these explanations[3]. I selected the prompt responses of reference B (15 prompts) and reference D (11 prompts), each for both prompt designs (DA and SQM) and both versions of the translations, i.e. with and without errors. Thus, in total, I examined 104 prompt responses. The aim of this analysis is to gain more insight into the scoring behaviour of ChatGPT with regards to discourse-level errors, i.e. this is by no means a comprehensive or systematic analysis.

Of all the examined explanations for incorrect translations (i.e. 52 responses), only four explicitly mention that the respective pronoun was mistranslated. In three out of these four cases, the score (90) is not worse than in instances where no pronoun translation errors are explicitly mentioned. In one of the explanations, the pronoun mistranslation is called a "minor error" (see example).

(6.4) Response text by ChatGPT:
90.
The machine translation preserves most of the meaning and is mostly fluent, with only minor errors such as the use of "er" instead of "es" in the last sentence.

In all other instances, the responses are overall very similar. First, ChatGPT gives a general feedback which is usually the following: "The machine translation preserves most of the meaning and is mostly fluent". Some explanations additionally mention more specific information such as "only minor errors in word choice and grammar" or "minor differences in wording and phrasing compared to the human reference translation".

[3]CSV files containing the prompt responses are available on GitHub: `https://github.com/naomiblkr/doc-mt-metrics-eval/tree/main/GPT_score_contrastive/gpt_responses`

Quite frequently the explanation also mentions explicit words that differ from the reference translation, saying that said words could be improved. Moreover, errors and explicit suggestions for possible improvements are mentioned just as frequently for the correct translations as for the perturbed translations (see examples in tables 13 and 14 in appendix). This could indicate that the explanations are the result of a mostly surface-level comparison with the reference translation. Although there are also several explanations that mention the source text, which would suggest that not only the reference translation is taken into consideration. In a few cases the translation direction is mixed up in the explanation (see example).

(6.5) Response text by ChatGPT:
90.
The machine translation is very good and captures the meaning of the original sentence accurately. However, there are a few minor errors in grammar and word choice that could be improved upon. For example, "Vernehmungsbefehl" should be translated as "questioning warrant" instead of "interrogation warrant". Additionally, the use of "feststellte" instead of "betonte" in the second sentence changes the tone slightly and could be seen as less forceful. Overall, the translation is very close to the human reference and conveys the same message effectively.

In some instances, the explanations mention explicit errors that do not occur in the current sentence being scored, i.e. the last sentence of the prompt, but are actually in the preceding two context sentences. Thus, this suggests that Doc-GEMBA does not solely consider the final sentence of the prompt when assigning a score but also the added context sentences. Finally, examining the segment-level scores reveals that, for most segments, there is no difference in scores of the correct and incorrect translations. Overall, the explanations provided do not show that Doc-GEMBA is capable of reliably detecting pronoun translation mistakes. Thus, although the experiment covers only one specific error category, it remains questionable whether ChatGPT has the ability to recognize and accurately score discourse-level MT errors.

6.6 Scoring Output of Sentence- and Document-level MT systems

The final experiment is intended to test ChatGPT's capability of accurately evaluating output of document-level MT systems and whether Doc-GEMBA can discriminate MT output from sentence-level and document-level systems. Ultimately, one

of the reasons for developing document-level MT metrics is to accurately evaluate high quality MT systems. Document-level systems might only make minor improvements compared to state-of-the-art sentence-level systems such as better fluency and consistency. Thus, metrics need to be sensitive to such minor changes to accurately score MT output. Therefore, this experiment obtains system-level scores for the output of a sentence- and document-level MT system. If Doc-GEMBA is capable of accurately evaluating at the document-level, it would be expected that the output of the document-level system receives a better score. Of course, this hypothesis is based on the assumption that the document-level MT system is in fact of higher quality than the sentence-level system.

6.6.1 Data

The MT output utilized for this experiment was kindly provided by Samuel Läubli who trained sentence- and document-level MT systems with the same data (en→de WMT22). Thus, differences in the performance of the systems should not be due to different training data. The provided MT hypotheses were obtained by translating a testset containing 9'587 sentences (500 documents) with the respective MT system. The average number of sentences per document is 19.2. The testset consists of document-level data from Europarl v10, News Commentary v16 and Tilde Rapid (WMT21). I used the output of the sentence-level system (SENT) and of two document-level MT systems: DOC-ORIG (which was trained respecting document boundaries) and DOC-LINE (which was trained ignoring document boundaries).

6.6.2 Experimental Setup

I utilized the document-level DA prompt as introduced in chapter 6.1 to get segment-level scores between 0 and 100. The responses by ChatGPT were then parsed to only extract the score from each generated text. System-level scores were computed by averaging segment-level scores. If no score was returned by ChatGPT, which was the case for roughly 10 out of the 9'587 sentences, then the respective segment was not considered when computing the system-level score. The temperature was set to 0 and if no response was generated, it was increased gradually.

