University of Zurich^UZH

Department of Computational Linguistics

Master's Thesis
Faculty of Arts and Social Sciences

Schneeschrauben, Himmelbäume, Gartengebäcke

Investigating and Enhancing the Sensitivity of Trained Neural Metrics to German Compounds

Supervisor: Prof. Dr. Rico Sennrich

Author:
Sarah Elisabeth Kiener
Gotthardstrasse 30
6410 Goldau
09-110-958

June 1, 2023

Abstract

The recent progress in the field of Neural Machine Translation (NMT) called for more sophisticated evaluation metrics that are capable of accurately assessing high-quality Machine Translation (MT) outputs. In response to this need, a new generation of MT evaluation metrics has been put forth. These metrics are themselves trained neural networks and are currently considered state-of-the-art. However, they have been repeatedly shown to suffer from certain blind spots that cause them to unpredictably assign high scores to bad translations. As these metrics become more widely adopted, the NLP community is at risk of optimising towards their weaknesses. Hence, it is of crucial importance to uncover, and ideally remove, the pathologies of neural metrics.

This thesis contributes to the investigation of the shortcomings of neural metrics by scrutinising their sensitivity to German compounds. Following the approach by Amrhein and Sennrich (2022), I will use the metrics under study as utility function in sampling-based Minimum Bayes Risk (MBR) decoding to explore and quantify their deficiencies.

In a case study on COMET-20, I will show that it is not sensitive enough to German compounds and frequently rewards nonsensical translations with high scores. Having identified this blind spot, I will delve into strategies to address it. I will show that the underlying language model plays a major role in the behaviour of the metric. Replacing the multilingual language model of COMET-20 with a monolingual German one, substantially enhances the metric's sensitivity to errors in German compounds, nouns, named entities and numbers. Pre-training the language model with Whole Word Masking further promotes the metric's sensitivity to mistakes in compounds and improves the quality of MBR-decoded translations. However, the blind spots cannot be fully removed.

Further, I will address the issue of metric overfitting and propose an approach to alleviate the problem. Combining the scores of two metrics during MBR decoding does not only improve the translation quality, but also counteracts the overfitting effect.

Moreover, I will show that when comparing two identical segments, neural metrics assign unforeseeable scores that vary from segment to segment. This behaviour has implications for the implementation of MBR decoding.

Finally, the results indicate that the system-level ranking of MT systems, commonly used as evaluation measure for metrics, is not an appropriate method to suitably capture the quality of an MT evaluation metric.

Zusammenfassung

Die jüngsten Fortschritte auf dem Gebiet der neuronalen maschinellen Übersetzung erfordern neue, präzisere Evaluationsmetriken, die in der Lage sind, qualitativ hochwertige maschinelle Übersetzungen (mÜ) korrekt zu bewerten. Als Reaktion darauf wurden Metriken entwickelt, die selbst trainierte neuronale Netzwerke sind und den neuesten Stand der Technik reflektieren. Allerdings weisen diese Metriken gewisse blinde Flecken auf und bewerten schlechte Übersetzungen oft unerwartet gut. Da diese Metriken zunehmend verwendet werden, läuft die NLP-Gemeinschaft Gefahr, auf deren Schwächen hin zu optimieren. Um dies zu verhindern, sollten jene Schwächen möglichst bald erkannt und idealerweise beseitigt werden.

Die vorliegende Masterarbeit leistet einen Beitrag zur Erforschung dieser Schwächen, indem sie die Sensitivität neuronaler Metriken gegenüber deutschen Komposita beleuchtet. Dazu werden die untersuchten Metriken als Nutzenfunktion im stichprobenbasierten Minimum Bayes Risk (MBR) Decoding verwendet (vgl. Amrhein and Sennrich, 2022).

Eine Fallstudie zeigt auf, dass COMET-20 zu wenig sensibel auf Fehler in deutschen Komposita reagiert und unsinnige Übersetzungen mit einer guten Bewertung belohnt. Verschiedene Strategien zur Beseitigung dieses blinden Fleckens werden untersucht. Dabei wird deutlich, dass das Verhalten der Metrik wesentlich durch das ihr zugrunde liegende Sprachmodell beeinflusst wird. Wenn das multilinguale Sprachmodell von COMET-20 durch ein monolinguales, deutsches ersetzt wird, erhöht sich die Sensitivität für Fehler in deutschen Komposita, Nomen, Eigennamen und Zahlen. Wird das Sprachmodell zudem mit Whole Word Masking vortrainiert, erhöht sich die Sensitivität für Komposita weiter. Auch die Qualität der MBR-dekodierten Übersetzungen verbessert sich. Die blinden Flecken werden jedoch nicht vollständig beseitigt.

Ausserdem werde ich auf das Problem des “Metric Overfitting” eingehen und einen möglichen Lösungsansatz präsentieren. Durch das Kombinieren der Bewertungen zweier Metriken während des MBR-Decodings wird nicht nur die Übersetzungsqualität besser, sondern auch der Overfitting-Effekt abgeschwächt.

Zudem werde ich aufzeigen, dass neuronale Metriken nicht immer gleich reagieren, wenn sie zwei identische Segmente miteinander vergleichen. Die Punktzahlen, die sie dabei vergeben, sind unvorhersehbar und variieren von Segment zu Segment. Dies hat Auswirkungen auf die Implementierung des MBR-Decodings.

Schliesslich deuten die Ergebnisse darauf hin, dass die Erstellung einer Rangliste für mÜ-Systeme, wie sie häufig zur Evaluation von Metriken verwendet wird, keine geeignete Methode ist, um die Qualität einer mÜ-Metrik angemessen zu erfassen.

Acknowledgement

First of all, I would like to express my sincere gratitude to my supervisor Prof. Dr. Rico Sennrich for his kind support and valuable advice throughout my thesis. His inspiring ideas and his expertise guided this work and his constructive feedback was a constant source of motivation for me.

I would like to express my heartfelt thanks to Chantal Amrhein for her kind helpfulness and constructive support over the past months. Her detailed explanations and her precious feedback contributed greatly to the completion of this thesis.

Many thanks to Branden Chan, Stefan Schweter and Timo Möller for sharing their GBERT-Data model with me. By providing me with their unpublished model, they rendered the experiments and the investigation of the research questions of this thesis possible.

I thank my parents and my sister from the bottom of my heart for supporting me in so many ways throughout my studies. Their constant encouragement and precious advice were a great help throughout this long and sometimes challenging journey.

Finally, my wholehearted thanks to all the others who encouraged and supported me throughout the process of writing this thesis.

Contents

List of Figures

List of Tables

List of Acronyms

BEER	Better Evaluation as Ranking
BERT	Bidirectional Encoder Representations from Transformers
BLEU	Bilingual Evaluation Understudy
BLEURT	Bilingual Evaluation Understudy with Representations from Transformers
CharacTER	Translation Edit Rate on Character Level
ChrF	Character n-gram F-score
COMET	Crosslingual Optimized Metric for Evaluation of Translation
DA	Direct Assessment
ESIM	Enhanced Sequential Inference Model
GBERT	German BERT
GBLEURT	German BLEURT
GCOMET	German COMET
GPT	Generative Pre-Trained Transformer
HTER	Human-mediated Translation Edit Rate
MAP	Maximum A Posteriori
MBR	Minimum Bayes Risk
METEOR	Metric for Evaluation of Translation with Explicit Ordering
MLM	Masked Language Modelling
MQM	Multidimensional Quality Metrics
MSE	Mean Squared Error
MT	Machine Translation
NE	Named Entity
NER	Named Entity Recognition
NLP	Natural Language Processing
NMT	Neural Machine Translation
NSP	Next Sentence Prediction
QE	Quality Estimation
RoBERTa	Robustly Optimized BERT Pre-training Approach
ROUGE	Recall-Oriented Understudy for Gisting Evaluation
RUSE	Regressor Using Sentence Embeddings
SMOR	Stuttgart Morphological Analyzer
TER	Translation Edit Rate

WMT	Workshop on Machine Translation
XLM	Cross-lingual Language Model
Zmorge	Zurich Morphological Analyzer for German

1 Introduction

In recent years, the field of neural machine translation (NMT) has seen considerable progress, which significantly improved the translation quality achieved by machine translation (MT) systems.

To detect and quantify the improvement, the translation quality needs to be measured with objective criteria. This task is complex, as there is not just one correct translation option. Rather, for a given source sentence, many different valid translations exist. Human judgement is commonly regarded as gold-standard. However, human quality scores are expensive and time-consuming to collect, and they are to some degree subjective. As a faster and cheaper alternative, various automatic evaluation metrics have been proposed to assess translation quality and guide decisions during the development of an MT system. These automatic measures are usually based on the comparison of the MT hypothesis with a human reference translation and sometimes with the source. Ideally, such an evaluation metric is able to not only reward translation hypotheses that coincide well with the surface form of the reference, but also hypotheses that use synonyms, a different but correct word order or that are paraphrases of the reference. In addition, a good metric should punish the use of antonyms and wrong polarity.

Traditional evaluation metrics, that measure the lexical overlap between the hypothesis and the reference (e.g., BLEU (Papineni et al., 2002), ChrF (Popović, 2015)), are incapable of capturing such nuances. In consequence, with the continuous improvements of MT systems, they increasingly fall short in assessing their quality correctly (Ma et al., 2019; Mathur et al., 2020; Kocmi et al., 2021; Freitag et al., 2021a, 2022b). Hence, the advancements in NMT called for new, more elaborate evaluation metrics.

In response to this need, a novel approach of deploying a trained neural network as a metric has been proposed. These neural metrics are based on large pre-trained language models, such as BERT (Devlin et al., 2019) or XLM-RoBERTa (Conneau et al., 2020), that are fine-tuned in a regression task on human quality scores. They leverage the pre-trained language model to encode the hypothesis, the reference and/or the source and to map the embeddings of the two or three segments being compared into the same shared feature space. This allows them to evaluate the *semantic similarities* between the embeddings and to assign a quality score to the hypothesis accordingly. Hence, they are able to recognize and reward synonyms and paraphrases with different word choice, word order and length (Kocmi et al., 2021; Freitag et al., 2022a), even though they still rely to some degree on surface-level overlap with the reference (Amrhein et al., 2022). Nonetheless, top-performing neural metrics such as BLEURT (Sellam et al., 2020a) and COMET (Rei et al., 2020a) correlate well with human judgements and are considered state-of-the-art (Kocmi et al., 2021; Freitag et al., 2022b).

The recent progress in evaluation metrics led to the exploration of alternative decoding strategies, such as Minimum Bayes Risk (MBR) decoding, that – in contrast to the commonly used beam search decoding – takes advantage of the latest improvements in MT evaluation metrics.

In sampling-based MBR decoding, unbiased samples are drawn from the probability distribution of the MT model. These samples are used as candidate pool and are compared against each other to find the *consensus translation* (Eikema and Aziz, 2020). A utility function assesses the similarity between the candidates. Typically, an evaluation metric serves as utility function as they are designed to measure the similarity between two segments. By incorporating powerful metrics in the decoding procedure, MBR decoding leverages the information provided by the metric and has the potential to further improve translation quality (Fernandes et al., 2022).

An additional advantage of MBR decoding is that it offers the possibility to shed light on blind spots of the metric used as utility function. As neural metrics like COMET and BLEURT are becoming more widespread, it is crucial to scrutinise their strengths and particularly their weaknesses. As they guide researchers and developers in deciding which model is best and which directions of research are most promising, adopting them without knowing their downsides entails the risk of biased decision-making and of optimising towards blind spots of the metric.

A thorough examination of novel metrics is important for several reasons. First, neural metrics are built on top of large language models that are pre-trained on huge amounts of natural language data and learn any kind of bias present in the data sets, including cultural biases regarding gender, race and religion (Chan et al., 2020; Kocmi et al., 2021). These biases can be reflected in the metric scores (Amrhein et al., 2022). Second, trained neural metrics are black boxes that do not explain why they assign a certain score to a translation. Sometimes, they fail unpredictably attributing high scores to bad translations (Amrhein and Sennrich, 2022).

Biases introduced by black box neural metrics are often subtle and more difficult to uncover than biases of lexical overlap-based metrics. Amrhein and Sennrich (2022) propose to use sampling-based MBR decoding to reveal blind spots of the metric used as utility function. Neural evaluation metrics are usually trained on beam search outputs. In contrast, the unbiased samples in MBR decoding are generally of a lower quality and contain different kinds of errors that are less frequent in beam search outputs. When confronted with such unusual errors, neural metrics may show an unexpected behaviour that hints at their weaknesses.

With this method, Amrhein and Sennrich (2022) demonstrated that COMET-20 is not sensitive enough towards errors in named entities (NEs) and numbers. Furthermore, they notice that COMET-20 tends to choose hypotheses with nonsensical German compounds and polarity errors. However, their evidence for these potential weaknesses is only anecdotal, as it was not the focus of their work.

1.1 Research Questions

In this thesis, I extend the work by Amrhein and Sennrich (2022) examining the sensitivity of neural metrics towards German compounds. I will provide a systematic investigation of whether German compounds are indeed another blind spot of neural metrics or whether the appearance of nonsensical compounds in MBR outputs must be attributed to other factors, e.g. the lower quality of the samples in the candidate pool.

For my experiments, I will focus on translations from English to German for several reasons. Firstly, in German, the formation of compounds is a productive morphological process that frequently serves to denote new objects or concepts, e.g. *Coronavirus-Richtlinien* (coronavirus guidelines). Secondly, I speak these two languages fluently which is necessary to judge whether a German compound is a valid translation of the English source text. Thirdly, working with one of the language pairs studied by Amrhein and Sennrich (2022) allows for a comparison of my results with their work. To further ensure that my work is comparable to theirs, I will use the same data sets, MT model and COMET model as they did.[1]

In the first part of this thesis, I will examine whether and to what extent German compounds are a blind spot of COMET. To this end, I will use COMET-20 as utility function in MBR decoding. The obtained translations are then compared to beam search outputs on the one hand, and to MBR-decoded translations obtained with ChrF and ChrF++ as utility functions on the other hand. The examination reveals that the MBR-decoded translations obtained with COMET-20 as utility function contain a considerably larger amount of nonsensical, mistranslated compounds than translation generated with ChrF and ChrF++ as utility function or via beam search decoding. This is a clear indication that German compounds are indeed a weakness of COMET-20.

Having identified this blind spot, the question arises whether it is possible to increase COMET's sensitivity towards the incorrect formation of German compounds. Amrhein and Sennrich (2022) showed that retraining COMET-20 on synthetic data is not sufficient to erase its blind spots. Therefore, I will experiment with exchanging one of the building blocks of COMET: the underlying pre-trained language model. COMET-20 is built on the multilingual XLM-RoBERTa that is pre-trained with the Masked Language Modelling objective (Devlin et al., 2019). Like most language models, XLM-RoBERTa uses Sub-

[1] The materials and methods are described in Section 3.1.

Word Masking (SWM) in this pre-training task.[2] Hence, if a word consists of multiple sub-words, only one sub-word is masked, while the others are visible to the model making the prediction of the masked token considerably easier.

German compounds are complex words that consist of at least two constituents, which are most likely split up into two or more sub-words during pre-training. It is therefore possible that models pre-trained with SWM concentrate on individual sub-words only, neglecting the relationship between them.

One way to force the model to focus on an *entire* word, is by using Whole Word Masking (WWM) in pre-training (Devlin et al., 2019).[3] In WWM, all sub-words of a word are masked simultaneously which makes the task to reconstruct the word more challenging. The model is forced to predict *all* parts of a word and most likely gains a better notion of what a word is and which sub-words belong together. Hence, it is reasonable to assume that models pre-trained with WWM are better at dealing with German compounds.

To investigate the effect of WWM on the translation of compounds, I will train new evaluation models and use them as utility function in MBR decoding. I use GBERT (Chan et al., 2020) as the underlying language model. GBERT is a monolingual German BERT model that is available in two variants: One is pre-trained with Sub-Word Masking (GBERT$_{\mathrm{SWM}}$), the other one with Whole-Word Masking (GBERT$_{\mathrm{WWM}}$).

On top of GBERT, I will add a regression layer and train it to predict a quality score for a hypothesis given a reference. For each of the two variants of GBERT, I train two different metric models: One that follows the slightly modified training procedure of COMET-QE (Rei et al., 2021), the other one follows the procedure of BLEURT (Sellam et al., 2020a). I call the resulting metrics GCOMET$_{\mathrm{SWM}}$, GCOMET$_{\mathrm{WWM}}$, GBLEURT$_{\mathrm{SWM}}$ and GBLEURT$_{\mathrm{WWM}}$.

In contrast to GCOMET and GBLEURT, the original COMET-20 and BLEURT-20 metrics rely on multilingual language models. To assess the effect of multilinguality on the translation of compounds, I additionally train COMET$_{\mathrm{Contrastive}}$. It is trained in exactly the same way as GCOMET, but with the multilingual XLM-RoBERTa$_{\mathrm{Base}}$ pre-trained with SWM as the underlying language model.

Regarding the effect of multilinguality, the hypothesis is less clear than in the case of WWM. On the one hand, multilinguality often improves performance on various tasks especially for low-resource languages, as the model benefits from cross-lingual transfer (Conneau and Lample, 2019; Sellam et al., 2020b). On the other hand, the model capacity is divided between several languages, which can have a negative impact on the model's performance. This phenomenon is known as the *curse of multilinguality* (Conneau et al., 2020; Pu et al., 2021). As German is a high-resource language, I expect that a multilingual model deteriorates the quality of German translations, particularly that of compounds.

[2] Sub-Word Masking and Whole Word Masking are explained in Section 2.4.

[3] See github.com/google-research/bert/commit/0fce551

To confirm or reject the hypotheses regarding the effect of WWM and multilinguality, I will conduct a series of experiments using the newly trained metrics as utility function in MBR decoding. Firstly, the overall quality of the MBR-decoded translations will be measured in terms of various automatic evaluation metrics. Secondly, I will analyse which utility function generates the most nonsensical compounds in its translations. Thirdly, a targeted sensitivity analysis as proposed by Amrhein and Sennrich (2022) is conducted to quantify how sensitive the different metric models are towards targeted changes in German compounds.

Additionally, the data set created by Amrhein and Sennrich (2022) is used to assess the sensitivity of the new metrics towards common nouns, named entities and numbers. This analysis sheds light on the question of whether a metric based on a monolingual language model pre-trained with WWM is able to remove the blind spots identified in COMET-20 or whether these persist.

On a side note, two other questions that arose during the conductance of the experiments will be answered. The first one concerns the implementation of MBR decoding. When the candidate and the support sets are identical, each candidate also occurs in the support set. When comparing a candidate against the support set, we can either include the comparison of the candidate to itself or exclude this comparison. This thesis will investigate whether including or excluding the comparison of two identical segments affects the metric's choice of the best translation and hence the MBR output.

The second question concerns the issue of "metric overfitting" (Fernandes et al., 2022). When we optimise towards a neural metric used as utility function in MBR decoding and then apply it to evaluate the translation quality of the generated outputs, it does no longer produce reliable scores. Rather, it rates translations generated with itself as utility function overly optimistically. The automatic evaluation conducted in 5.1 reveals a strong overfitting effect for COMET-20. Hence, I will investigate 1) whether this effect is similarly strong for $\text{GCOMET}_{\text{WWM}}$ and 2) which factors contribute the most to the observed overfitting.

In addition, I will experiment with a novel approach. As neural metrics have certain blind spots, they sometimes unpredictably assign high scores to bad translations. However, different metrics have distinct blind spots. To overcome the deficiencies of a given metric, I will *combine* two metrics in MBR decoding. Hence, a candidate is only chosen as best translation if it receives a high score from both metrics. I will investigate whether the combination of metrics 1) will improve translation quality and 2) alleviate the observed overfitting effect.

The code and other materials used in this thesis are publicly released on GitHub.[4] The newly trained metric models can be downloaded from Google Drive.[5]

[4] github.com/sarahkiener/compound-sensitivity

[5] drive.google.com/compound-sensitivity_models

1.2 Thesis Structure

After the research questions have been presented, Chapter 2 introduces the concept of MBR decoding, the different types of MT evaluation metrics, their respective strengths and weaknesses and elucidates the relationship between the evaluation metrics and MBR decoding. Finally, the chapter explains the differences between Sub-Word Masking and Whole Word Masking. In Chapter 3, I will show that German compounds are a blind spot of COMET-20. Chapter 4 explains how the novel metrics, GCOMET and GBLEURT were trained providing a description of their building blocks and architecture. Further, it outlines the experiments that were conducted to investigate whether monolingual models and WWM can improve the translations of German compounds. Chapter 5 presents the results of these experiments. These results are discussed and interpreted in Chapter 6. Finally, Chapter 7 summarises the most important contributions of this thesis and provides an outlook on future work.

2 Related Work

Recent work has proposed Minimum Bayes Risk (MBR) decoding as an alternative decision rule to beam search decoding in neural machine translation (NMT). MBR decoding relies on a utility function, which is typically an evaluation metric that measures the similarity between a candidate sentence and a (pseudo-)reference. This chapter discusses both the current state of research in the field of MBR decoding as well as the progress that has been made regarding machine translation (MT) evaluation metrics. In addition, the chapter presents studies experimenting with Whole Word Masking as a pre-training technique for large language models.

2.1 Minimum Bayes Risk Decoding

Modern NMT systems commonly rely on beam search - a tractable approximation to the maximum a posteriori (MAP) decision rule - as decoding algorithm (Eikema and Aziz, 2020; Müller and Sennrich, 2021). MAP decoding seeks to find among the set of all possible hypotheses the translation hypothesis that is most probable under a given model. Hence, MAP decoding aims at identifying the *mode* of the distribution that the NMT model has learnt. Since considering every possible translation hypothesis is not feasible in practice, approximations such as beam search (Graves, 2012; Sutskever et al., 2014) are needed (Eikema and Aziz, 2020; Müller and Sennrich, 2021; Amrhein and Sennrich, 2022; Fernandes et al., 2022). However, various studies (Ranzato et al., 2015; Sountsov and Sarawagi, 2016; Koehn and Knowles, 2017; Lee et al., 2019; Ott et al., 2018; Khayrallah and Koehn, 2018; Stahlberg and Byrne, 2019; Kumar and Sarawagi, 2019; Müller et al., 2020) report serious pathologies of the beam search algorithm such as: *the beam search curse*, i.e. translation quality decreases with a larger beam size, *length bias*, i.e. the true length of translations is often underestimated, *skewed word frequencies*, i.e. frequent tokens in the training data are found disproportionately often in the model outputs while rare tokens are underrepresented, *copying of input sentences*, i.e. if source copies appear on the target side of the training data, copy hypotheses are overrepresented in the beam search output, *low domain robustness*, i.e. the model tends to produce hallucinations under domain shift, *exposure bias*, i.e. the mismatch between training and test time arising from training the model on gold-standard reference sentences that are not available at test time where the model suddenly has to deal with its own predictions, and *non-admissible heuristic search bias*, i.e. candidates are pruned based on their current score without regard to their future score (cf. Eikema and Aziz, 2020; Müller and Sennrich, 2021). Moreover, beam search outputs are usually ranked below human references in professional evaluations (Freitag et al., 2021a) while the NMT model itself rates its outputs as more likely than human translations (Ott et al., 2018). Hence, the estimated probability and the actual translation quality do not always correlate (Freitag et al., 2022a).

Instead of blaming the NMT model or the training algorithm for those biases, Eikema and Aziz (2020) attribute these deficiencies to the *inadequacy of the mode*. They argue that in NMT the search space is so large that the most probable translations together account for only very little probability mass. Due to this flat probability distribution, the model's choice of the best translation, i.e. the mode, is somewhat arbitrary and the beam search output is a relatively rare outcome under the model. This is in line with Stahlberg and Byrne (2019), who find that in many cases the mode is the empty sequence. Hence, Eikema and Aziz (2020) conclude that MAP decoding relying heavily on the mode might be the source of many of the observed biases while the distribution of the model might represent well the properties of the training data in terms of length, lexical and word order statistics. By analysing unbiased samples from the model obtained via ancestral sampling, they demonstrate that this is indeed the case.

To exploit the available information about the model's learnt distribution and to make an informed decision in identifying the best translation among the set of unbiased samples, Eikema and Aziz (2020) suggest to use Minimum Bayes Risk (MBR) decoding as a decision rule, a concept from statistical decision theory which takes into account the model distribution holistically. MBR decoding has already been deployed successfully in statistical machine translation (Goel and Byrne, 2000; Kumar and Byrne, 2004; Tromble et al., 2008) and was used in NMT in combination with beam search decoding (Stahlberg et al., 2017; Shu and Nakayama, 2017). It is well-suited if we trust a model in expectation but not its mode in particular.

The goal of MBR decoding is to choose from the entire set of possible translations H the translation candidate y^* that minimizes the expected cost (risk), given a source sentence x, the true probability distribution P and a loss function L that compares the translation candidate h_i to the true translation h_j. Amrhein and Sennrich (2022) formalize the problem as:

$$y^* = \underset{h_i \in H}{\operatorname{argmin}} \sum_{h_j \in H} P(h_j|x) L(h_i, h_j) \tag{2.1}$$

However, the true probability distribution is unknown and it is unfeasible to sum over the entire set of possible hypotheses. Therefore, in practice, the problem is approximated based on the probability distribution of the translation model and a subset of hypotheses. Eikema and Aziz (2020) suggest to draw a candidate set C of unbiased hypotheses from the model distribution via ancestral sampling. With the same method, they create a support set S that can either be identical to the candidate set C or contain different samples. Then, they apply a utility function $u(h_i, h_j)$ that assesses each hypothesis h_i from the candidate set C against each pseudo-reference h_j from the support set S and picks the candidate that maximises the expected utility as the optimum decision y^*. Following this procedure, Amrhein and Sennrich (2022) reformulate the problem as:

$$y^* = \underset{h_i \in C}{\operatorname{argmax}} \frac{1}{|S|} \sum_{h_j \in S} u(h_i, h_j) \tag{2.2}$$

The sizes of the candidate and support sets as well as the utility function u are hyperparameters of the algorithm (Müller and Sennrich, 2021). In MT, the utility function is typically a measure of similarity. In principle, any automatic MT evaluation metric can serve as utility function in MBR decoding (cf. 2.2). The advantage of this method over MAP decoding is that, even though the translation distribution might be very flat, the expected utility for the different candidates might be quite distinct and candidates that are very dissimilar to others (e.g. the empty string, hallucinations) are ruled out. Thus, MBR decoding can be viewed as a method to find the *consensus translation* that is closest on average to all likely candidates and that is able to identify the candidates that share statistics with the reference translations (Müller and Sennrich, 2021; Eikema and Aziz, 2022; Kumar and Byrne, 2004).

Various studies (Eikema and Aziz, 2020; Müller and Sennrich, 2021; Freitag et al., 2022a; Eikema and Aziz, 2022; Fernandes et al., 2022) could show that MBR decoding indeed leverages this potential and is able to perform on par with or even outperform beam search decoding in terms of both state-of-the-art automatic metrics and human assessment. Furthermore, sampling-based MBR decoding overcomes some of the drawbacks of beam search, because sampling allows for more exploration of the distribution and a higher diversity in the candidate set than beam search decoding (Fernandes et al., 2022). Most importantly, sampling-based MBR decoding does not suffer from an equivalent of the beam search curse (Eikema and Aziz, 2022) and it is more robust to domain shift, hallucinations and copy noise (Müller and Sennrich, 2021). In addition, it alleviates, but does not entirely resolve the token frequency bias (Müller and Sennrich, 2021).

Despite those improvements over beam search, MBR decoding also exhibits a few drawbacks. A major downside is its inefficiency compared to beam search. The cost of assessing the utility function grows quadratically with the number of samples in the candidate and support sets. Various strategies to overcome this limitation were proposed (Eikema and Aziz, 2022; Fernandes et al., 2022; Freitag et al., 2022a; Amrhein and Sennrich, 2022). Moreover, Müller and Sennrich (2021) show that MBR decoding cannot mitigate the length bias encountered frequently in beam search outputs. Rather, it inherits the length bias associated with its utility function.

Nonetheless, MBR decoding is a valuable alternative to beam search that mitigates several beam search biases and thus has the potential to further improve MT quality. In addition, it leverages the progress in evaluation metrics by incorporating them as utility function. Hence, these metrics are crucial to the performance of MBR decoding.

2.2 Automatic Evaluation Metrics for Machine Translation

As described in the previous section, MBR decoding requires a utility function that assesses the similarity between a candidate translation and a (pseudo-)reference. Such similarity metrics have been proposed long before the advance of MBR as a decoding strategy in NMT. As a fast and cheap proxy for human assessment scores (Sellam et al., 2020a), they evaluate the translation quality of sentences produced by an MT system. Most of these metrics are reference-based. They compare the MT output to the reference translation and assign a quality score to it. This score does not only allow to judge the overall quality of a system, but also enables the comparison between the quality of two MT systems (Kocmi et al., 2021).

During decades, lexical overlap-based metrics that measure the surface similarity between a hypothesis and a reference were deployed by default (Marie et al., 2021). However, as machine translations became more and more sophisticated, these metrics increasingly fell short in assessing their quality correctly (Ma et al., 2019; Mathur et al., 2020; Kocmi et al., 2021; Freitag et al., 2021a, 2022b). In consequence, the field has seen significant progress in recent years. Neural metrics have been proposed that rely on pre-trained word embeddings in order to capture the semantic similarity between the hypothesis, the reference and, in some multilingual metrics, the source. While the first generation of neural metrics compares the embeddings of the candidate and the reference in the embedding space, the newest metrics are themselves neural networks trained on human quality assessment data (Freitag et al., 2022a). While most recent approaches (Kocmi and Federmann, 2023) experiment with leveraging large language models, such as GPT (Radford et al., 2018; Brown et al., 2020), to score the quality of a translation hypothesis, this thesis focuses on metrics that have been specifically designed for the MT evaluation task. The majority of these MT metrics can be categorized into four groups that are described in more detail in the following subsections.[1]

2.2.1 Metrics Based on Lexical Overlap

Metrics that rely on lexical overlap compare the surface form of the hypothesis to the surface form of the reference, usually in terms of word or character n-grams. These metrics are cheap to compute and can be applied to any language which made them a popular choice for MT evaluations for many years. Further, their performance is predictable and allows us to determine which substring has the largest impact on the score (Kocmi et al., 2021). On the downside, string-based metrics are heavily dependent on the quality of the reference translation as they completely ignore the source sentence. Moreover, they are unable to recognize paraphrases and do not take the severity of errors into account (Kocmi

[1] Other metrics exist that do not belong into one of the four groups discussed here. One example is PRISM (Thompson and Post, 2020) that deploys a sequence-to-sequence paraphraser, that is itself a trained multilingual NMT model, conditioned on the reference to score an MT hypothesis.

et al., 2021; Fernandes et al., 2022; Freitag et al., 2021a,b).

By far the most commonly used metric during the last two decades (Marie et al., 2021) is the Bilingual Evaluation Understudy (BLEU) developed by Papineni et al. (2002). BLEU measures lexical overlap as the precision of n-gram matches between the hypothesis and the reference with $n \leq 4$. To punish translations that are too short, a brevity penalty is added. Recent studies demonstrate that BLEU coincides poorly with human quality judgements and is unable to differentiate correctly between high performing MT systems. It should therefore be deprecated as the standard metric (Ma et al., 2018, 2019; Mathur et al., 2020; Kocmi et al., 2021; Freitag et al., 2021b, 2022b).

Another popular string-based metric is ChrF and its variants (Popović, 2015). Instead of word n-grams, ChrF uses the character n-gram F-score to assess the candidate translation against the reference. Its variant ChrF++ additionally includes word unigrams and bigrams (Popović, 2017). While suffering from the same drawbacks as other string-based metrics, ChrF has a clearly higher correlation with human quality judgements than BLEU (Ma et al., 2018, 2019; Mathur et al., 2020).Therefore, Kocmi et al. (2021) recommend ChrF as standard evaluation metric along with COMET.

Many other string-based evaluation metrics have been proposed, among them ROUGE (Lin, 2004), METEOR (Lavie and Denkowski, 2009), TER (Snover et al., 2006), CharacTER (Wang et al., 2016) and BEER (Stanojević and Sima'an, 2015), but it is beyond the scope of this work to discuss them in detail.

2.2.2 Embedding-Based Metrics

With the advance of neural networks, metrics have been developed that exploit the new possibilities. They make use of pre-trained word and sentence representations and compare the embeddings of the hypothesis with those of the reference. Examples of such metrics are BERTScore (Zhang et al., 2020) and BERTr (Mathur et al., 2019), that both leverage contextual BERT embeddings (Devlin et al., 2019) to calculate a score based on the cosine similarity between the hypothesis and the reference, and YiSi-1 (Lo, 2019) that aggregates inverse-document-frequency-weighted lexical semantic similarities based on pre-trained embeddings. Metrics based on embedding-overlap are robust and perform well even with little or no training data (Sellam et al., 2020a). They generally correlate better with human judgements than string-based metrics, but are outperformed by trained neural metrics (Ma et al., 2019; Mathur et al., 2020), that, in addition to semantic similarities, exploit information present in human quality scores like DA or MQM (Rei et al., 2020a).