As discussed before, the approach of incorporating two preceding context sentences might lead to the last two sentences of each document being weighted less. However, this should only be the case if ChatGPT does not recognize the last sentence of the segment (or does not understand the prompt) and thus scores all three sentences

per prompt. To examine the impact of this, I additionally computed scores with a slightly different approach for the output of the DOC-LINE and sentence-level systems. For each document, the final two sentences were scored again to counteract this issue. Let s_n be the last sentence of a document, then the following two prompts were added: one where s_n was concatenated with one preceding sentence, i.e. $\langle s_{n-1}; s_n \rangle$, and another with only s_n. For example, for a document consisting of five sentences, the sentences would be prompted in the following pattern: [1], [1,2], [1,2,3], [2,3,4], [3,4,5], [4,5], [5] instead of [1], [1,2], [1,2,3], [2,3,4], [3,4,5]. However, this only works for computing system-level scores as this approach will result in more segment-level scores than there are segments in the document because two additional scores are calculated per document.

6.6.3 Results

Results for all three systems are reported for both the original Doc-GEMBA-DA prompt defined in chapter 6.1 as well as the approach which ensures equal weighting of the last two sentences per document as introduced in chapter 6.6.2. The system-level score for DOC-ORIG was only computed with the standard approach.

	Input / Context Incorporation	
System	Standard approach	Equal weighting
DOC-ORIG	90.79	-
DOC-LINE	91.77	91.69
SENT	**91.83**	**91.78**

Table 9: System-level Doc-GEMBA-DA scores for output of document-level and sentence-level MT systems. Scores are reported for the standard approach (6.1) and the approach which ensures equal weighting of all segments (6.6.2)

The results show that, in both scenarios, Doc-GEMBA gives the output of the sentence-level system the better system-level score. For the standard approach, DOC-ORIG has the lowest score out of all three systems while DOC-LINE and the sentence-level system (SENT) obtained very similar scores with a difference of roughly 0.06. Weighting all sentences per document equally resulted in only marginally different scores: scores for both DOC-LINE and SENT are slightly lower. However, the sentence-level system still obtains a better score than the document-level system with equal weighting of the sentences; the difference in scores between the two systems being marginally larger (0.09 difference). Like in previous experiments, scores are overall very high, which is to some extent expected with high

quality MT output – but it also leaves very little room for scoring improved systems, especially considering that the sentence-level system received a score of almost 92 out of 100.

As briefly mentioned in the previous chapter, for some sentences, ChatGPT did not return a score instead generating an answer that no translation was provided in the prompt. Examining these instances showed that a translation (i.e. source, hypothesis and reference) was in fact provided, however, the segments only consisted of a single word or abbreviation such as "DE" (i.e. it was the first sentence of a document were no context is added). It seems thus that ChatGPT does not always recognize very short translation segments.

Overall, the results show that Doc-GEMBA cannot successfully discriminate between the output of document- and sentence-level systems. Of course, this finding is based on the assumption that the output of the document-level systems is of higher quality than the sentence-level system output. We cannot conclude whether ChatGPT is inherently incapable of accurately scoring output of document-level MT systems because aspects such as different prompt designs might lead to different results (see chapter 7.3).

7 Discussion

In this chapter each of the research questions is revisited by discussing and synthesizing the results. Potential limitations of the respective experiments are also pointed out.

7.1 Research Question 1

The first research question aimed to assess whether pre-trained metrics that are expanded to the document-level as proposed by Vernikos et al. [2022] utilize context and have a true understanding of context cues. The first experiment thus tested context usage of Doc-Prism, Doc-BERTScore and Doc-COMET.

The results in chapter 5 show that utilizing random context sentences (from the same testset but not necessarily the same document) leads to improved correlations with human judgments for the document-level metrics compared to their sentence-level counterparts. Even without using any context sentences (i.e. simply adding an empty string) correlations improve compared to the sentence-level metrics – although at times only marginally. Thus, whether the preceding context sentences are concatenated with each sentence or other, non-meaningful context is utilized does not make a clear difference.

These findings imply that the improved correlation with human quality judgments of the document-level metrics compared to their sentence-level version is due to non-linguistic factors and cues (e.g. added noise). Therefore, we can argue that the metrics do not demonstrate a true, interpretable understanding of context. These results are similar to the findings about non-interpretable context usage in context-aware MT [Kim et al., 2019], as discussed in chapter 2.1.1. What this means for the reliability of the metrics is unclear. On the one hand, the mostly improved correlations of Doc-Prism, Doc-BERTScore and Doc-COMET compared to their sentence-level version suggests that adding context sentences to the embedding is useful. On the other hand, the high correlations even after corrupting context sentences raise the question of why this is the case and whether the metrics can

detect and accurately score discourse-level errors which is crucial for document-level metrics.

As the results after removing and shuffling context sentences were rather clear, I did not conduct any further experiments with context corruptions. However, other experiments – such as using context from a different domain – could potentially provide more insights. The results also made it somewhat redundant to further explore whether the metrics mainly leverage hypothesis, reference or (for Doc-COMET) source context sentences.

7.2 Research Question 2

The aim of the second research question was to examine the sensitivity of the document-level metrics proposed by Vernikos et al. [2022] to discourse-level translation errors. For this research question, I focused on Doc-COMET, i.e. the other document-level metrics were not evaluated. Additionally, I aimed to explore whether Doc-COMET is more susceptible to discourse-level errors than COMET, i.e. whether the additional context improves sensitivity to such errors. While many document-level errors seem to be relatively rare, mistranslation of such phenomena can not only impact the overall fluency of the translation but also its accuracy and coherence. Thus, it is crucial for automatic metrics to accurately score discourse-level errors.

To answer this research question, mistranslations of anaphoric occurrences of the pronoun "it" were added to 76 segments of human reference translations. In all instances the antecedent is not in the same sentence but in one of the preceding context sentences, i.e. context is potentially needed to correctly score the mistakes. The sentences with and without the added errors were then scored against a second human reference. While the approach of adding an error to a human reference made it possible to isolate the specific phenomenon (i.e. the metric score should only be impacted by the specific error and not by any other errors), one drawback is that this is not the most realistic scenario. In practise, an MT output can contain multiple errors per segment.

In chapter 5.2, the sensitivity of Doc-COMET as well as the original, sentence-level version of COMET to the errors was tested. The results show that both COMET and Doc-COMET correctly assign the better score to the translations without errors in almost all scenarios. However, for both metrics, the differences in the scores is only minor. Because each sentence being scored contains a relatively important mis-

take, the score discrepancy between the version with and without the errors should be larger. Therefore, although the metrics show some awareness of discourse-level errors they are arguably not sensitive enough to these specific pronoun translation mistakes. Due to the scoring ranges of COMET and Doc-COMET not being clearly defined, it is difficult to draw final conclusions. Reporting segment-level scores could have provided additional information about the differences in scores and thus might have helped with interpreting the scores.

The fact that sentence-level COMET achieved similar results as Doc-COMET (i.e. also gave the correct translations marginally better scores) suggests that extending COMET to the document-level might not improve its sensitivity to discourse-level errors. However, as discussed in chapter 2.3.4, Vernikos et al. [2022] reported improved accuracy for the reference-free version of Doc-COMET compared to COMET on contrastive testsets covering discourse-level phenomena. In their paper, the authors reported the accuracy (i.e. percentage where metric gave the higher score to the correct translation) while I computed absolute scores (and used different data), thus these results cannot be compared directly. It is also possible that the length of the sentences/segments being scored might have impacted scores. The segments from the testsets that I used for my experiments are relatively long, thus errors might not impact scores as strongly as errors in shorter segments.

Of course, the experiment only examines the metrics' susceptibility to one specific discourse-level phenomenon. To draw more decisive conclusions, other linguistic phenomena would have to be evaluated, especially for other language pairs. It should also be mentioned that a potential issue with the setup of this experiment is that the correct pronoun translation can often be inferred from the reference translation, i.e. it is questionable whether context cues are truly needed for scoring this specific discourse-level error. However, this is not always the case because if the antecedent is translated differently than in the reference translation, the respective pronoun might also differ. Thus, while Doc-COMET is arguably not sensitive enough to pronoun translation errors, it is still unclear whether adding context sentences has any impact on the metric's ability of scoring these or similar discourse-level errors.

7.3 Research Question 3

The final research question was whether ChatGPT is a useful document-level MT evaluation metric. Because this is a very general question, it was divided into three sub-questions. First, I evaluated how well system-level scores assigned by Doc-GEMBA correlate with human judgments. Second, I explored the metric's

sensitivity to discourse-level errors and its usage of context cues. Third, I tested if ChatGPT can discriminate between the output of sentence-level and document-level MT systems.

The results are contradictory to some degree. On the one hand, adapting the sentence-level prompts proposed by Kocmi and Federmann [2023] to include inter-sentential context leads to improved system-level, pair-wise accuracy. The high pair-wise accuracy indicates that ChatGPT has the capability of discriminating and ranking MT systems accurately and that adding document-level context improves the performance. However, further experiments and analyses reveal some of ChatGPT's and/or Doc-GEMBA's limitations concerning document-level MT evaluation.

Firstly, the distribution of segment-level scores illustrates that ChatGPT very frequently assigns scores of 90 (and 95 in the case of the DA-based prompt). This leaves little room for improved hypotheses obtaining higher scores. The fact that so many segment-level scores are the same is also indicative of possibly worse performance at the segment-level. It especially raises the question whether Doc-GEMBA is capable of detecting and accurately scoring minor differences or accuracy errors – which then in turn also influence the overall coherence and fluency – at the segment-level. Results by Kocmi and Federmann [2023] have shown that GEMBA performs worse at the segment-level than at the system-level, suggesting that this is likely also an issue for Doc-GEMBA.