2.2.3 Trained Neural Metrics

Most recent metrics are trained regression models based on pre-trained contextual embeddings and fine-tuned on human quality evaluation scores. They are able to recognize synonyms and paraphrases with different sentence structures, word choices and length compared to the reference translation. Hence, they are less dependent on the translation quality of the reference and have the potential to rate hypotheses with infrequent but correct tokens higher than string-based metrics as they are not forced to match surface n-grams in the reference (Kocmi et al., 2021; Freitag et al., 2022a). Nevertheless, they still considerably rely on surface-overlap with the reference and hence the translation quality of the reference still has an impact on the metric's performance (Amrhein et al., 2022; Rei et al., 2020b). Trained metrics can be adapted to score task-specific properties like fluency, faithfulness, grammar and style (Sellam et al., 2020a). Nonetheless, those metrics have certain drawbacks (Kocmi et al., 2021; Amrhein and Sennrich, 2022). They depend on their training data which can be a source of biases introduced into the model. Additional problems for the quality evaluation arise if an MT system is trained on the same data or incorporates the same pre-trained model as the evaluation metric does. Moreover, trained evaluation metrics are black boxes that do not always behave as expected. For example, they could exhibit difficulties with a specific domain or prefer fluency over adequacy. Finding the source of unexpected behaviour is difficult in such metrics. Finally, trained metrics can only be applied to a limited set of languages and their performance can differ for each language. Nevertheless, they are currently the state-of-the-art in the field (Kocmi et al., 2021; Freitag et al., 2022a).

Early examples of such metrics are RUSE (Shimanaka et al., 2018) that uses pre-trained embeddings, and ESIM (Chen et al., 2017; Mathur et al., 2019) that learns the sentence representations from scratch. Both are then fine-tuned on human judgements of the WMT Shared Tasks.

Currently, COMET (Rei et al., 2020a) is considered the state-of-the-art evaluation metric (Kocmi et al., 2021). The COMET-family comprises various neural metrics, including models trained on Direct Assessment data, Multidimensional Quality Metrics (MQM) and a reference-free quality estimation model. This thesis focuses on the "standard" COMET-20 metric as this is the metric scrutinized by Amrhein and Sennrich (2022), on whose work this thesis builds. COMET builds on the multilingual XLM-RoBERTa model (Conneau et al., 2020) to encode the candidate and the reference along with the source sentence in the same cross-lingual space. It is then fine-tuned in a regression task on Direct Assessment data from the WMT Metrics Shared Task (Rei et al., 2020a). For this thesis, I worked with `wmt20-comet-da` (cf. 4.1.3), which was the recommended default model during the past three years. Only recently, the new default model `wmt22-comet-da` was released (Rei et al., 2022) which is able to predict error spans annotating the words with `OK` or `BAD` tags.

BLEURT (cf. 4.1.2) is ranked among the top-performing metrics as well. The first English-only version of BLEURT (Sellam et al., 2020a) was based on BERT (Devlin et al., 2019), the current default model BLEURT-20 (Pu et al., 2021) uses multilingual RemBERT (Chung et al., 2021) embeddings. BLEURT is a fine-tuned regression model trained in a three-step procedure. After the regular BERT pre-training, it is further pre-trained on synthetic, randomly perturbed sentence pairs. Finally, it is fine-tuned on DA scores from the WMT Metrics Shared Tasks. In contrast to COMET, BLEURT encodes the MT hypothesis and the reference together as a *sentence pair* ignoring the source sentence.

2.2.4 Reference-Free Quality Estimation Metrics

As an alternative to the reference-based evaluation metrics, several reference-free quality estimation (QE) metrics have been developed over the past years. These metrics calculate the similarity across different languages by comparing the source sentence to the candidate translation. Their main advantages are that they do not require a reference translation to score a hypothesis and are hence not affected by a potential reference-bias (Mathur et al., 2020). Due to the increased difficulty of the quality estimation task, QE metrics usually exhibit a lower correlation to human judgements than reference-based metrics. Nonetheless, their correlation to human scores is constantly growing and the top-performing systems such as COMET-QE (Rei et al., 2021) and OpenKiwi-MQM (Kepler et al., 2019; Rei et al., 2021) already achieve competitive results (Freitag et al., 2021b; Specia et al., 2021). This indicates that references do not provide as much valuable information as previously thought (Rei et al., 2021; Kocmi et al., 2021). Future work might well dispense with references that are expensive and time-consuming to collect (Rei et al., 2021).

2.3 Uncovering Blind Spots in MT Metrics Through MBR Decoding

The previous sections have shown that on the one hand, MBR decoding has the potential to overcome the deficiencies of beam search decoding. On the other hand, by incorporating the latest evaluation metrics into the decoding procedure, MBR can be considered quality-aware decoding that takes the recent tremendous progress in MT metrics into account (Fernandes et al., 2022), while beam search entirely ignores these advancements. Using more powerful metrics that coincide better with human judgement in the decoding procedure potentially leads to improved translation quality (Fernandes et al., 2022; Freitag et al., 2022a).

Freitag et al. (2022a) demonstrated that this is indeed the case. Translations obtained via MBR decoding with the powerful neural BLEURT metric as utility function receive higher scores from human annotators than translations produced with lexical metrics or by beam search. The latter two show a strong preference for most likely tokens in order to increase the chance of matching the surface form of the reference. In contrast, MBR decoding with BLEURT chooses hypotheses that are not among the highest log likelihood candidates according to the NMT model, but that are regarded as more appropriate by human annotators.

Nonetheless, several pitfalls have been identified for MBR decoding with neural metrics. First, MBR decoding relies heavily on the quality of the candidate set. Thus, a good NMT model is needed to produce high-quality candidates and to enable MBR to outperform beam search (Freitag et al., 2022a). Second, Fernandes et al. (2022) mention the issue of "metric overfitting". When explicitly optimizing for a fine-tuned metric in MBR decoding, this metric is no longer a reliable indicator for the final evaluation of translation quality. Typically, a fine-tuned metric scores translations produced by MBR using the same metric as utility function overly optimistically. By doing so, their correlation with human judgements drops. This behaviour points to a potential overfitting effect. Instead of improving translation quality, the systems learn to exploit pathologies inherent in a certain metric (Fernandes et al., 2022).

To uncover such blind spots of neural evaluation metrics, Amrhein and Sennrich (2022) propose a novel approach. They suggest to use the metric under study as utility function in sampling-based MBR decoding to gain insights into its weaknesses. In this setup, the hypotheses in the candidate and support pool are usually of lower quality than the beam search outputs, on which the metric model was trained. Hence, when scoring the candidates, the metric is confronted with error types that are rarely encountered in beam search sentences. These infrequent error types might trigger the metric to behave unexpectedly and to assign high scores to bad translations. By inspecting the final MBR-decoded output, i.e. the hypotheses chosen by the metric, Amrhein and Sennrich (2022) identify linguistic phenomena that are especially challenging for the metric to judge correctly. Having found a potential weakness in a metric, they compare the frequency of this error type in the MBR outputs to the frequency of the same error in beam search translations and in MBR outputs generated with a different utility function. This comparison reveals whether the identified error type is indeed a blind spot of the metric under study, or whether it poses a challenge to MT models and evaluation metrics in general.

Applying this procedure, Amrhein and Sennrich (2022) show that COMET-20 is not sensitive enough to errors in named entities and numbers. To investigate these failures of COMET-20 more systematically, they propose an MBR-based sensitivity analysis. Keeping the support set constant, they apply targeted changes to the translation hypotheses and include them in the candidate pool along with the correct reference. In the course of MBR decoding, a reliable metric is expected to assign a considerably higher score to the correct translations than to the perturbed candidates. By measuring the difference between the scores assigned to the correct and the perturbed candidates and averaging them across all examples of one perturbation type, Amrhein and Sennrich (2022) demonstrate that COMET-20 tends to ignore errors in numbers and named entities rating them only slightly lower than the correct references. These biases seem to be inherent to COMET-20, as they cannot be removed by a simple retraining on synthetic data.

This thesis relies on the described approaches by Amrhein and Sennrich (2022) to investigate the sensitivity of COMET-20 and the newly trained neural metrics towards German compounds.

2.4 Whole Word Masking in Language Model Pre-Training

Since the introduction of BERT in 2019 (Devlin et al., 2019), large language models have become popular in the field of Natural Language Processing (NLP). These models are typically pre-trained on a massive corpus of unlabelled monolingual data. Later, they are fine-tuned on a labelled data set for a specific downstream task.

When training BERT, Devlin et al. (2019) introduced a new pre-training objective to learn from unlabelled data: Masked Language Modelling (MLM). This objective was later adopted by many variants of BERT as well as by other language models such as XLM (Conneau and Lample, 2019; Conneau et al., 2020; Liu et al., 2020b; Chan et al., 2020)

In the Masked Language Modelling task, 15 % of the tokens in the input sequence are randomly masked by replacing them with the special [MASK] token. The model is then asked to reconstruct these masked tokens. This strategy allows to train *bidirectional* representations. Without masking, the bidirectional conditioning would leak the information on the other words in the sequence making it very easy for the model to reconstruct any token. On the downside, this training strategy restricts the learning to the 15 % of masked tokens (Chan et al., 2020). Another drawback is that the special [MASK] token introduces a discrepancy between pre-training and fine-tuning where the [MASK] token does not occur. To alleviate this problem, Devlin et al. (2019) opt for three different "masking" strategies. Of the 15 % of input tokens chosen for masking:

1. 80 % are replaced with the [MASK] token,
2. 10 % are replaced with a random token,
3. 10 % remain unchanged and are not replaced.

To better deal with rare words, BERT splits the input words into sub-word units (Sennrich et al., 2016) using WordPiece embeddings (Wu et al., 2016). These sub-words (or tokens) form the basis on which BERT operates. The original BERT model, was pre-trained with **Sub-Word Masking (SWM)**, i.e. masking was applied on sub-word level. As the tokens to be masked are selected randomly, it is possible that only one sub-word of a word is masked while the other sub-words of that word are visible to the model. The following example illustrates how words in the input sequence are split up into several sub-words, marked with ##. The line **SWM** shows a possible outcome of randomly masking some sub-word tokens:

Input:	The	cat	likes	play	##ing	with	the	ball	of	wool	.
SWM:	[MASK]	cat	likes	play	[MASK]	with	the	ball	of	[MASK]	.

The example illustrates that masking only some tokens of a word reduces the complexity of the prediction task. Predicting the second [MASK] is very easy, as the beginning of that word is visible along with the other words in the sequence. In contrast, the prediction of the third [MASK] is more difficult: Is it a ball of *wool*, *yarn* or even *light*? The first

[MASK] is also easy to reconstruct, as the definite article is often encountered as the first word of an English sentence. Here, the simplicity of the prediction task is not owed to the masking strategy, but to the frequency, with which the masked token occurs.

After the release of the first BERT model, its authors noticed that Sub-Word Masking tends to oversimplify the prediction task. They propose to use **Whole Word Masking (WWM)** instead.[2] With WWM, *all* the tokens of a word are masked at once making it more difficult for the model to reconstruct the word, while everything else remains constant, i.e. the overall masking rate is the same as in SWM and each token is still predicted independently. Hence, the example above is masked as follows with WWM:

Input:	The	cat	likes	play	##ing	with	the	ball	of	wool	.
WWM:	The	cat	likes	[MASK]	[MASK]	with	the	ball	of	[MASK]	.

Figure 1 further illustrates the differences between the two masking strategies and clarifies how the language model predicts the masked tokens.

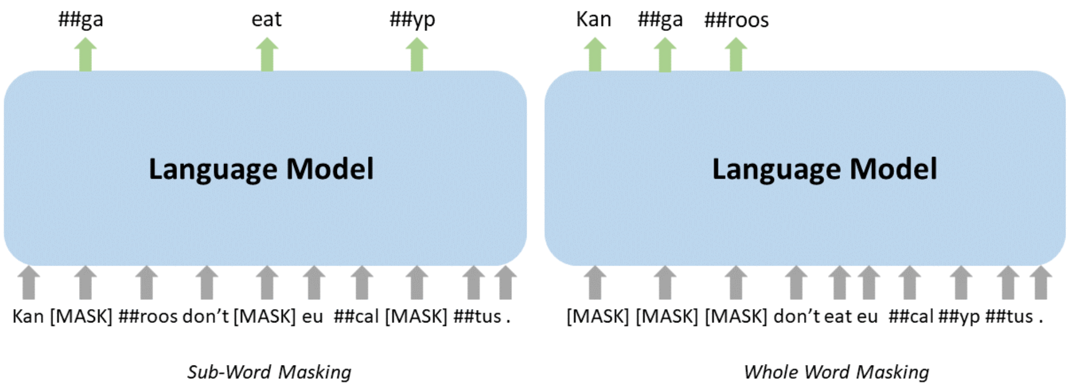
##ga
eat
##yp
Kan
##ga
##roos
Language Model
Language Model
Kan [MASK] ##roos don't [MASK] eu ##cal [MASK] ##tus .
[MASK] [MASK] [MASK] don't eat eu ##cal ##yp ##tus .
Sub-Word Masking
Whole Word Masking

Figure 1: Sub-Word Masking and Whole Word Masking for the sentence *Kangaroos don't eat eucalyptus.*

The authors of BERT conclude that WWM is beneficial to Question Answering and Natural Language Inference.[3] Despite this finding, most variants of BERT, such as RoBERTa (Liu et al., 2020b) and RemBERT (Chung et al., 2021) among many others, and language models complementing the training objectives of BERT, like ELECTRA (Clark et al., 2020), still use Sub-Word Masking or Substitution respectively in their training procedure.

So far, WWM has mainly been investigated for Chinese language models. Unlike English subwords that can be further split up into single characters, a Chinese word consists of several characters that are atomic units. Therefore, Whole Word Masking is needed to dissolve the association between characters and make the model learn something mean-

[2] See github.com/google-research/bert/commit/0fce551

[3] See https://github.com/google-research/bert

ingful (Dai et al., 2022). WWM was shown to have a beneficial effect on a wide range of Chinese NLP tasks (Liu et al., 2020a; Dai et al., 2022; Zhou et al., 2020; Cui et al., 2021).

Concerning German, to the best of my knowledge, only Chan et al. (2020) studied the effect of Whole Word Masking in language model pre-training. They train three German BERT (GBERT) models, each of the size of $\text{BERT}_{\text{Base}}$. Two of them are of particular interest to this thesis. The first one is pre-trained with Sub-Word Masking ($\text{GBERT}_{\text{SWM}}$), while the second one uses Whole Word Masking ($\text{GBERT}_{\text{WWM}}$)[4] in pre-training. Everything else (training data, hyperparameters etc.) is identical for both models. This allows to assess the effect of the different masking strategies on the performance of the language model.

Evaluating the models on a coarse- and fine-grained hate speech classification task and on Named Entity Recognition, Chan et al. (2020) conclude that WWM has a beneficial effect. When compared only to models of the same size, $\text{GBERT}_{\text{WWM}}$ even reaches a new state-of-the-art for NER and the coarse hate speech classification task. Hence, Whole Word Masking is a promising approach for German language models, and has the potential to improve their performance on various NLP tasks.

In this thesis, I will work with the two presented $\text{GBERT}_{\text{SWM}}$ and $\text{GBERT}_{\text{WWM}}$ models to shed light on the effect that WWM has on trained neural MT evaluation metrics. The GBERT models as well as the experiments will be described in detail in Chapter 4.

[4] Chan et al. (2020) call the SWM model $\text{GBERT}_{\text{Data}}$ and the WWM model $\text{GBERT}_{\text{Data + WWM}}$. I rename the models as $\text{GBERT}_{\text{SWM}}$ and $\text{GBERT}_{\text{WWM}}$ respectively as these names are more appropriate to the research questions addressed in this thesis.

3 Analysis of Compound Translations in MBR-Decoded Outputs

As outlined in the previous chapters, trained neural evaluation metrics achieve superior correlations with human quality judgements and reach a new state-of-the-art. Nonetheless, Amrhein and Sennrich (2022) demonstrated that trained neural metrics have certain blind spots and are insensitive to specific linguistic phenomena. These weaknesses can cause the metric to unpredictably assign overly optimistic scores to bad translations. As these metrics become more widely adopted and are even recommended as a default to evaluate the quality of MT outputs (Kocmi et al., 2021), it is crucial to uncover and study their weaknesses, before we can safely optimize towards them.

In their analysis, Amrhein and Sennrich (2022) found first, yet unsystematic evidence that COMET-20 tends to choose German hypotheses containing nonsensical compounds. So far, this potential weakness of COMET-20 was not yet studied in detail. The aim of this thesis is to fill this gap and to provide a systematic analysis of COMET-20's performance regarding German compounds.

In German, composition is a complex and productive morphological process. Newly formed compounds are frequently encountered in German texts. They might denote new objects or concepts, e.g., *Coronapandemie (corona pandemic)*, or refer to a very specific context or scenario, such as *Horror-Sturz* which denotes a severe sports accident. In the formation of compounds, lexemes loose their inflectional suffixes such that two (or more) word *stems* are linked together. Between the constituents a linker element, the so-called *Fugenelement*, can appear, as in *Arbeit-**s**-tag (work day).*

Nominal compounds are the most frequent compound type in German. While other types such as verbal, adjectival and adverbial compounds exist, they are clearly less prevalent. In a nominal compound, the last component is always a noun, while the other constituents may belong to different parts of speech, e.g. *Koch-topf* (*cooking pot*) where the first component is a verb and *Gegen-teil* (*contrary, opposite*) where the first component is a preposition. However, most commonly, two (or more) nouns are merged.