Findings in chapter 6.4 where sentences with large score discrepancies between GEMBA and Doc-GEMBA were analyzed revealed further issues with segment-level scoring and context awareness. The analysis showed that Doc-GEMBA does not seem capable of assigning more accurate scores than GEMBA for instances were context awareness should help with e.g. the disambiguation of words. However, due to the small number of sentences analyzed and because only sentences with large score discrepancies were analyzed these findings are not conclusive and are likely somewhat biased. The analysis exposed another potential issue of Doc-GEMBA: the propagation of errors in context sentences. Although, this presumably only impacts segment-level scores but not system-level scores. The problem might also be avoided by adapting the pattern in which sentences are scored, i.e. how context sentences are incorporated.

Moreover, the results of chapter 6.5, show that Doc-GEMBA does not seem to be sensitive enough to discourse-level pronoun translation mistakes. Whether this is due to Doc-GEMBA not utilizing context clues or simply because ChatGPT generally seems to perform worse at the segment-level is unclear. However, the analysis of the prompt responses (chapter 6.5.2) further underlines that Doc-GEMBA only rarely

recognizes the pronoun translation errors and if it does, they are only judged as a minor errors which do not impact the score. Similarly, findings concerning the usage of ChatGPT for document-level MT by Wang et al. [2023b] – who report lower accuracy of ChatGPT in contrastive testing compared to other translation models – support the conclusion that ChatGPT struggles with resolving discourse-level ambiguities. However, as discussed in chapter 3.2.3, their findings also indicate that using GPT-4 could lead to improvements in this aspect.

The results of the final experiment (see chapter 6.6) revealed the arguably biggest issue of Doc-GEMBA (and potentially ChatGPT) as a document-level metric. Doc-GEMBA was unable to accurately discriminate output from sentence-level and document-level MT systems, assigning the output of the sentence-level system the better score. Thus, the metric is not useful for discriminating or ranking high-quality document-level systems. Again, we cannot conclude that ChatGPT is inherently unable to perform this task or if the performance could be improved with a different prompt design (such as adding a sentence boundary tags).

Finally, there is a more general issue with using ChatGPT as an MT quality evaluator that is not specific to document-level evaluation but is nonetheless relevant. Scores generated by a GPT model might not be comparable with previous scores due to model updates and/or randomness in the generated answers; which in turn makes it difficult to compare or reproduce results. As a consequence, while there is potential in using ChatGPT (and other GPT models) for document-level MT evaluation, it is questionable how useful it is in practice.

7.4 Future Work

The present thesis could be expanded upon and the findings also brought up more questions left to explore both for the development as well as the evaluation of document-level MT metrics. To better evaluate document-level MT metrics, more discourse-level phenomena could be examined using contrastive testsets. This would be especially interesting and relevant for other language pairs where specific discourse-level aspects are more prominent.

Regarding human annotations of MT quality that are used as the gold standard for computing correlations with metrics' scores, there are - to my awareness - no "true" document-level judgments. For example, with MQM-annotations, although human raters have access to document-level context, errors are only annotated at the segment-level. It might be interesting to see how assigning an overall fluency

and coherence score per document (in addition to segment-level error annotations) impacts results and whether such human judgments could also be used to train document-level metrics. A similar approach of combining a document-level fluency and coherence score with sentence-level accuracy scores could also be interesting for developing or improving metrics. However, it is questionable how useful this would be because these categories cannot be separated completely as sentence-level accuracy also impacts the overall coherence and fluency of translations.

With regards to GPT as an MT evaluator, it would be interesting to test other GPT models, e.g. GPT-4 which should have better context awareness. It could also be interesting to experiment with different prompt designs and other approaches for incorporating document-level context. Using other language pairs would also be relevant, as findings by Hendy et al. [2023] about utilizing GPT for MT suggest that it only achieves competitive quality in high-resource settings, thus this likely also applies to MT evaluation.

8 Conclusion

In this Master's thesis, I first evaluated to what extent context is used in pre-trained document-level MT metrics (Doc-COMET, Doc-BERTScore and Doc-Prism). Results showed that the context usage of these metrics is not interpretable, i.e. they show no true understanding of context cues. Moreover, it is likely that the metrics are not sensitive enough to discourse-level errors. Overall, the metrics' inability to understand inter-sentential context cues suggests that they are not useful for document-level MT evaluation and the development or improvement of such metrics remains an open issue.

In the second part of this thesis, I investigated whether ChatGPT is a useful document-level MT metric. To this end, I designed two prompts that incorporate two preceding context sentences when asking ChatGPT to score each segment of a document. The prompts for Doc-GEMBA are based on the sentence-level metric GEMBA [Kocmi and Federmann, 2023] and are designed to return scores between 0 and 100. Results of the evaluation are two-fold. On the one hand, improved pair-wise accuracy was achieved (89.7% for both prompt designs), suggesting that the metric is capable of correctly ranking MT systems. On the other hand, further experiments as well as qualitative analyses revealed that Doc-GEMBA is not susceptible enough to specific discourse-level errors and might not be capable of leveraging context cues for disambiguation. Finally, the metric was also unable to accurately discriminate output from document-level and sentence-level MT systems. Thus, while there seems to be potential in using ChatGPT for document-level MT evaluation, there currently are some limitations. However, some weaknesses might be improved upon by utilizing other models such as GPT-4 and by adapting the prompt design.

References

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

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

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

E. Avramidis and V. Macketanz. Linguistically motivated evaluation of machine translation metrics based on a challenge set. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 514–529, Abu Dhabi, United Arab Emirates (Hybrid), 2022. Association for Computational Linguistics.

D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. 2015. 3rd International Conference on Learning Representations, ICLR 2015.

S. Banerjee and A. Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan, 2005. Association for Computational Linguistics.

R. Bawden, R. Sennrich, A. Birch, and B. Haddow. Evaluating discourse phenomena in neural machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1304–1313, New Orleans, Louisiana, 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1118.

T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020.

X. Chen, D. Wei, H. Shang, Z. Li, Z. Wu, Z. Yu, T. Zhu, M. Zhu, N. Xie, L. Lei, S. Tao, H. Yang, and Y. Qin. Exploring robustness of machine translation metrics: A study of twenty-two automatic metrics in the WMT22 metric task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 530–540, Abu Dhabi, United Arab Emirates (Hybrid), 2022. Association for Computational Linguistics.

A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.747.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423.

P. Fernandes, K. Yin, G. Neubig, and A. F. T. Martins. Measuring and increasing context usage in context-aware machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1:*

Long Papers), pages 6467–6478, Online, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.505.

M. Freitag, D. Grangier, and I. Caswell. BLEU might be guilty but references are not innocent. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 61–71, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.5.

M. Freitag, G. Foster, D. Grangier, V. Ratnakar, Q. Tan, and W. Macherey. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474, 2021a. doi: 10.1162/tacl_a_00437.

M. Freitag, R. Rei, N. Mathur, C.-k. Lo, C. Stewart, G. Foster, A. Lavie, and O. Bojar. Results of the WMT21 metrics shared task: Evaluating metrics with expert-based human evaluations on TED and news domain. In *Proceedings of the Sixth Conference on Machine Translation*, pages 733–774, Online, 2021b. Association for Computational Linguistics.

M. Freitag, R. Rei, N. Mathur, C.-k. Lo, C. Stewart, E. Avramidis, T. Kocmi, G. Foster, A. Lavie, and A. F. T. Martins. Results of WMT22 metrics shared task: Stop using BLEU – neural metrics are better and more robust. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 46–68, Abu Dhabi, United Arab Emirates (Hybrid), 2022. Association for Computational Linguistics.

Z. Gong, M. Zhang, and G. Zhou. Document-level machine translation evaluation with gist consistency and text cohesion. In *Proceedings of the Second Workshop on Discourse in Machine Translation*, pages 33–40, Lisbon, Portugal, 2015. Association for Computational Linguistics. doi: 10.18653/v1/W15-2504.

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

A. Hendy, M. Abdelrehim, A. Sharaf, V. Raunak, M. Gabr, H. Matsushita, Y. J. Kim, M. Afify, and H. H. Awadalla. How good are gpt models at machine translation? a comprehensive evaluation, 2023.

P. Isabelle, C. Cherry, and G. Foster. A challenge set approach to evaluating machine translation. In *Proceedings of the 2017 Conference on Empirical*

Methods in Natural Language Processing, pages 2486–2496, Copenhagen, Denmark, 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1263.

M. Jalili Sabet, P. Dufter, F. Yvon, and H. Schütze. SimAlign: High quality word alignments without parallel training data using static and contextualized embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1627–1643, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.147.

Y. Jiang, T. Liu, S. Ma, D. Zhang, J. Yang, H. Huang, R. Sennrich, R. Cotterell, M. Sachan, and M. Zhou. BlonDe: An automatic evaluation metric for document-level machine translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1550–1565, Seattle, United States, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.111.

N. Kalchbrenner and P. Blunsom. Recurrent continuous translation models. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1700–1709, Seattle, Washington, USA, 2013. Association for Computational Linguistics.

Y. Kim, D. T. Tran, and H. Ney. When and why is document-level context useful in neural machine translation? In *Proceedings of the Fourth Workshop on Discourse in Machine Translation (DiscoMT 2019)*, pages 24–34, Hong Kong, China, 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-6503.

T. Kocmi and C. Federmann. Large language models are state-of-the-art evaluators of translation quality, 2023.

T. Kocmi, C. Federmann, R. Grundkiewicz, M. Junczys-Dowmunt, H. Matsushita, and A. Menezes. To ship or not to ship: An extensive evaluation of automatic metrics for machine translation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 478–494, Online, 2021. Association for Computational Linguistics.

P. Koehn. *Evaluation*, page 217–246. Cambridge University Press, 2009. doi: 10.1017/CBO9780511815829.009.

P. Koehn and C. Monz. Manual and automatic evaluation of machine translation between European languages. In *Proceedings on the Workshop on Statistical*

Machine Translation, pages 102–121, New York City, 2006. Association for Computational Linguistics.

S. Läubli, R. Sennrich, and M. Volk. Has machine translation achieved human parity? a case for document-level evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4791–4796, Brussels, Belgium, 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1512.