This thesis will shed light on the behaviour of COMET-20 towards German nominal compounds. As the evidence suggesting that COMET-20 may struggle with selecting correct German compounds is still anecdotal, I will first conduct a systematic examination to determine whether this is indeed a weakness of the metric, or whether other metrics, used as utility functions in MBR decoding, exhibit similar difficulties in handling German compounds.

To this end, I will follow the procedure proposed by Amrhein and Sennrich (2022) (cf. 2.3). To ensure the comparability to their results (cf. 5.3), I use the same models, data and methods as they did.

3.1 Materials and Methods

The following paragraphs describe the translation model, the implementation of MBR decoding and the data sets used by Amrhein and Sennrich (2022). In addition, I propose new methods to specifically examine the translation of German compounds.

3.1.1 Translation Model

As this work focuses on German compounds, a translation model is needed that translates into German. To keep the comparability with the results of Amrhein and Sennrich (2022) high, their translation model for the language pair en→de is used. It is a Transformer Base model (Vaswani et al., 2017) trained with the `nematus`[1] framework (Sennrich et al., 2017). It consists of 6 encoder layers, 6 decoder layers and 8 attention heads. It uses an embedding and hidden state dimension size of 512 and a feed-forward network dimension of 2048. The sub-word vocabulary amounts to a size of 32k and is computed with byte pair encoding (Sennrich et al., 2016) using SentencePiece (Kudo and Richardson, 2018). The maximum sequence length during training and decoding is limited to 200 tokens (Amrhein and Sennrich, 2022).

The model is trained on the parallel data of the WMT 2018 News Shared Task (Bojar et al., 2018), excluding the ParaCrawl corpus. Hence, the training data includes the Europarl, the News Commentary and the Rapid-2016 corpora containing 5.9 million sentence pairs. After deduplicating the corpus, the model was trained on approximately 5.6 million sentence pairs (Amrhein and Sennrich, 2022). For a more detailed description of the translation model including hyperparameters, I refer the reader to Amrhein and Sennrich (2022).

Amrhein and Sennrich (2022) then used the trained model to generate the samples that serve as candidate and support sets in MBR decoding based on the evaluation data set described in the next section.

[1] github.com/EdinburghNLP/nematus

3.1.2 Data set for Evaluation and MBR Decoding

Amrhein and Sennrich (2022) evaluate their translation model on the test set of the WMT 2021 News Shared Task (Akhbardeh et al., 2021). The reasons for choosing this data set for evaluation are three-fold.

1. The data set was not part of the training data of COMET-20. Thus, neither the source sentences and translations of them nor any quality scores assigned to them have been previously observed by COMET-20.
2. The data set does not contain any translationese source sentences. All source sentences were originally written in the source language and translated by professional translators into the target language.
3. The data set contains two references for each source sentence for the translation direction under study, en→de.

The test set for en→de consists of 1002 sentence triplets, a source and two human reference translations. Based on these source sentences, Amrhein and Sennrich (2022) use the translation model described in section 3.1.1 to generate 100 unique translation hypotheses for each source sentence. This results in a set of 100,200 hypotheses in total.

Throughout this thesis, I will use this set of hypotheses as input to the MBR decoding, whereby it serves as both candidate pool and support set. The source and/or the reference sentences of the test set are used for the semi-automated analysis (cf. 3.3), the computation of various automatic metrics (cf. 3.4 and 4.3.2) and the sensitivity analysis (cf. 4.3.5).

3.1.3 COMET Model

Following Amrhein and Sennrich (2022), the `wmt20-comet-da` model (Rei et al., 2020b), referred to as COMET-20, is used as utility function in MBR decoding. Another reason for selecting this version of COMET is its status as the recommended default model at the time of planning the thesis.

COMET-20 predicts a Direct Assessment (DA) score for a translation hypothesis given the source sentence and the reference. It was trained on the DA scores from the WMT17 to WMT19 data sets (Bojar et al., 2017; Ma et al., 2018, 2019) and submitted to the WMT20 Metrics Shared Task (Mathur et al., 2020).[2] It is considered to correlate very well with human quality scores (Mathur et al., 2020; Freitag et al., 2021b, 2022b; Kocmi et al., 2021).

[2] For more details about the model and the data sets see Chapter 4.

3.1.4 Implementation of MBR Decoding

For sampling-based MBR decoding, this work follows the procedure described in Amrhein and Sennrich (2022). As proposed by Eikema and Aziz (2020, 2022), they let the translation model generate 100 unbiased samples for each source segment. Each sample is unique to avoid that it receives a higher score simply because it occurs multiple times in the support set. The same 100 samples are used as both the candidate and the support set.

As COMET includes the source sentence x in its analysis, Amrhein and Sennrich (2022) reformulate the MBR problem as:

$$y^* = \underset{h_i \in C}{\operatorname{argmax}} \frac{1}{|S|} \sum_{h_j \in S} u(x, h_i, h_j) \tag{3.1}$$

All the analyses and experiments in this thesis with a COMET model as utility function, use the MBR implementation of Amrhein and Sennrich (2022)[3] or a slightly modified version of it (cf. 4.3.1). In contrast to the MBR implementation of the original COMET repository[4], the implementation by Amrhein and Sennrich (2022) has the advantage that the support set and the candidate set can contain different hypotheses and a different number of sentences. Moreover, their implementation is very efficient, as they encode each segment only once with XLM-RoBERTa$_{\text{Large}}$ and cache the embeddings. This allows them to compute the scores of all candidate-support pairs in parallel.

In the subsequent sections and chapters, the MBR-decoded outputs obtained with COMET-20 as utility function (MBR$_{\text{COMET-20}}$) are compared to the outputs produced with ChrF and ChrF++ (Popović, 2015, 2017) as utility functions. For the MBR decoding with these two lexical overlap-based metrics, the implementation[5] by Eikema and Aziz (2022) is applied.

3.2 Manual Exploration of Compound Translations in MBR$_{\text{COMET-20}}$ Outputs

To gain a first overview over compound translations produced with MBR$_{\text{COMET-20}}$, the outputs are inspected manually. To this end, the MBR implementation by Amrhein and Sennrich (2022) is run with the candidate and support sets described in 3.1.2. The purpose of this preliminary analysis is to collect further evidence for the observations of nonsensical German compounds reported by Amrhein and Sennrich (2022) and to examine whether the formation of nonsensical compounds follows a certain pattern.

[3] github.com/chanberg/COMET-mbr

[4] github.com/Unbabel/COMET

[5] github.com/Roxot/mbr-nmt

The analyses in this thesis focus exclusively on *nominal compounds*. Wrongly formed adjectival compounds are much rarer and other types of compounds occur even less frequently. As the encountered samples of the non-nominal compound types are too small to be representative on the one hand and have only a small impact on the results on the other hand, I decided to exclude them from the analysis.

The results of this first, still unsystematic exploration, indicate that COMET-20 indeed shows difficulties with assessing German compounds correctly. Several error types appear frequently in the MBR-decoded translations chosen by COMET-20. Table 1 contains a few illustrative examples.

ID	MBR COMET	Reference	Source
3	Hundemarkt	Hundepark	dog park
7	Video-Fotografie	Videoaufnahme	video footage
71	Schraubschwendern	Schraubenziehern	screwdrivers
244	Gartengebäcken	Garten-Gimmicks	gardening gimmicks
816	Straßenbündnisse	Straßenbahnen	trams
74	Löschwagen	Lieferwägen	vans
118	Schlauchmitteln	Arzneimitteln	drugs
256	Himmelbäume	Obstbäumen	fruit trees
51	Alkoholschwester	alkoholischen Bruder	alcoholic brother
135	Pflegeleiterin	Interimstrainer	caretaker manager
408	Zieldifferenz	Tordifferenz	goal difference
676	Twitter-Posten	Twitter-Post	Twitter post
843	Strahlzentrum	strahlendes Zentrum	beaming center

Table 1: Examples of mistranslated German compounds

In many cases, COMET-20 chooses a compound whose first constituent was translated correctly while the subsequent components are incorrect or vice versa (two upper sections of the table). Moreover, COMET-20 does often not recognize gender mistakes (third section). It exhibits difficulties with polysemous words as well as loanwords and sometimes analyses concept boundaries erroneously (forth section). Interesting insights can be gained from the different kinds of mistakes encountered in compounds. However, this topic is not further discussed here as it is not the major focus of the thesis. An extensive description and discussion of the various error types can be found in Appendix B.

It is noteworthy that in many cases the correct German compound is found in one or more candidates produced by the MT system. In fact, the correct compound is often even among the most frequent variants in the candidate pool (cf. 4.3.5). In these cases, it is clearly the metric used as utility function that fails in selecting the correct variant. However, in other cases, it is the MT model that fails in translating the compound adequately. Consequently, none of the candidates contains the correct compound. In these cases, the metric only has the choice between wrong variants. This issue is further addressed in 4.3.5.

3.3 Semi-Automated Analysis of Compound Translations in $\text{MBR}_{\text{COMET-20}}$ Outputs

The preliminary manual exploration of the $\text{MBR}_{\text{COMET-20}}$ output revealed several kinds of mistakes associated with the translation and formation of German compounds. However, before we can conclude that German compounds are indeed a blind spot of COMET-20, a more systematic analysis of compound translation and a comparison to other decoding methods is required.

To that end, the newly formed compounds are extracted from the $\text{MBR}_{\text{COMET-20}}$ translations of the WMT 2021 News Shared Task test set (cf. 3.1.2). The number of incorrectly formed compounds that are bad translations is then compared to the number of ill-formed compounds in the outputs of three other decoding strategies:

- **Beam Search:** Most state-of-the-art MT systems use beam search for decoding. This algorithm is not known to exhibit specific difficulties with the formation of German compounds. Hence, it is expected that it will perform better than $\text{MBR}_{\text{COMET-20}}$ regarding the translation and formation of compounds.

- **$\textbf{MBR}_{\textbf{ChrF}}$:** To analyze whether incorrect German compounds are a specific blind spot of COMET-20 or whether they are an artifact of the candidate pool used in MBR decoding, the $\text{MBR}_{\text{COMET-20}}$ outputs are compared to the outputs obtained with two other utility functions. The first of them is ChrF. Out of the string-based evaluation metrics, ChrF and its variants show the highest correlation with human judgements and are therefore recommended by Kocmi et al. (2021) as secondary default metric to complement COMET.

- **$\textbf{MBR}_{\textbf{ChrF++}}$:** ChrF++ is an optimized variant of ChrF that correlates even better with human judgements (Popović, 2017). It is expected that $\text{MBR}_{\text{ChrF++}}$ selects similar candidates as MBR_{ChrF}, but that the output qualtiy of $\text{MBR}_{\text{ChrF++}}$ is slightly higher. The comparison of $\text{MBR}_{\text{COMET-20}}$ to the basic string-based metric ChrF and its more sophisticated variant ChrF++ will reveal on the one hand whether string-based metrics are more stable regarding German compounds than the neural metric COMET-20 and on the other hand how well MBR performs in terms of compound formation compared to beam search.

To gain a rough overview over the quality of the different outputs, they will first be evaluated in terms of various automatic metrics. The aim of this evaluation is to conduct a sanity check, which assesses whether all outputs are of a reasonable and comparable quality. Four different evaluation metrics are chosen for this purpose. Two of them are string-based metrics, namely BLEU and ChrF++. The other two are the neural metrics BLEURT-20 and COMET-20. As explained in Chapter 2.2, different types of metrics exhibit distinct strengths and weaknesses. While the scores generated by string-based metrics are predictable and allow to examine which sub-string has the largest impact on the quality score, neural metrics are able to reward paraphrases and synonyms. To gain a comprehensive picture of the quality of the MT outputs, both types of metrics were taken into account. BLEU was chosen because it has long been the most popular MT evaluation metric. However, ChrF++ generally correlates better with human judgement and is therefore included in the analyses. BLEURT-20 and COMET-20 were selected because they are currently considered state-of-the-art with the highest correlation with human quality scores.

Then, the formation of compounds in the outputs obtained with the aforementioned four decoding strategies is analysed and compared. The focus of this work is on the incorrectly formed, nonsensical compounds. To ensure that all mistranslated compounds are included in the analysis, a broad definition of *compound* is adopted. A word is considered as a compound whenever two (or more) constituents are identifiable, even if one (or more) of these constituents are nonsensical or not entirely correct word stems. An example of an ill-formed compound in which neither of the components is a correct word stem is *Schraub-schwender*, an attempt to translate *screw driver* (*Schraubenzieher* in German). The first compound part lacks the ending *-en* of the correct stem *Schrauben.* Such deletions of characters, as well as insertions of additional characters, are a commonly observed phenomenon in the formation of incorrect compounds. The second part of *Schraub-schwender* is completely nonsensical. Nonetheless, the two “stems” are still clearly identifiable as two parts of a compound. Hence, it is included in the analysis.

Extracting compounds automatically is a complex task and identifying incorrectly formed compounds that cannot be found in a dictionary or a data corpus is even more challenging. Therefore, the extraction of ill-formed compounds is approached in a multi-step procedure that is explained in more detail in the subsequent paragraphs:

1. Automated extraction of unknown words[6] from the MT outputs

2. Semi-automatic retrieval of all compounds from the list of unknown words

3. Manual selection of nonsensical compounds from the list of compounds

3.3.1 Extraction of Unknown Words

Incorrectly formed, nonsensical compounds are typically newly invented by the MT model and are not encountered in an existing data corpus. Hence, in order to identify such novel, potentially nonsensical items, the words in the MBR-decoded translations are compared to the words in the training data, i.e. the WMT 2018 News Shared Task corpus excluding the ParaCrawl dataset, and to the words in the two German references of the WMT 2021 News Shared Task test set. The training corpus consists of large amounts of data. Hence, a common, canonical word is very likely to appear in it. The reference translations

[6] Throughout this thesis, a simple operational definition of the term *word* is used: A *word* is defined as the character sequence between two white spaces.

correspond to the sentences translated via MBR decoding. They enrich the vocabulary with proper names, numbers and new, but correct compounds that appear in the test sentences, but that might not be found in the training corpus. A simple pre-processing is applied to remove punctuation characters, except for hyphens that often form an integral part of German compounds. This results in a vocabulary with 2,399,484 distinct words. When a word in the MT output is not found in this large vocabulary, it is reasonable to infer that the word is a newly formed, non-canonical item.

Finally, each MBR-decoded sentence is compared to its corresponding source sentence, to exclude untranslated proper names as well as other copied words from the analysis. Copying of input words is a distinct phenomenon from that of forming novel words and should be studied separately.

If a machine-translated word is neither found in the large German vocabulary nor in the source sentence, it is added to the list of "unknown words". The number of unknown words encountered in the output is a first indicator of the reliability of a certain decoding strategy. If the number of unknown words is high, it is very likely that a large part of them is nonsensical.

3.3.2 Identifying Compounds in the List of Unknown Words

Having identified the unknown, newly created words, the next step is to retrieve the compounds from that list. This purpose requires a morphological analysis. As outlined above, automatically conducting a morphological analysis of German words is not trivial, especially when productive word formation processes such as composition are involved.

Only few tools are designed to handle morphologically productive phenomena in German. One of them is the Zurich Morphological Analyzer for German (Zmorge) by Sennrich and Kunz (2014). Zmorge is based on the Stuttgart Morphological Analyzer (SMOR) (Schmid et al., 2004) that uses finite-state transducers. SMOR splits words into their morphemes and annotates them with the corresponding tag. In addition to inflection, it covers productive word formation processes such as derivation and composition (Schmid et al., 2004). Zmorge uses a slightly modified version of SMOR, with more conventional lemmatization and tags, and combines it with a large, constantly updated lexicon extracted from the German Wikitionary that provides additional morphological knowledge (Sennrich and Kunz, 2014).

Zmorge is used to classify the deduplicated set of unknown words into five different categories described below. The results returned by Zmorge are then analysed for certain tags[7] listed in Table 2.