S. Lee, J. Lee, H. Moon, C. Park, J. Seo, S. Eo, S. Koo, and H. Lim. A survey on evaluation metrics for machine translation. *Mathematics*, 11(4), 2023. ISSN 2227-7390. doi: 10.3390/math11041006.

C.-k. Lo. YiSi - a unified semantic MT quality evaluation and estimation metric for languages with different levels of available resources. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 507–513, Florence, Italy, 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-5358.

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

Q. Ma, J. Wei, O. Bojar, and Y. Graham. Results of the WMT19 metrics shared task: Segment-level and strong MT systems pose big challenges. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 62–90, Florence, Italy, 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-5302.

B. Marie, A. Fujita, and R. Rubino. Scientific credibility of machine translation research: A meta-evaluation of 769 papers. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7297–7306, Online, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.566.

S. Maruf, A. F. T. Martins, and G. Haffari. Selective attention for context-aware neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3092–3102,

Minneapolis, Minnesota, 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1313.

N. Mathur, T. Baldwin, and T. Cohn. Tangled up in BLEU: Reevaluating the evaluation of automatic machine translation evaluation metrics. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4984–4997, Online, 2020a. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.448.

N. Mathur, J. Wei, M. Freitag, Q. Ma, and O. Bojar. Results of the WMT20 metrics shared task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 688–725, Online, 2020b. Association for Computational Linguistics.

K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, USA, 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135.

M. Popović. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal, 2015. Association for Computational Linguistics. doi: 10.18653/v1/W15-3049.

Y. Qin and L. Specia. Truly exploring multiple references for machine translation evaluation. In *Proceedings of the 18th Annual Conference of the European Association for Machine Translation*, pages 113–120, Antalya, Turkey, 2015.

R. Rei, C. Stewart, A. C. Farinha, and A. Lavie. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.213.

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

R. Rei, J. G. C. de Souza, D. Alves, C. Zerva, A. C. Farinha, T. Glushkova, A. Lavie, L. Coheur, and A. F. T. Martins. COMET-22: Unbabel-IST 2022

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

T. Sellam, D. Das, and A. Parikh. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.704.

K. Semenov and O. Bojar. Automated evaluation metric for terminology consistency in MT. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 450–457, Abu Dhabi, United Arab Emirates (Hybrid), 2022. Association for Computational Linguistics.

R. Sennrich. How grammatical is character-level neural machine translation? assessing MT quality with contrastive translation pairs. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 376–382, Valencia, Spain, 2017. Association for Computational Linguistics.

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

Y. Tang, C. Tran, X. Li, P. Chen, N. Goyal, V. Chaudhary, J. Gu, and A. Fan. Multilingual translation with extensible multilingual pretraining and finetuning. *CoRR*, abs/2008.00401, 2020.

Y. Tang, C. Tran, X. Li, P.-J. Chen, N. Goyal, V. Chaudhary, J. Gu, and A. Fan. Multilingual translation from denoising pre-training. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3450–3466, Online, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.304.

B. Thompson and M. Post. Automatic machine translation evaluation in many languages via zero-shot paraphrasing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 90–121, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.8.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

G. Vernikos, B. Thompson, P. Mathur, and M. Federico. Embarrassingly easy document-level MT metrics: How to convert any pretrained metric into a document-level metric. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 118–128, Abu Dhabi, United Arab Emirates (Hybrid), 2022. Association for Computational Linguistics.

D. Vilar, G. Leusch, H. Ney, and R. E. Banchs. Human evaluation of machine translation through binary system comparisons. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 96–103, Prague, Czech Republic, 2007. Association for Computational Linguistics.

E. Voita, R. Sennrich, and I. Titov. When a good translation is wrong in context: Context-aware machine translation improves on deixis, ellipsis, and lexical cohesion. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1198–1212, Florence, Italy, 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1116.

J. Wang, Y. Liang, F. Meng, Z. Sun, H. Shi, Z. Li, J. Xu, J. Qu, and J. Zhou. Is chatgpt a good nlg evaluator? a preliminary study, 2023a.

L. Wang, C. Lyu, T. Ji, Z. Zhang, D. Yu, S. Shi, and Z. Tu. Document-level machine translation with large language models, 2023b.

T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*, 2020.

A Tables

Tables 10, 11 and 12 show three examples where pronoun translation errors were created as described in chapter 5.2.2. In these examples, only one preceding context sentence is shown, however, the perturbed translations were created in such a way, that any number of context sentences can be used. The full overview of the data can be found in the GitHub repository.

Example 1	
Wrong	Sie wurde 1905 aufgestellt und stand vor dem alten Gerichtsgebäude Anne County, am gleichen Ort wo einst Sklaven versteigert wurden.
Correct	Es wurde 1905 aufgestellt und stand vor dem alten Gerichtsgebäude Anne County, am gleichen Ort wo einst Sklaven versteigert wurden.
Context	Arbeiter entfernten das 27 Fuß (8,2 Meter) hohe Denkmal am Samstag morgen, berichteten die Medien.
Reference	Es wurde 1905 aufgestellt und stand vor dem alten Gerichtsgebäude von Princess Anne County, an der Stelle, an der einst Sklavenauktionen stattfanden.
Reference Context	Arbeiter bauten das 8,2 Meter hohe Denkmal am Samstagmorgen ab, berichteten die Medien.
Source	It was installed in 1905 and stood outside the old Princess Anne County courthouse, in the same place that slave auctions were once held.
Source Context	Workers took down the 27-foot (8.2-meter) tall monument on Saturday morning, media outlets reported.