According to the appearance of these tags in the analysis returned by Zmorge, the unknown words are grouped automatically into five classes:

[7] The list of tags is cited from pub.cl.uzh.ch/users/sennrich/zmorge/

tag	description
`<#>`	marks the compound boundary
`<->`	marks the joining element (*Fugenelement*) in compounds
`<TRUNC>`	marks hyphenation
`<~>`	marks other morpheme boundaries

Table 2: Zmorge tags related to compounds

- **Non-words:** If Zmorge is unable to morphologically analyze a certain word and returns `no result for <word>`, it is assumed that the respective item is nonsensical and does not constitute a word.
- **Non-nouns / non-compounds:** As this work focuses exclusively on nominal compounds, only words tagged as nouns are of interest. Hence, if none of the analyses by Zmorge contains the tag `<+NN>`, the item is added to the "no noun" category.
- **Known compounds:** If one of the analyses returned by Zmorge contains either the tag `<#>` or `<TRUNC>` along with `<+NN>` it is a compound. For compounds that exist as lexemes in its vocabulary, Zmorge returns at least one analysis where the compound is not split up and treated as a conjunct item. In this case, the word in question is considered as known compound. An example for such an analysis is given below[8]:
  `Schraub<~>en<#>zieh<~>er<+NN><Masc><Nom><Sg>`
  `Schraube<~>n<#>zieh<~>er<+NN><Masc><Nom><Sg>`
  `Schraubenzieher<+NN><Masc><Nom><Sg>`
- **New Compounds:** If at least one of the morphological analyses includes the tag `<#>` together with `<+NN>` and the word is not known to Zmorge as a lexeme, it is categorized as new compound. An example is shown below:
  `Kett<~>en<#>säule<+NN><Fem><Nom><Sg>`
  `Kette<->n<#>säule<+NN><Fem><Nom><Sg>`
- **Hyphenated Compounds:** If the constituents of a compound are separated by a hyphen, marked with `<TRUNC>`, it is assigned to this category.
  `{Corona}-<TRUNC>Pandemie<+NN><Fem><Nom><Sg>`

Inspecting the results of this automatic classification reveals that the accuracy of the analysis varies considerably between the different categories. While the majority of the items assigned to *new compounds* and *hyphenated compounds* are indeed newly formed compounds, many items in the class *known compounds* are actually nonsensical such as *Nacht-brücke (night bridge)*. In contrast, the category *non-compounds* contains a few items

[8] In this and the following examples, only a selection of the analyses by Zmorge is shown.

that are compounds. Most problematic is the category *non-words* that includes many nonsensical compounds for which Zmorge could not find a valid morphological analysis such as *Polizeigelicht*, a mistranslation of *police custody* (*Polizeigewahrsam* in German). However, for the purpose of this work, the inclusion of nonsensical compounds in the analysis is crucial.

Therefore, the automatic classification by Zmorge is manually corrected to identify *all* unknown compounds. As composition is a productive morphological process in German, novel compounds might be correctly formed, valid translations of the source text. Therefore, the collected compounds are manually classified into *correct* and *incorrect.* If necessary, the source and sometimes the reference are consulted to decide if a compound is correct or not. The resulting list of mistranslated compounds forms the basis for comparing the performance of different decoding strategies regarding the translation and formation of German compounds.

3.3.3 Limitations of the Approach

The chosen approach exclusively extracts compounds from the German MT hypotheses, completely ignoring the source sentence. While in most cases, the extracted novel compound corresponds to a compound in the source as well as in the reference, various exceptions, where this correspondence does not apply, were observed.

Firstly, a German compound sometimes corresponds to a simple noun in the English source. An example is the English word *tram* that translates to German as *Straßenbahn.* Any wrong, nonsensical translations such as *Straßenbündnis* are included in the analysis of mistranslated compounds. Secondly, the opposite case where an English source compound corresponds to another syntactical construction in German, was also observed. For example, the most idiosyncratic translation of *the Chengdu consulate* is *das Konsulat in Chengdu.* The German compound *das Chengdu-Konsulat* sounds unnatural and is included in the list of nonsensical compounds. In other cases, German compounds appear in the translation hypothesis, although neither the source nor the reference words are compounds.[9] These three cases are methodologically unproblematic as they concern a nonsensical form of a German compound and are hence included in the analysis.

However, such mismatches between syntactical constructions in the source and the target language complicate the task of semi-automatically identifying mistranslated compounds and hint at the limitations of the chosen approach.

An English compound might be translated with an inappropriate syntactical construction in the hypothesis. For example, *no-confidence motion* corresponds to the German compound *Misstrauensantrag*, but is translated as *vertrauensloser Antrag* in the MT output. The approach is unable to capture wrong translations of English compounds, if they

[9] See Tables 23 and 27 in Appendix B for examples.

consist of canonical but inappropriate words.

Moreover, canonical compounds that appear in the hypothesis are discarded from the analysis, as it is assumed that well-formed compounds are correct translations. However, there is no guarantee that this is indeed the case. On the contrary, the hypotheses sometimes contain a canonical but inappropriate compound. For instance, *baboon attacks* is wrongly translated as *Bombenangriffe (bomb attacks)*. Even though this translation conveys a very different meaning, it is disregarded in the analysis as *Bombenangriffe* is a canonical word.

These limitations need to be taken into account when the results of the analysis are interpreted. The mere number of incorrectly formed compounds alone is not sufficient to draw a reliable conclusion on a metric's blind spot. Nonetheless, the frequency of ill-formed compounds in the MT output are a first indicator of which metrics exhibit special difficulties, as it can be assumed that the mentioned limitations apply to all decoding strategies to a similar extent.

3.4 Results

This section presents the results of the analyses described in the previous paragraphs. To ensure that the general quality of the MT outputs obtained with different decoding strategies is comparable, a quality evaluation in terms of various MT metrics was conducted. The results are shown in Table 3.

	BLEU	ChrF++	BLEURT-20	COMET-20
Beam Search	**26.220**	**52.961**	**62.841**	**0.2601**
MBR ChrF++	20.941	48.330	52.114	-0.1084
MBR ChrF	20.390	48.156	51.825	-0.1225
MBR COMET-20	16.980	44.580	57.281	0.2440

Table 3: Evaluation of the translation quality in terms of automatic metrics

According to all four evaluation metrics, the beam search outputs are clearly of the highest quality. The metrics agree that the MBR-decoded outputs do not achieve the quality of the beam search outputs, but three of the four metrics consider them to be of reasonable quality. While the two metrics based on lexical overlap between the hypothesis and the reference, BLEU and ChrF++, prefer outputs generated with a surface-overlap metric as utility function, the neural metric BLEURT-20 favorizes outputs obtained with $\text{MBR}_{\text{COMET-20}}$. Interestingly, the evaluation metric COMET-20 shows a strong preference for MBR outputs generated with itself as utility function and for beam search outputs. Translations generated with string-based utility functions are deemed to be of much inferior quality. This quality gap is clearly larger than the one observed in the BLEURT-20 ratings.

Having confirmed that the MBR-decoded outputs are of reasonable quality, the compounds in the four MT output sets, each containing 1002 sentences, are analysed according to the procedure described in 3.3. The results of this analysis are shown in Table 4.

	Unknown Words	New Compounds	Mistranslated Compounds
Beam Search	**461**	**374**	**241**
MBR ChrF++	790	476	386
MBR ChrF	844	511	415
MBR COMET-20	855	553	447

Table 4: Unknown words, newly formed compounds and mistranslated compounds found in the output sentences of different decoding strategies

The results demonstrate that MBR decoding, regardless of which utility function was applied, generates by far more unknown words, and in consequence more novel and ill-formed compounds, than beam search decoding. Nonetheless, the utility function still has an effect on the number of unknown words. COMET-20 seems to be less sensitive to new, and possibly incorrectly formed words and compounds in the translation candidates. It selects candidates containing unknown words more often than the utility functions ChrF++ and ChrF. In the analysed 1002 sentences, the $\text{MBR}_{\text{COMET-20}}$ outputs contain almost twice as many unknown words and mistranslated compounds than the beam search outputs.

As expected, a large portion of the unknown words are compounds since composition is a productive morphological process in German that is frequently used in the news domain. The exact proportions are given in Table 5.

	Compounds/Unknown	Mistranslated/Compounds
Beam Search	81.12 %	64.45 %
MBR ChrF++	60.25 %	81.09 %
MBR ChrF	60.54 %	81.21 %
MBR COMET-20	64.68 %	80.83 %

Table 5: Portion of compounds among the unknown words and portion of mistranslated compounds among the compounds

The results highlight that especially in the beam search outputs most unknown words are compounds. This rate is clearly lower for the MBR outputs. In addition, it stands out that more than 80 % of the unknown compounds found in the MBR outputs are incorrectly formed, nonsensical translations. Regarding the beam search outputs, the proportion of mistranslations among unknown compounds is still high, but clearly lower than in MBR decoding.

To gain further insights into which types of unknown words occur most frequently with a certain decoding strategy, the results are analysed per morphological category in Table 6.

The table shows the number of unknown items (Its.) found in each of the five categories presented in 3.3.2 and the number of compounds (Comp.) identified among these items.

Decoding	New		Known		Hyphenated		Non-Nouns		Non-Words	
	Its.	Comp.	Its.	Comp.	Its.	Comp.	Its.	Comp.	Its.	Comp.
Beam Search	157	156	6	3	155	153	27	1	116	61
MBR ChrF++	301	275	4	2	93	87	74	12	318	100
MBR ChrF	322	298	6	4	98	92	81	9	337	108
MBR COMET	310	295	14	12	125	121	71	11	335	114

Table 6: The number of items (Its.) in each morphological category after the automatic classification with Zmorge and the number of manually identified compounds (Comp.) in each category

When compared to beam search, all three MBR-based decoding strategies output about twice as many new and nonsensical compounds (classified as *new* and *non-words* by Zmorge). Moreover, their outputs contain about three times as many items that are not nouns, mainly unknown adjectives and verbs. In comparison to the other decoding methods, MBR$_{\text{COMET-20}}$ outputs include more items that are considered as "known compounds" by Zmorge. However, as described in 3.3.2, most of these items are actually nonsensical.

Interestingly, beam search shows a much stronger preference for hyphenated compounds than MBR decoding. More than 40 % of the unknown compounds in the beam search outputs are hyphenated, while this portion is about 20 % for all three MBR utility functions.

3.5 Discussion

The results presented in the previous section highlight several interesting phenomena that are discussed in the following paragraphs.

3.5.1 Automatic Evaluation of Translation Quality

The automatic evaluation with different MT metrics revealed that string-based metrics prefer MBR outputs obtained with a string-based metric as utility function while neural metrics favor outputs generated with a neural metric as utility function. The effect is strongest if a metric evaluates MBR outputs produced with itself as utility function, as the results for ChrF++ and, in particular, COMET-20 show. This "metric overfitting" has been previously reported by Fernandes et al. (2022). They observe that the "metric gap" is largest for COMET-20. This finding coincides with the results in Table 3. However, the strength with which COMET-20 favors translations generated with itself as utility function is unexpectedly high and by far exceeds the metric gap reported in Fernandes et al. (2022). This problem is further addressed in 6.1.

3.5.2 Translation of Compounds

The results in Table 4 illustrate that MBR decoding generates more unknown and, consequently, more incorrectly formed compounds than beam search. This finding is in line with my expectations. Unbiased MBR decoding is based on a candidate pool generated via ancestral sampling on sub-word level. In contrast to beam search, ancestral sampling is not a path-finding algorithm and hence lacks a mechanism to control that the drawn and juxtaposed sub-words result in a correct and meaningful word. As a result, the word formation process is more susceptible to errors such as insertion or deletion of characters or wrong combinations of compound constituents. In consequence, the available candidates from which the utility function can choose are generally of a lower quality than beam search translations and contain more unknown, non-canonical words. Hence, it is inevitable that the final outputs exhibit a higher number of incorrect, unknown and possibly nonsensical words. Improving the quality of the candidate pool without introducing a strong bias is of crucial importance for future deployment of MBR decoding, but it is not the focus of this work as it is outside the responsibility of the utility metrics.

What is of interest here, is the question of how the different utility functions of MBR decoding perform regarding nonsensical compounds, as this comparison allows to reveal potential blind spots of a metric. The results in Tables 4 and 6 highlight that COMET-20 is insensitive to newly formed words, and in particular to ill-formed compounds. The translations chosen by COMET-20 exhibit a large amount of unknown words that is in a similar range as that of ChrF and clearly higher than that of ChrF++. The performance difference between COMET-20 and ChrF++ is even more striking regarding mistranslated compounds. COMET-20 even performs worse than ChrF.

These results are surprising as COMET-20 is widely praised for its high correlation with human judgement, while string-based metrics are criticized for correlating poorly with human quality scores (Fernandes et al., 2022; Kocmi et al., 2021; Freitag et al., 2021b,a). It is particularly unexpected that ChrF outperforms COMET-20 regarding the sensitivity to poorly translated compounds, as ChrF is a relatively simple metric based on character overlap, while its variant ChrF++ is slightly more sophisticated.

Finally, the results in Table 6 underline that beam search decoding generates especially many hyphenated compounds. The reasons for this preference are unclear and I leave it to future work to investigate them.

In conclusion - even though the analysis should be taken with a grain of salt as outlined in 3.3.3 - the results indicate that the supposition of Amrhein and Sennrich (2022) is correct and nonsensical compounds are indeed a weakness of COMET-20. This unpredictable behavior is concerning as COMET was proposed as the new default evaluation metric that should be widely adopted in future work (Kocmi et al., 2021). Therefore, it is crucial to understand COMET's blind spots and weaknesses.

For this purpose, further experiments were conducted aiming at 1) quantifying the sensitivity of COMET-20 towards compounds and 2) finding strategies that enhance the sensitivity of neural metrics towards German compounds. The experimental setup is described in the next chapter.

Moreover, the preliminary analysis raised questions related to the issue of metric overfitting. As an aside of the subsequent experiments, the metric gap of COMET-20 observed in the automatic evaluation will be further investigated as well.

4 Enhancing the Sensitivity towards German Compounds

As the preliminary MBR-based analysis in Chapter 3 uncovered, COMET-20 is less sensitive to incorrectly formed German compounds than the string-based evaluation metrics ChrF and ChrF++. Thus, the question arises whether it is possible to increase COMET's sensitivity towards the correct formation of German compounds.

To elevate COMET's sensitivity towards specific linguistic phenomena, I will experiment with exchanging one of the building blocks of COMET-20: the underlying language model, XLM-RoBERTa$_{\text{Large}}$ (Conneau et al., 2020). This multilingual model is pre-trained with the Masked Language Modelling objective using Sub-Word Masking (SWM). Hence, the model only has to predict one sub-word (or token) of a word, while the other sub-words are visible to the model which considerably facilitates the task.

German compounds are complex words that consist of at least two constituents, which are most likely split up into two or more sub-words during pre-training. It is therefore possible that models pre-trained with SWM do not sufficiently learn about the relationship between the different components of a compound. The observed error types described in 3.2 and in Appendix B are in line with this assumption. The MBR$_{\text{COMET-20}}$ outputs contain many examples of compounds where one constituent was translated correctly, while the other one is wrong.

One way to force the model to pay attention to an *entire* word, is by using Whole Word Masking (WWM) in pre-training. As explained in 2.4, WWM masks all sub-words of a word simultaneously, which renders the task to reconstruct the word more challenging. Even though each sub-word is still predicted independently, the model is forced to produce *all* parts of a word. Thus, it is likely that the model gains a better notion of what a word is and which sub-words belong together.

Hence, it is reasonable to assume that models pre-trained with WWM are better at dealing with German compounds. This and the subsequent chapters shed light on this hypothesis. To investigate the effect of WWM on the translation of compounds, I will train new neural metrics and use them as utility function in MBR decoding.

The German language model GBERT (Chan et al., 2020) will form the basis of the newly trained metrics, as it is available in two flavours: pre-trained 1) with Sub-Word Masking and 2) with Whole Word Masking while everything else is kept constant.

Different types of metric models will be trained. On the one hand, the new metrics should be comparable to COMET-20 to which Chapter 3 was dedicated. For this purpose, two German COMET (GCOMET) models are trained, based on the two GBERT variants. To use the multilingual COMET framework with a monolingual German model, some adaptations in its architecture are necessary (cf. 4.2.2). Hence, GCOMET is not entirely comparable to COMET-20. To allow for a fairer comparison, $\text{COMET}_{\text{Contrastive}}$ is trained. The only difference to GCOMET is that it is based on the multilingual XLM-$\text{RoBERTa}_{\text{Base}}$. It serves to investigate the effect of different underlying language models and to compare the multilingual to the monolingual setting.

On the other hand, two German BLEURT (GBLEURT) models will be trained, based on the two variants of GBERT. In contrast to COMET, BLEURT was originally designed for the monolingual evaluation setting and can be easily adapted to the purpose of this work.

Training two distinct types of metrics that are otherwise as similar as possible, using the same underlying language model and the same training data, allows for insights into the effects of different metric architectures. This will shed light on the question whether German compounds are generally a blind spot of neural metrics or whether they are a specific weakness of COMET. Further, it allows us to investigate whether the different architectures have an impact on the metrics' sensitivity towards other linguistic phenomena and whether the effects of WWM vs. SWM are the same for both metric types.

To answer these questions, various experiments will be conducted. They will not only measure the sensitivity of different metric types towards specific linguistic phenomena, but they will also elucidate the issue of metric overfitting and explore different implementations of MBR decoding.

Before the experimental setup and the experiments are described, the theoretical framework of the deployed models and the data sets are outlined in the next section.

4.1 Materials and Methods

To better understand the models and frameworks that were used for the experiments, their theoretical background is elucidated in this section. The metrics trained for this thesis are based on German BERT (GBERT) embeddings described in 4.1.1. Then, the BLEURT and COMET frameworks are delineated in 4.1.2 and 4.1.3 respectively. Finally, the datasets used to train and evaluate the GBLEURT and GCOMET models are described in 4.1.4.

4.1.1 GBERT

To investigate the effect of WWM in comparison to SWM, I work with the embeddings of two German BERT models (GBERT) provided by Chan et al. (2020) to train my German

COMET and BLEURT metrics. GBERT$_{\text{WWM}}$ is pre-trained with Whole Word Masking and is publicly available on Hugging Face as `deepset/gbert-base`.[1] GBERT$_{\text{SWM}}$ is pre-trained in the exact same way, except that it uses Sub-Word Masking instead of WWM. This model is not publicly available. The authors of GBERT shared it with me upon my request.

I chose to work with the two GBERT models, because they enable a direct comparison between SWM and WWM and their effect on the final performance of the fine-tuned metric model on a downstream task, i.e. scoring the quality of a translation candidate. Moreover, GBERT uses a *cased* vocabulary, which is important when investigating the model's sensitivity towards errors in German compounds, as German nouns always start with a capital letter.

The GBERT models use the same architecture and training objectives as the original BERT model (Devlin et al., 2019). BERT, and hence GBERT, use a *bidirectional* approach to pre-train deep representations that are conditioned on both the left and right context of a word in all layers (Devlin et al., 2019). BERT takes as input either a *single sentence* or a *sentence pair*. Each sequence starts with the special [CLS] token. Its final hidden state is regarded as an aggregated summary of the information contained in the encoded sequence and is used as input to classification and regression tasks. If the input is a sentence pair, sentence A and sentence B are separated by the special [SEP] token. BERT represents an input token as the sum of three embeddings: 1) the token embedding, 2) the segment embedding that indicates whether the token belongs to sentence A or B and 3) the position embedding that indicates the token's position within the entire sequence (Devlin et al., 2019).

The pre-training of GBERT involves two different objectives: Next Sentence Prediction (NSP) and Masked Language Modelling (MLM). In the NSP task, the model learns to understand the relationship between sentences by predicting whether sentence B is the next sentence that follows sentence A. For this task, GBERT encodes *sentence pairs*. In the MLM objective, GBERT predicts the masked tokens. The masking can either be applied on sub-word or on whole word level (cf. 2.4).

Both GBERT$_{\text{WWM}}$ and GBERT$_{\text{SWM}}$ have been pre-trained on the same large-scale corpus consisting of German monolingual, unlabelled data sets with a total size of 163.4 GB. Specifically, the German parts of the corpora (see Suárez et al., 2019; Tiedemann, 2012; Ostendorff et al., 2020) listed in Table 7 were used as training data. As the majority of these texts were scrapped from the internet, Chan et al. (2020) emphasise that the GBERT models might have learned various kinds of biases present in the data. One should keep that in mind when analysing results from experiments with language models.

To train the GBERT models, Chan et al. (2020) use the official Tensorflow training script[2]

[1] huggingface.co/deepset/gbert-base

[2] github.com/google-research/bert

Dataset	Description	Size in GB
OSCAR	Pre-processed Common Crawl texts	145
OPUS	Texts from various domains (movie subtitles, speeches, books, etc.)	10
Wikipedia	Pre-processed Wikipedia texts	6
OpenLegalData	Court decisions	2.4

Table 7: Training data of the GBERT models.

of BERT. The models correspond to BERT$_{\text{Base}}$ in size, having 110 million parameters. Apart from the different masking strategies, GBERT$_{\text{SWM}}$ and GBERT$_{\text{WWM}}$ are trained with the same hyperparameters (cf. Chan et al., 2020).

4.1.2 BLEURT

BLEURT[3] is a reference-based neural evaluation metric designed to predict human judgements of translation quality as precisely as possible. The original version of BLEURT (Sellam et al., 2020a) is based on BERT embeddings and is limited to English sentences. To evaluate a certain MT hypothesis, it encodes this hypothesis along with the reference as a *sentence pair.*

The training of BLEURT involves three different steps:

1. the normal BERT pre-training as defined in Devlin et al. (2019),
2. further pre-training on a large corpus of synthetic data to enhance BLEURT's robustness,
3. fine-tuning BLEURT in a regression task on task-specific ratings from the WMT Metrics Shared Task.

The regular BERT pre-training is not discussed here and I refer the reader to Devlin et al. (2019) and to Section 4.1.1 for details. The additional pre-training step aims at fostering BLEURT's robustness to variation in the references, potential errors in the MT hypotheses and domain shift. The synthetic data is perturbed to imitate the variation and mistakes introduced by MT models, e.g. phrase substitution, lexical alterations, paraphrases, noise, omissions, sentence truncation and void predictions. Sellam et al. (2020a) define four pre-training tasks on the synthetic data set and combine the different loss functions in a weighted sum.

Finally, the model is fine-tuned for quality evaluation in a regression task. BERT (Devlin

[3] **Bil**ingual **E**valuation **U**nderstudy with **R**epresentations from **T**ransformers

et al., 2019) converts a reference sentence $\boldsymbol{x}$ with the tokens $(\boldsymbol{x}_1, ..., \boldsymbol{x}_r)$ and a MT hypothesis $\tilde{\boldsymbol{x}}$ consisting of the tokens $(\tilde{\boldsymbol{x}}_1, ..., \tilde{\boldsymbol{x}}_p)$ into a sequence of contextualized vectors, formalized by Sellam et al. (2020a) as:

$$BERT(\boldsymbol{x}, \tilde{\boldsymbol{x}}) = \boldsymbol{v}_{[CLS]}, \boldsymbol{v}_{x_1}, ..., \boldsymbol{v}_{x_r}, \boldsymbol{v}_{\tilde{x}_1}, ..., \boldsymbol{v}_{\tilde{x}_p} \tag{4.1}$$

where $\boldsymbol{v}_{[CLS]}$ represents the special [CLS] token of BERT that is prepended to every encoded sequence. On top of this CLS vector, Sellam et al. (2020a) add a linear layer that learns to predict the quality scores assigned by human annotators to a translation hypothesis. Hence, Sellam et al. (2020a) describe the regression task as:

$$\hat{y} = \boldsymbol{f}(\boldsymbol{x}, \tilde{\boldsymbol{x}}) = \boldsymbol{W}\tilde{\boldsymbol{v}}_{[CLS]} + \boldsymbol{b} \tag{4.2}$$

where $\hat{y}$ denotes the predicted rating, $\boldsymbol{W}$ the weight matrix and $\boldsymbol{b}$ the bias vector. In the regression task, both the BERT parameters as well as the linear layer are fine-tuned on the supervised data set using the mean squared error (MSE) as regression loss (Sellam et al., 2020a):

$$l_{supervised} = \frac{1}{N} \sum_{N=1}^{N} ||y_i - \hat{y}||^2 \tag{4.3}$$

The regression task is trained on the WMT Metrics Shared Task data sets from the years 2017 to 2019 using only to-English language pairs. At the time of its release, BLEURT established a new state-of-the-art regarding the correlation with human judgement in terms of Kendall's Tau (Sellam et al., 2020a).

Later, Sellam et al. (2020b) extended BLEURT to other languages by using embeddings from the multilingual mBERT[4] model (Devlin et al., 2019), which was pre-trained on Wikipedia data in 102 languages (Mathur et al., 2020). For this multilingual BLEURT model, they omitted the second step of further pre-training and directly fine-tuned the regression layer on top of the CLS vector. Subsequent work by Pu et al. (2021) used the multilingual RemBERT embeddings (Chung et al., 2021) and distilled the large BLEURT model into a smaller model. Their experiments lead to the development of the currently recommended multilingual BLEURT-20.

4.1.3 COMET

The COMET[5] framework has been developed by Rei et al. (2020a) to predict various kinds of human assessment scores, such as Direct Assessment (DA), Multidimensional Quality Metrics (MQM) or Human-mediated Translation Edit Rate (HTER). COMET is a *mulitlingual* framework based on pre-trained embeddings of XLM-RoBERTa (Conneau et al., 2020). The multilingual embedding space allows COMET to consider not only

[4] github.com/google-research/bert

[5] **C**rosslingual **O**ptimized **M**etric for **E**valuation of **T**ranslation

the MT hypothesis and the reference, but also the *source sentence*, which leads to higher correlation with human judgements. This stands in contrast to BLEURT and most other reference-based metrics that ignore the source sentence.

Many different models have been trained under the COMET framework, estimator models as well as translation ranking models, reference-based as well as reference-free models. In the following paragraphs, I will focus on the two models that are relevant to my own work:

- `wmt20-comet-da` (COMET-20) is a COMET regression model trained to predict a DA score given the source, the MT hypothesis and the reference.
- `wmt20-comet-qe-da` (COMET-QE) is a reference-free COMET regression model trained to predict a DA score given the source and the MT hypothesis.

At the time when I started to work on this thesis, these two models were the recommended default models. Only recently, they were replaced by the newest COMET models (Rei et al., 2022).

Both models are based on three major building blocks: 1) a cross-lingual encoder, 2) a pooling layer and 3) a predictive neural network that regresses on human DA scores.

Cross-lingual encoder: The pre-trained cross-lingual encoder produces token-level embeddings for the source, the hypothesis and the reference. Thanks to the multilinguality of the encoder, the source is mapped to the same feature space as the hypothesis and the reference (Rei et al., 2020a). COMET-20 and COMET-QE both rely on XLM-RoBERTa$_{\text{Large}}$ to encode the input sentences (Rei et al., 2020b), but the COMET framework can be used with other encoders such as BERT or XLM.

Pooling Layer: Different layers of the pre-trained encoder may contain different kinds of linguistic information (Zhang et al., 2020). To leverage as much linguistic knowledge as possible, Rei et al. (2020a) use a layer-wise attention mechanism to combine the embeddings of various encoder layers for each token. Then, they apply average pooling to convert the sequence of token embeddings into a single segment-level representation for each input.

Estimator Model: Rei et al. (2020a) extract the following features from the segment-level embeddings of the source $\boldsymbol{s}$, the hypothesis $\boldsymbol{h}$ and the reference $\boldsymbol{r}$:

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

These combined features emphasize the differences between the embeddings in the semantic feature space. Rei et al. (2020a) concatenate them to the embeddings of the hypothesis

$\boldsymbol{h}$ and the reference $\boldsymbol{r}$. The source embedding $\boldsymbol{s}$ is not included as the semantic feature space between different languages is only loosely aligned (Rei et al., 2020a). Hence, this method results in the final feature vector:

$$\boldsymbol{x} = [\boldsymbol{h}; \boldsymbol{r}; \boldsymbol{h} \odot \boldsymbol{s}; \boldsymbol{h} \odot \boldsymbol{r}; |\boldsymbol{h} - \boldsymbol{s}|; |\boldsymbol{h} - \boldsymbol{r}|] \tag{4.4}$$

The input vector $\boldsymbol{x}$ is fed into a feed-forward regressor which is trained to minimize the mean squared error (see Equation 4.3) between the predicted score and the actual human score (Rei et al., 2020a). Hence, the COMET estimator model uses the same training objective as BLEURT. However, the features in their input vectors differ: While BLEURT inputs the [CLS] embeddings of a sentence pair ($\boldsymbol{r}$, $\boldsymbol{h}$), COMET extracts additional features that combine the information from the different embeddings and concatenates them to $\boldsymbol{r}$ and $\boldsymbol{h}$.

COMET-20 regresses on z-normalized Direct Assessment scores from the WMT 2017 to 2019 data sets including 24 language pairs (Rei et al., 2020a,b). It correlates well with human judgements on segment-level and system-level (Mathur et al., 2020), differentiates successfully between high-performing MT systems and is currently regarded as the top-performing metric (Kocmi et al., 2021).

COMET-QE is a reference-free quality estimation model. It is based on the same architecture as COMET-20 with the difference that it does not include the reference. In consequence, it extracts only two instead of four combined features: the element-wise source product $\boldsymbol{h} \odot \boldsymbol{s}$ and the absolute element-wise source difference $|\boldsymbol{h} - \boldsymbol{s}|$. In contrast to the reference-based COMET models, COMET-QE includes the source embedding $\boldsymbol{s}$ in the final feature vector in exchange for the reference embedding. Hence, the input vector to the feed-forward regression layer is:

$$\boldsymbol{x} = [\boldsymbol{h}; \boldsymbol{s}; \boldsymbol{h} \odot \boldsymbol{s}; |\boldsymbol{h} - \boldsymbol{s}|] \tag{4.5}$$

The dimensions of the feed-forward layer are reduced to match the reduced size of the input vector (Rei et al., 2020b, 2021). COMET-QE was trained on DA scores from the WMT 2017 to 2019 data sets. In the WMT 2020 Metrics Shared Task, it showed a competitive performance and was the only system able to differentiate correctly between human translations and MT output (Mathur et al., 2020).

4.1.4 Data Sets

Following both Sellam et al. (2020a) and Rei et al. (2020b), I train the GBLEURT and GCOMET models on the Direct Assessment data sets from the WMT 2017 to 2019 Metrics Shared Tasks (Bojar et al., 2017; Ma et al., 2018, 2019) and evaluate them on the official WMT 2020 Metrics Shared Task data set (Mathur et al., 2020). The WMT Metrics Shared Task is considered the reference benchmark for the evaluation of metrics that assess the quality of MT outputs (Sellam et al., 2020b).

The data sets provide for each segment the source sentence, one or more reference translations and the hypotheses generated by the MT systems that participated in the WMT News Shared Task of the respective year. The adequacy of the translation hypotheses is rated by human annotators on a continuous scale from 0 to 100. The resulting Direct Assessment scores are normalized for each annotator yielding a z-score (Mathur et al., 2020).

GBLEURT, GCOMET and COMET$_{\text{Contrastive}}$ are trained to predict the z-scores given a German hypothesis and a German reference. For the training, I use the to-*German* portion of the WMT 2017 - 2019 Metrics Shared Task data sets as provided by the authors of COMET on their Github page.[6] The number of segments in the training set and the language pairs from which the segments were extracted are described in Table 8.

WMT	segments	language pairs
2017	7 025	en→de
2018	10 208	en→de
2019	42 118	en→de, fr→de
total	59 351	en→de, fr→de

Table 8: Training data for GBLEURT, GCOMET and COMET$_{\text{Contrastive}}$

As development set, 100 source sentences and their corresponding hypotheses and references were chosen from the WMT 2020 Metrics Shared Task data set as provided by Unbabel.[7] As each source segment was translated by various participating MT systems, the development set contains 722 hypotheses. These segments overlap with the test set. However, the performance on the development set is exclusively used for early stopping. It is not used for hyperparameter fine-tuning or to guide any other decision related to the training of the metric models. Therefore, no data leakage occurs during the training process.

To evaluate the metric models, the official test set `newstest2020` of the WMT 2020 Metrics Shared Task is used, as provided by its organizers.[8] Since the trained models are monolingual, they were only evaluated on to-German segments. These 19 852 German translation hypotheses were generated by 14 different MT systems.

[6] github.com/Unbabel/COMET/blob/master/data/README.md

[7] github.com/Unbabel/COMET/blob/master/data/README.md

[8] WMT20_data: `newstest2020txt-v2.tar.gz`

4.2 Experimental Setup

Having outlined the theoretical background, this section describes the German BLEURT (GBLEURT) and the German COMET (GCOMET) models trained to investigate the major research question of this work: Does Whole Word Masking increase the sensitivity of neural metrics towards mistakes in German compounds?

To extend the scope of the thesis beyond the monolingual setting, $\text{COMET}_{\text{Contrastive}}$ is trained, that is based on the multilingual $\text{XLM-RoBERTa}_{\text{Base}}$, to elucidate the effect of a multilingual compared to a monolingual model.

4.2.1 GBLEURT

The BLEURT architecture is chosen because it is a *reference*-based neural evaluation metric that does not rely on the source segment. Hence, it was originally designed for the monolingual setting studied in this work. It is well-suited to be used in combination with the monolingual GBERT models.

As explained in 4.1.2, BLEURT relies on three different building blocks. Firstly, it builds on a large pre-trained language model. For the purpose of this thesis, two metric models are trained that rely on two variants of GBERT (Chan et al., 2020): $\text{GBERT}_{\text{SWM}}$ and $\text{GBERT}_{\text{WWM}}$. In analogy, the two resulting metric models are dubbed $\text{GBLEURT}_{\text{SWM}}$ and $\text{GBLEURT}_{\text{WWM}}$ respectively. Both variants of GBERT use a cased vocabulary and are of the size of $\text{BERT}_{\text{Base}}$ having 12 layers, 768 hidden units and 12 attention heads (Chan et al., 2020).

Secondly, BLEURT is additionally pre-trained on a large corpus of synthetic data to promote its robustness. This step is omitted for GBLEURT, as such a massive pre-training is beyond the scope and the resources of this work. In addition, Sellam et al. (2020a) conclude that, even though skipping the additional pre-training harms the model slightly, BLEURT is still competitive and Sellam et al. (2020b) omit this step for the multilingual BLEURT. Hence, the resulting GBLEURT models are nevertheless expected to be a good choice and to exhibit a competitive performance.