Table 10: Example 1 anaphoric ”it” translation error

Example 2	
Wrong	Er sagte, dass die Maßnahme „Tausende von britischen Reiseplänen ins Chaos stürzt“
Correct	Sie sagte, dass die Maßnahme „Tausende von britischen Reiseplänen ins Chaos stürzt“
Context	British Airways sagte, dass sie enttäuscht sind über die neuesten Änderungen an den Reiseempfehlungen und Regeln, wobei sie jedoch sagten, dass ihre Flüge weiterhin stattfinden werden.
Reference	Sie sagte, dass der Schritt „die Reisepläne von Tausenden von Briten ins Chaos stürzt“
Reference Context	British Airways äußerte sich „enttäuscht“ über die jüngsten Änderungen der Reisehinweise und -regeln der Regierung, obwohl die Fluggesellschaft sagte, dass ihre Flüge weiterhin durchgeführt würden.
Source	It said the move was ”throwing thousands of Britons’ travel plans into chaos.”
Source Context	British Airways said it was ”disappointed” about the latest changes to the government’s travel advice and rules, although the airline said its flights were continuing to operate.

Table 11: Example 2 anaphoric ”it” translation error

Example 3	
Wrong	Er fügte hinzu, dass nur Passagiere mit einem Zertifikat für solch einen Test in das Land gelassen werden.
Correct	Es fügte hinzu, dass nur Passagiere mit einem Zertifikat für solch einen Test in das Land gelassen werden.
Context	Ab Dienstag müssen Reisende, die mit dem Flugzeug aus Rumänien nach Griechenland einreisen, den griechischen Grenzbehörden einen negativen PCR-Tests für den Coronavirus vorlegen, der höchstens 72 Stunden vor der Einreise in unser südliches Nachbarland durchgeführt wird, teilte das Außenministerium mit.
Reference	Weiter hieß es, dass nur Passagiere, die ein Zertifikat für diesen Test hätten, einreisen dürften.
Reference Context	Ab Donnerstag müssten Flugreisende von Bulgarien nach Griechenland griechischen Grenzbeamten einen PCR-Test mit negativem Ergebnis für den Coronavirus vorweisen, der längstens 72 Stunden vor der Einreise bei unserem südlichen Nachbarn gemacht werden müsse, gab das Außenministerium bekannt.
Source	They added that only passengers who have a certificate for such a test will be allowed in the country.
Source Context	From Tuesday, those traveling by plane from Bulgaria to Greece will have to present to the Greek border authorities a negative result from a PCR test for coronavirus, performed up to 72 hours before their entry into our southern neighbor, the Ministry of Foreign Affairs announced.

Table 12: Example 3 anaphoric ”it” translation error

Perturbed Reference B, DA-based prompt	
Score	Response text
90	90. The machine translation accurately conveys the meaning of the original sentence and uses appropriate grammar and vocabulary. However, there are some minor differences in word choice and sentence structure compared to the human reference.
90	90. The machine translation accurately conveys the main points of the sentence and maintains fluency, but there are some minor errors in word choice and grammar. For example, "zusammengetan" is not the best translation for "collude" and "er zur Zahlung" should be "sie zur Zahlung" to refer to the bank.
90	90. The machine translation is very close to the human reference, with only minor differences in wording and grammar. The meaning is preserved and the sentence is fluent.
90	90. The machine translation accurately conveys the meaning of the original sentence and uses appropriate grammar and vocabulary. However, there are some minor differences in phrasing and word choice that could be improved upon.
90	90. The machine translation is very close to the human reference, with only minor differences in word choice and sentence structure. The meaning and grammar are preserved, and the translation is fluent.
90	90. The machine translation accurately conveys the meaning of the source text and uses appropriate grammar and vocabulary. However, there are some minor differences in phrasing and word choice that make the translation slightly less fluent than the human reference.
90	90. The machine translation is very close to the human reference, with only minor differences in word choice and sentence structure. The meaning and fluency are preserved, and the grammar is correct.
90	90. The machine translation is very close to the human reference, with only minor differences in wording and grammar. The meaning is preserved and the sentence is fluent.