Thirdly, GBLEURT is fine-tuned on a regression task to predict human quality scores given a reference and an MT hypothesis. The fine-tuning follows the procedure in Sellam et al. (2020a), described in 4.1.2. The underlying GBERT model encodes the reference $\boldsymbol{x}$ and the hypothesis $\tilde{\boldsymbol{x}}$ as a *sentence pair*. This results in a sequence of contextualized vectors for the reference tokens $(\boldsymbol{x}_1, ..., \boldsymbol{x}_r)$ and the hypothesis tokens $(\tilde{\boldsymbol{x}}_1, ..., \tilde{\boldsymbol{x}}_p)$. As GBERT belongs to the family of BERT models, it uses the special [CLS] token at the beginning of the encoded sequence. The representations of the reference and the hypothesis are separated by the [SEP] token.

$$GBERT(\boldsymbol{x}, \tilde{\boldsymbol{x}}) = \boldsymbol{v}_{[CLS]}, \boldsymbol{v}_{x_1}, ..., \boldsymbol{v}_{x_r}, \boldsymbol{v}_{[SEP]}, \boldsymbol{v}_{\tilde{x}_1}, ..., \boldsymbol{v}_{\tilde{x}_p} \tag{4.6}$$

On top of the final hidden state of the [CLS]-vector, the linear regression layer is added as described in Equation 4.2. The regression layer as well as the parameters of GBERT are fine-tuned on Direct Assessment scores for German segments. The training data corresponds to the German portion of the data sets from the WMT 2017 - 2019 Metrics Shared Task as described in 4.1.4. Mean squared error is used as regression loss as formalized in Equation 4.3.

The final architecture of the GBLEURT models is depicted in Figure 2.

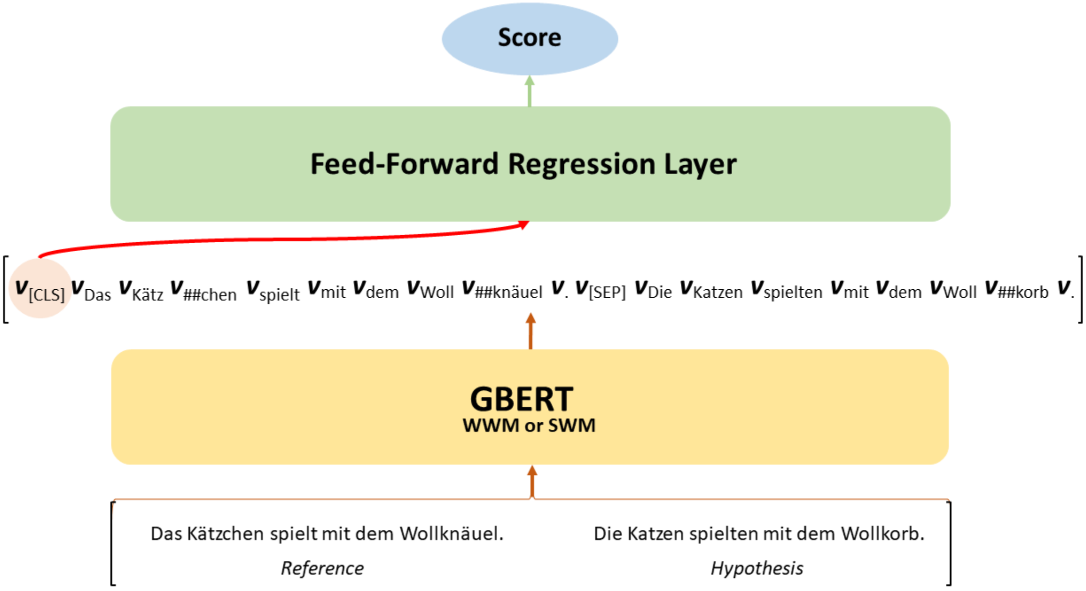
Score
Feed-Forward Regression Layer
v[CLS] vDas vKätz v##chen vspielt vmit vdem vWoll v##knäuel v. v[SEP] vDie vKatzen vspielten vmit vdem vWoll v##korb v.
GBERT
WWM or SWM
Das Kätzchen spielt mit dem Wollknäuel.
Reference
Die Katzen spielten mit dem Wollkorb.
Hypothesis

Figure 2: Architecture of GBLEURT

For training, the same hyperparameters as in Sellam et al. (2020a) were used, with two exceptions. As the training set is considerably smaller than that of the original BLEURT, the number of training steps for fine-tuning is reduced from 40 000 (Sellam et al., 2020a) to 20 000 for GBLEURT. With the given training corpus and batch size, 20 000 training steps correspond to iterating through the entire corpus approximately ten times. Training and evaluation are run in parallel. According to the reduced number of training steps, the model is evaluated and saved every 500 steps on the validation data (in contrast to BLEURT which is evaluated every 1 500 steps (Sellam et al., 2020a)). The full list of hyperparameters is shown in Table 16 in Appendix A.

The training scripts were implemented using the Hugging Face library[9]. The models are trained on a NVIDIA GeForce GTX TITAN X for approximately 37 hours in the case of GBLEURT$_{\mathrm{WWM}}$ and 39 hours in the case of GBLEURT$_{\mathrm{SWM}}$. After training, the best model checkpoint is chosen based on its performance on the validation data. For GBLEURT$_{\mathrm{SWM}}$, the best checkpoint was obtained after 1 000 steps. For GBLEURT$_{\mathrm{WWM}}$, the model reached its best performance after 3 500 steps.

[9] huggingface.co/

The [CLS]-vector on which the regression layer is built has a fixed-width of 512 tokens. In order to avoid that the model is trained on truncated examples, sentence pairs (i.e. the reference and hypothesis together) with more than 512 tokens are filtered out from the training data prior to training. The filtering, though, revealed that none of the sentence pairs in the training corpus exceeded 512 tokens. In contrast, the test set contains such examples. However, to ensure that the model produces predictions for every input example, filtering is not applied at inference time. Rather, sentence pairs exceeding the maximum sequence length are truncated.

4.2.2 GCOMET

Based on the observations by Amrhein and Sennrich (2022), the preliminary analysis of metrics' blind spots focused on COMET-20 as exemplary showcase of neural metrics. As the analysis demonstrated that COMET-20 is not sensitive enough towards mistakes in German compounds, it is of particular interest to investigate whether said sensitivity could be improved when COMET builds on an encoder pre-trained with WWM.

However, GBERT that is available in the two pre-training flavours SWM and WWM is a *monolingual* German model. Consequently, it cannot readily be used with COMET that relies on a *cross-lingual* encoder. To enable the usage of GBERT with COMET, the framework has to be adapted to the monolingual setting. These adaptations are described in the following paragraphs.

In the monolingual setting, the source is discarded from the analysis and the dimensions of COMET have to be reduced accordingly. Such a dimension reduction to adjust COMET to only two instead of three input segments was already implemented by Rei et al. (2020a,b) to obtain a *reference-free* quality estimation metric that considers the source and the hypothesis.

For the purpose of this thesis, the *reference-free* COMET-QE model `wmt20-comet-qe-da` is adapted to the monolingual setting and transformed into a *source-free*, *reference-based* model. As outlined in 4.1.3, the COMET framework relies on three major building blocks. For GCOMET, these blocks were adjusted as follows:

Firstly, the cross-lingual encoder of COMET-QE is replaced with the monolingual GBERT model. Instead of the source, the reference is encoded along with the hypothesis. Like for GBLEURT, two variants are trained based on the two versions of GBERT: GCOMET$_{\text{SWM}}$ and GCOMET$_{\text{WWM}}$.

Secondly, a pooling layer is added that extracts the most relevant linguistic information from each encoder layer and summarises the token embeddings in a single segment-level representation. For GCOMET, the pooling layer of the COMET-QE model is used without modification.

Thirdly, the estimator model is trained in a regression task to predict human quality scores given a reference and an MT hypothesis. In the COMET framework, the input vector to the regression layer is a concatenation of segment-level representations and extracted features. The feature vector for GCOMET is constructed in analogy to that of COMET-QE (cf. Equation 4.5) replacing the source with the reference. That means, that the element-wise reference product $\boldsymbol{h} \odot \boldsymbol{r}$ and the absolute element-wise reference difference $|\boldsymbol{h} - \boldsymbol{r}|$ are extracted and concatenated to the embeddings of the hypothesis $\boldsymbol{h}$ and the reference $\boldsymbol{r}$. This results in the final feature vector:

$$\boldsymbol{x} = [\boldsymbol{h}; \boldsymbol{r}; \boldsymbol{h} \odot \boldsymbol{r}; |\boldsymbol{h} - \boldsymbol{r}|] \tag{4.7}$$

This feature vector is fed into the feed-forward regression layer. Like GBLEURT, the two GCOMET regression models are trained on Direct Assessment scores for the German segments of the data sets from the WMT 2017 - 2019 Metrics Shared Task (cf. 4.1.4). Mean squared error as defined in Equation 4.3 is used as regression loss.

The architecture of the GCOMET models is illustrated in Figure 3.

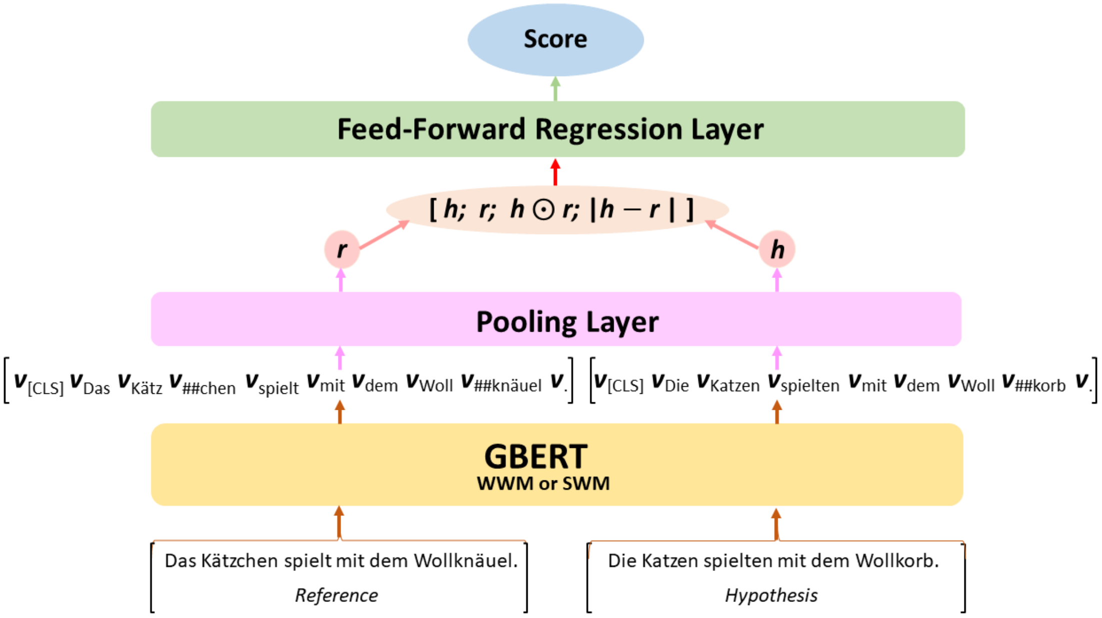
Score
Feed-Forward Regression Layer
[h; r; $h \odot r$; $|h - r|$]
r
h
Pooling Layer
[$v_{[CLS]}$ v_{Das} $v_{Kätz}$ $v_{\#\#chen}$ v_{spielt} v_{mit} v_{dem} v_{Woll} $v_{\#\#knäuel}$ $v_{.}$]
[$v_{[CLS]}$ v_{Die} v_{Katzen} $v_{spielten}$ v_{mit} v_{dem} v_{Woll} $v_{\#\#korb}$ $v_{.}$]
GBERT
WWM or SWM
Das Kätzchen spielt mit dem Wollknäuel.
Reference
Die Katzen spielten mit dem Wollkorb.
Hypothesis

Figure 3: Architecture of GCOMET

To train GCOMET, the same hyperparameters were used as for `wmt20-comet-qe-da` (cf. Rei et al., 2021). The most relevant hyperparameters are summarised in Table 17 in Appendix A. GCOMET is implemented as a PyTorch Lightning[10] model like all other models of the COMET family. The models were trained on a NVIDIA GeForce RTX 3090 for around 2.5 hours. After training, the model with the best performance on the validation data was chosen. For GCOMET$_{\mathrm{SWM}}$ as well as for GCOMET$_{\mathrm{WWM}}$, the best model was obtained after 11 130 training steps.

[10] pytorchlightning.ai

4.2.3 COMET$_{\text{Contrastive}}$

The newly trained GCOMET models are monolingual German models, while the original COMET-20 is a multilingual model. This difference might play a role in the model's sensitivity towards German compounds and is worth to be investigated. However, GCOMET and COMET-20 are not directly comparable. Firstly, COMET-20 is trained on DA scores for *all* languages of the WMT 2017 - 2019 Metrics Shared Tasks, while GCOMET has only seen the *German* portions of these data sets. Secondly, COMET-20 receives additional information from the source segment. Thirdly, the underlying language model of COMET-20, XLM-RoBERTa$_{\text{Large}}$, is much larger than that of GCOMET as shown in Table 9 (cf. Conneau et al. (2020)).

Model	Languages	Layers	Hidden States	Attention Heads	Vocab	Params
GBERT	1	12	768	12	31k	110M
XLM-R$_{\text{Base}}$	100	12	768	12	250k	270M
XLM-R$_{\text{Large}}$	100	24	1024	16	250k	550M

Table 9: Sizes of different language models

Hence, when comparing GCOMET to COMET-20 it would remain unclear to what extent observed differences are owed to the effect of multilinguality and to what degree they must be attributed to the bigger size of the underlying language model, the larger amount of training data and the additional information from the source.

In order to assess the impact of multilinguality on the model's sensitivity to specific linguistic phenomena more reliably, COMET$_{\text{Contrastive}}$ was trained. It is designed to be as comparable as possible to GCOMET, differing only in the chosen encoder.

COMET$_{\text{Contrastive}}$ is based on XLM-RoBERTa$_{\text{Base}}$ (Conneau et al., 2020), which is more comparable in size to GBERT (cf. Table 9). XLM-RoBERTa$_{\text{Base}}$ covers 100 languages and is pre-trained on a clean Common Crawl corpus with a masked language modelling objective using sub-word masking.

COMET$_{\text{Contrastive}}$ is trained with the same hyperparameters and on the same data sets as the two GCOMET models (see Table 17 in Appendix A). The best model was obtained after 18 550 training steps. Like GCOMET, COMET$_{\text{Contrastive}}$ encodes the reference and the MT hypothesis. The source segment is disregarded in order to avoid that the model has access to additional information.

4.2.4 Evaluation

Before the newly trained metrics are deployed in MBR decoding, their performance is evaluated on a standard test set, to verify that they are of reasonable quality. To ensure the comparability with other models, the official evaluation procedure of the WMT 2020

Metrics Shared Task (Mathur et al., 2020) is run.[11] The evaluation measures the system-level Pearson correlation (r) between the predicted DA scores and the actual human scores defined by Mathur et al. (2020) as:

$$r = \frac{\sum_{i=1}^{n}(H_i - \overline{H})(M_i - \overline{M})}{\sqrt{\sum_{i=1}^{n}(H_i - \overline{H})^2}\sqrt{\sum_{i=1}^{n}(M_i - \overline{M})^2}} \tag{4.8}$$

where $\overline{H}$ is the mean human assessment score of all systems in a certain translation direction and $\overline{M}$ is the corresponding mean score predicted by a given metric. H_i and M_i are the scores assigned to the i^{th} MT system by a human and a metric respectively. They are calculated as the average of all segment-level scores for the given i^{th} MT system.

Table 10[12] shows the system-level Pearson correlation of the newly trained metrics on the WMT 2020 Metrics Shared Task test set for en→de. The results are compared to the performance of other metrics briefly discussed in Chapters 2 and 4. The left part of the table reports the correlation over all 14 MT systems for the language pair en→de. The models trained for this thesis perform slightly worse than the previously existing multilingual neural metrics. As already noticed by Sellam et al. (2020b), multilinguality seems to boost correlation of metric models with human judgements. This effect can be observed for GBLEURT and the multilingual mBERT-L2, that is very comparable in size and training method, as well as for GCOMET and COMET$_{\text{Contrastive}}$. The fact that COMET$_{\text{Contrastive}}$ outperforms both GCOMET and GBLEURT underpins the beneficial impact of multilinguality on system-level correlation with human judgements.