Table 13: Examples ChatGPT Prompt Responses, Perturbed Reference B

Original Reference B, DA-based prompt

Score	Response text
90	90. The machine translation accurately conveys the meaning of the original sentence and uses appropriate grammar and vocabulary. However, there are some minor differences in word choice and sentence structure compared to the human reference.
90	90. The machine translation accurately conveys the main points of the sentence and maintains fluency, but there are some minor errors in word choice and phrasing. For example, "zusammengetan" is not the best translation for "collude" and "verurteilt" is not the best translation for "ordered."
90	90. The machine translation is very close to the human reference, with only minor differences in wording and grammar. The meaning is preserved and the sentence is fluent.
90	90. The machine translation accurately conveys the meaning of the original sentence and uses appropriate grammar and vocabulary. However, there are some minor differences in phrasing and word choice that could be improved upon.
90	90. The machine translation accurately conveys the meaning of the original sentence and uses appropriate vocabulary and grammar. However, there are some minor errors such as using "Limodosen" instead of "Getränkedosen" and "Konservendosen" instead of "Dosenöffner".
90	90. The machine translation accurately conveys the meaning of the source text and uses appropriate grammar and vocabulary. However, there are some minor differences in phrasing and word choice that make the translation slightly less fluent than the human reference.
90	90. The machine translation accurately conveys the meaning of the source text and is grammatically correct. However, there are some minor differences in word choice and sentence structure that could be improved upon.
90	90. The machine translation is very close to the human reference, with only minor differences in wording and phrasing. The meaning and grammar are preserved and the translation is fluent.

Table 14: Examples ChatGPT Prompt Responses, Original Reference B

B JSON Examples

Examples 1 and 2 illustrate some of the extracted translation segments (and context sentences) where there was a large discrepancy in scores assigned by GEMBA and Doc-GEMBA. The score assigned by the original, sentence-level GEMBA is named "sent_score" and the score assigned by Doc-GEMBA is called "doc_score". The complete JSON files – of which the first 30 items were used for the analysis in chapter 6.4 – can be found in the GitHub repository: `https://github.com/naomiblkr/Doc-GEMBA/tree/main/score_distribution`.

```
[{
    "score_diff": 25,
    "doc_score": 95,
    "sent_score": 70,
    "hyp_context": "Wenn Sie Ihre Rückerstattung nicht
        erhalten haben, nachdem wir einen Artikel als versandt
        markiert haben, senden Sie uns bitte eine Nachricht,
        damit wir die Rückerstattung bearbeiten können. F.
        Enthalten Sie eine Quittung?",
    "hyp_sent": "A. Nein, wir legen keine Quittungen in
        Pakete bei, es sei denn, sie werden angefordert.",
    "src_context": "If you haven't received your refund
        after we have marked an item as dispatched, please
        send us a message so that we can have the refund
        processed. Q. Do you include a receipt?",
    "src_sent": "A. No, we do not include receipts in
        packages unless requested."
},

{
    "score_diff": 25,
    "doc_score": 95,
    "sent_score": 70,
    "hyp_context": "Wenn Sie einen Mehrwertsteuerbeleg
        benötigen, kontaktieren Sie uns bitte und wir können
        Ihnen einen per E-Mail zusenden. F. Ich habe eine
        Weile auf meine Bestellung gewartet und sie ist
        nicht angekommen.",
    "hyp_sent": "Ist es vielleicht verloren?",
    "src_context": "If you require a VAT receipt,
        please contact us and we can email one over. Q.
        I have been waiting a while for my order and
        it hasn't arrived.",
    "src_sent": "Is it possibly lost?"
}]

```


Listing 1: JSON example of sentences with a score discrepancy of 25

```
[{
    "score_diff": 50,
    "doc_score": 70,
    "sent_score": 20,
    "hyp_context": "Es tut mir wirklich leid, dass Sie
        dieses Problem mit Ihrem eBook haben, aber ich
        bin bereit, Ihnen zu helfen. Ich werde mit ein
        paar Schritten teilen, um in Ihr Gerät zu spielen,
        okay?",
    "hyp_sent": "Ich bin in der Personalabteilung und
        habe in der Vergangenheit in der Gehaltsabrechnung
        gearbeitet.",
    "src_context": "I'm really sorry to know that you
        are having this issue with your eBook, but I'm
        willing to help you. I'll share with a couple
        of steps to perform into your device, okay?",
    "src_sent": "I'm in HR and have worked payroll in the past."
},{
    "score_diff": 70,
    "doc_score": 90,
    "sent_score": 20,
    "hyp_context": "Geschirrspüler. Bei der ersten Schicht
        hatten sie mich in der Nähe.",
    "hyp_sent": "Samstag.",
    "src_context": "Dishwasher. First shift they had me close.",
    "src_sent": "Saturday."
},{
    "score_diff": 40,
    "doc_score": 60,
    "sent_score": 20,
    "hyp_context": "Schritt 5: Gehen Sie durch die Wagen und
        suchen Sie nach etwas Kaltem und Produziertem.
        Schritt 6: Rückkehr sagte kalt / produzieren.",
    "hyp_sent": "Schritt 7: Bedecken Sie alle Nicht-Tür-Kälte,
        d. H. Käse / Fleisch / Gemüse usw.",
    "src_context": "Step 5: Go through the carts looking
        for anything cold and produce. Step 6: Return said
        cold/produce.",
    "src_sent": "Step 7: Cover all non-doored cold
        i.e. cheese/meat/veggies, etc."
}]
```


Listing 2: JSON example of sentences with a score discrepancy of 40 or more