However, this effect disappears when outlier systems are excluded from the evaluation. As shown in the right part of Table 10, the strongest monolingual model, GBLEURT$_{\text{WWM}}$ outperforms COMET$_{\text{Contrastive}}$. In addition, the performance gap between the monolingual GBLEURT and GCOMET models and the strong multilingual models BLEURT$_{\text{Extended}}$, mBERT-L2 and COMET-20 shrinks. Moreover, four of the five models trained for this thesis are not significantly outperformed by any other metric when outliers are removed from the analysis. This points towards the ability of the metrics to distinguish between similarly performing MT models. Hence, the results indicate that the five metric models show a very good performance and can be used for further experiments described in the subsequent chapters.

Furthermore, the results clearly indicate that WWM in pre-training is beneficial to the performance of the metric model. The WWM-metrics outperform their SWM-equivalents in both settings, with and without the outlier systems. Hence, pre-training metric models with WWM seems to be a promising approach that might improve the quality of outputs when deployed in MBR decoding.

Finally, the strong results of the lexical metric ChrF and the embedding-based metric

[11] github.com/WMT-Metrics-task/wmt20-metrics

[12] The Table corresponds to Table 6 in Mathur et al. (2020).

		en→de	
		all	-out
		14	11
Lexical	BLEU	0.928	0.825
	sentBLEU	0.934	0.823
	ChrF	**0.962**	**0.862**
	ChrF++	0.958	**0.850**
	YiSi-1	**0.971**	**0.887**
	PRISM	0.958	**0.851**
Neural	ESIM	**0.979**	**0.894**
	BLEURT$_{\text{Extended}}$	**0.969**	**0.870**
	mBERT-L2	**0.970**	**0.861**
	COMET-20	**0.972**	**0.863**
QE	COMET-QE	0.903	0.831
	OpenKiwi-XLMR	**0.968**	0.814
New Neural	GBLEURT$_{\text{WWM}}$	0.951	**0.855**
	GBLEURT$_{\text{SWM}}$	0.932	0.833
	GCOMET$_{\text{WWM}}$	0.942	**0.842**
	GCOMET$_{\text{SWM}}$	0.933	**0.840**
	COMET$_{\text{Contrastive}}$	0.955	**0.844**
	Human	**0.984**	**0.932**

The table shows the system-level Pearson correlation for en→de on the test set of the WMT 2020 Metrics Shared Task. In the column `all`, the output of 14 MT systems was included in the analysis. In the column `-out`, the outliers were removed leaving only 11 MT systems in the analysis. Metrics that were not significantly outperformed by another metric for a given language pair are highlighted in bold.

Table 10: System-level Pearson correlation on the WMT 2020 test set for en→de

YiSi-1 are remarkable. Both outperform all five neural metrics trained for this thesis in both evaluation settings. Despite the often claimed superiority of neural metrics (Kocmi et al., 2021; Freitag et al., 2022a; Fernandes et al., 2022), other metric types might still achieve strong correlations with human judgements.

4.3 MBR Decoding with GBLEURT and GCOMET

In the following sections, a series of experiments based on MBR decoding is run. Thereby, the newly trained metric models are deployed as utility function. The experiments follow the approach proposed by Amrhein and Sennrich (2022) that uses MBR decoding as an instrument to identify blind spots in an evaluation metric.

The main focus of the experiments is the question whether the translation of German compounds improves when the metric model is pre-trained with Whole Word Masking compared to pre-training with Sub-Word Masking. To this end, the same semi-automated analysis of compound translation is conducted as in 3.3. While this method provides a first overview of the frequency of mistranslated compounds as well as insights into different error types, it has certain drawbacks (cf. 3.3.3). To obtain more reliable results regarding the sensitivity of a metric towards mistakes in German compounds, an MBR-based sensitivity analysis is run as suggested by Amrhein and Sennrich (2022). Additionally, the MBR-based experiments are run with COMET$_{\text{Contrastive}}$ as utility function. The comparison between GCOMET$_{\text{SWM}}$ and COMET$_{\text{Contrastive}}$ allows to investigate the effect of monolinguality versus multilinguality on the translation of compounds.

When preparing the experiments, various questions raised regarding the implementation of MBR decoding and the reliability of automatic evaluation metrics. Inspired by these questions, additional experiments were designed to answer them. They are described in the following sections, before the major experiments related to the compound sensitivity of the different metrics are outlined.

4.3.1 Implementation of MBR Decoding

The implementation of the MBR decoding follows the procedure by Amrhein and Sennrich (2022) described in 3.1.4. For each of the 1002 sentences in the en→de test set of the WMT 2021 New Shared Task (cf. 3.1.2), Amrhein and Sennrich (2022) generated 100 unique, unbiased samples as both candidate and support pool. All subsequent experiments are based on these samples to ensure that the results are comparable to theirs. For COMET-20, the MBR implementation of Amrhein and Sennrich (2022) is used as formulated in Equation 3.1.

In contrast to COMET-20, GCOMET, COMET$_{\text{Contrastive}}$ and GBLEURT are source-free metrics that compare a reference to a hypothesis. Hence, MBR decoding is implemented[13] very similarly as in Amrhein and Sennrich (2022), but excluding the source:

$$y^* = \underset{h_i \in C}{\operatorname{argmax}} \frac{1}{|S|} \sum_{h_j \in S} u(h_i, h_j) \tag{4.9}$$

where h_i denotes the i^{th} candidate of the candidate set C, h_j refers to the j^{th} sample in the support set S and y^* is the candidate that maximises the expected utility.

For the purpose of this thesis, the candidate and support sets are identical. However, the MBR implementation would allow them to be different. For the COMET-based models, the embeddings of the hypotheses and pseudo-references are cached and the scores are computed in parallel. For GBLEURT, the embeddings cannot be cached, as the models

[13] See referenceless.py

encode *sentence pairs* rather than individual segments. Hence, the MBR implementation is less efficient than for COMET-based models.

As the candidate and support sets are identical, each candidate is included in the support set. When implementing the MBR decoding for GBLEURT, the question arose whether the candidate should be compared to itself, i.e. to all 100 samples in the support set, or only to the other 99 samples in the support set excluding itself.

As long as MBR decoding focused on lexical overlap-based metrics, this question was not relevant. When comparing two identical segments, lexical metrics assign a perfect score of 1 to them. As each candidate reliably obtains exactly one perfect score, this comparison does not affect the final outcome of the decoding.

However, when using a neural metric as utility function, including or excluding the comparison of the candidate to itself possibly affects the final average scores. The behavior of neural metrics is often unpredictable as they are black box metrics. They do not necessarily assign the same perfect score every time they evaluate two identical segments. In consequence, the average scores could be affected differently for each candidate. As a result, two different candidates could reach the highest average score depending on whether the comparison of the candidate to itself is included or excluded.

To the best of my knowledge, this effect has not yet been investigated. Previous MBR implementations with COMET and BLEURT (Fernandes et al., 2022; Amrhein and Sennrich, 2022; Freitag et al., 2022a) include the comparison of the candidate to itself. To shed light on this question, the MBR decoding for GBLEURT is implemented in two different variants. The first, called MBR$_{100}$, assesses each candidate against the full support set of 100 pseudo-references following Equation 4.9. The second variant (MBR$_{99}$) excludes the comparison of the candidate to itself. The MBR problem is reformulated as:

$$y^* = \underset{h_i \in C}{\operatorname{argmax}} \frac{1}{|S| - 1} \sum_{h_j \in S \setminus \{h_i\}} u(h_i, h_j) \tag{4.10}$$

The results of this experiment are reported in Section 5.1.

4.3.2 Automatic Evaluation of MBR Translation Quality

To ensure that the MBR implementations described in the previous section work properly, an automatic evaluation is run on their output sentences. The same evaluation metrics were chosen as in 3.4. Originally run as a sanity check, the automatic evaluation in 3.4 revealed a phenomenon known as “metric overfitting” (cf. 2.3, Fernandes et al. (2022)). The effect was particularly strong for COMET-20, which assigned overly optimistic scores to MBR outputs produced with itself as utility function. To scrutinize whether this effect is especially strong for COMET-20 or occurs with a similar strength for other neural metrics, GCOMET$_{\text{WWM}}$ is included in this experiment as additional output evaluation metric. The results of the automatic evaluation are shown in Section 5.1.

4.3.3 Combining Metrics in MBR Decoding

The observed metric overfitting inspired an additional experiment. As the findings by Amrhein and Sennrich (2022) and Chapter 3 of this work illustrate, neural metrics have blind spots. Sometimes, they unpredictably assign high scores to bad translations. However, as they fail unpredictably, it is reasonable to assume that different metrics will have difficulties with different segments.

Hence, combining two metrics in MBR decoding is expected to improve the quality of the output on the one hand. On the other hand, it might also alleviate the problem of metric overfitting. As the resulting outputs are only partially optimised towards a given metric, this metric is expected to rate their quality more adequately.

To investigate these hypotheses, COMET-20 and $\text{GCOMET}_{\text{WWM}}$ are combined in MBR decoding. Both metrics exhibit a strong overfitting effect and both are COMET-based models. However, they rely on different masking strategies and encoder models, one being multilingual, the other monolingual. Hence, they are sufficiently different to most likely exhibit distinct pathologies. When combined in MBR decoding, their weaknesses are expected to cancel each other out.

During MBR decoding, each candidate is rated by both metrics and the average score is calculated. If one metric assigns a high score to a bad candidate, that candidate should still receive a low score from the other metric. On average, it probably does not achieve the highest score. Rather, a candidate will be selected that obtains high scores from both metrics. The results of this experiment are reported in 5.1.

4.3.4 Semi-Automated Analysis of Compound Translations in MBR-Decoded Outputs

To gain a first overview over the sensitivity of the different metrics towards German compounds, the MBR output of the newly trained neural metrics is analysed semi-automatically as outlined in 3.3. Hence, unknown words are automatically extracted from the MBR-decoded outputs. Then, the compounds are identified, inspected manually and classified into correct and incorrect translations.

The hypothesis here is that translations obtained with a metric model that relies on WWM contain fewer incorrect compounds than translations generated with a SWM-based utility function, as WWM is expected to render the model more sensitive to incorrect compositions of several sub-words.

As explained in 4.3.1, the MBR decoding with the two GBLEURT models as utility function was implemented in two different flavours: MBR_{99} and MBR_{100}. As the resulting

MBR-decoded outputs differ only minimally between the two variants (cf. 5.1 and 6.1), only the output of MBR_{99} was analysed regarding the translation quality of German compounds.

The reason for this decision is that it is methodologically more correct not to compare a candidate to itself if this comparison does not result in a perfect score of 1 and hence might influence the outcome of MBR decoding. However, this approach stands in contrast to the decision made by Amrhein and Sennrich (2022), who opted for including the comparison of a candidate to itself in their MBR implementation. As I use their implementation for MBR decoding with COMET-based models, it should be kept in mind when interpreting the results that the outputs generated with a COMET-based metric as utility function were obtained with a slightly different approach than those generated with a GBLEURT model. However, in practice, the difference between the two approaches is minimal. Compared to MBR_{100}, MBR_{99} produces only one additional incorrect German compound when used with $\text{GBLEURT}_{\text{WWM}}$ as utility function, and one fewer incorrect compound when used with $\text{GBLEURT}_{\text{SWM}}$.

4.3.5 MBR-Based Sensitivity Analysis

The semi-automated analysis of compound translations in MBR outputs gives us a first glimpse into the sensitivity of the various metrics towards German compounds. However, as mentioned in 3.3.3, this approach has certain limitations. Consequently, the results are not entirely reliable.

In order to establish a more reliable basis for the insights presented in this thesis, an MBR-based sensitivity analysis was conducted, following the approach by Amrhein and Sennrich (2022). Keeping the support set constant, they apply targeted changes to a candidate translation. During MBR decoding, the metric used as utility function scores both the original, correct candidate as well as its perturbed versions. By comparing the assigned scores, one can gain insights about the model's sensitivity towards the linguistic phenomenon under study. If the correct candidate does not obtain a higher score or if the score differences are very small, the model is not sensitive enough towards the studied phenomenon.

To scrutinise the sensitivity of the different neural metrics towards German compounds, a challenge set with 25 examples is composed. The examples are chosen from the test set of the WMT 2021 News Shared Task that was used for MBR decoding with COMET in Amrhein and Sennrich (2022) as well as in this thesis. The 25 examples were selected according to the following criteria that apply cumulatively:

1. The reference translation contains (at least) one German compound, that either corresponds to an English compound in the source segment or to another linguistic construction.

2. MBR with COMET-20 failed to generate the correct German compound. Instead, it chose a candidate with a wrong translation of the compound.
3. Beam search generated a correct German translation of the compound.
4. The correct German compound is found in at least one of the 100 samples in the candidate set.

Cases where all four criteria are fulfilled are of particular interest, as the wrong compound translation can clearly be attributed to COMET-20. The translation model generating the candidates as well as beam search decoding succeeded in producing the correct translation, whereas COMET-20 was unable to recognise it.

To compose the challenge set, the 1-best beam search output containing an accurate translation of the German compound is used as the correct candidate. It is then perturbed replacing the correct compound with an incorrect alternative found in the candidate pool. The perturbations are grouped into two types.

- **Most frequent:** The three *incorrect* alternatives that occur most frequently in the 100 samples of the candidate pool are chosen. If several alternatives occur equally often, one of them is selected randomly as the most frequent.
- **Most similar:** The three *incorrect* alternatives from the candidate pool are chosen that are most similar to the correct compound either in terms of string-based similarity, e.g. *Hundepark* (dog park) and *Hundenpark*, or in terms of semantics, e.g. *Obstbäume* (fruit trees) and *Apfelbäume* (apple trees). The choice of the most similar alternatives is not always well-defined and to some degree subjective.

This results in six different perturbations that are applied to each example in the challenge set. In five examples, the most similar alternatives are also among the most frequent ones. In these cases, only five different perturbations are applied. The perturbations can either be compounds or another linguistic construction, e.g. a normal noun, adjective and noun constructions or genitive constructions. Some examples from the challenge set are shown in Table 11.

Target	Werkzeugkasten	Waldbrand	Feuerwerke	Pfefferspray
Freq 1	Werkzeuge	Wildlandbrand	Schussfeuerwerke	Pfeffersprühen
Freq 2	Toolboxen	Wildwasser	Fireworks	Pfeffersprüh
Freq 3	Werkzeugköpfe	freie Wildbahn	Schussfeuerwerk	Pfeffersprühmittel
Sim 1	Werkzeugkarten	Waldfeuer	Feuerenwerk	Pfefferspritzen
Sim 2	Werkzeugketten	Wildbrand	Feuerwerkszeuge	Pfefferspersay
Sim 3	Werkzeugkorbse	wilden Waldbrand	Feuerstellen	Pfeffersperre

The table shows four examples of correct target compounds that were perturbed in six different ways. The target was replaced with the three most frequent (Freq) incorrect alternatives and the three most similar (Sim) incorrect alternatives.

Table 11: Examples of perturbations applied to the target compound

The correct candidate and the six perturbed variants are then used as candidate set in MBR decoding. In addition, the candidate set contains a copy of the source segment that serves as a reference point to gain an idea of how large the sensitivity scores are if the metric assesses a candidate that is very different from the support set. The support set contains the two German reference translations of the respective segment. Amrhein and Sennrich (2022) found that COMET-20 is more sensitive towards errors in a candidate when used with a high-quality support set. In contrast, when the 100 MT generated samples are used as support, other discrepancies between the candidate and the support than the targeted changes may come into play and affect the results. Thus, the sensitivity scores are more reliable when the references are used as support.

In the course of MBR decoding, the metric under study compares each candidate to the support set and assigns a score to it. The scores of each perturbed candidate is subtracted from the score of the 1-best beam search output. This difference in scores represents the sensitivity of the utility function towards the targeted change. The higher the score, the more sensitive the metric is towards a change. The score differences are then averaged across the different perturbation types. Additionally, they are also averaged across both types of compound perturbation to obtain the general sensitivity to errors in compounds.

In this thesis, the difference between the score of the 1-best beam search output and the perturbed candidates is used as sensitivity score. This stands in contrast to Amrhein and Sennrich (2022), who use the *absolute* score difference as sensitivity scores. However, the absolute difference dilutes the distinction between correct score differences, i.e. the beam search output obtained a higher score than the perturbed candidate, and incorrect ones, i.e. the perturbed candidate was scored higher. To punish wrong assessments, the non-absolute score difference is used as sensitivity score in this thesis.

The sensitivity analysis is run with COMET-20 as well as with the newly trained metrics as utility function. The hypothesis is that metrics relying on a language model pre-trained with WWM are more sensitive towards errors in compounds.

In addition, the sensitivity analysis is also run on the challenge set composed by Amrhein and Sennrich (2022), that includes perturbations of common nouns, named entities and numbers as well as a copy of the source segment and a hallucinated segment. The challenge set includes different types of perturbations. On the one hand, entire nouns, named entities and numbers are substituted. On the other hand, minor changes are applied to the target word by inserting, deleting and substituting single characters.

The sensitivity of the metric models towards targeted changes in common nouns is of particular interest as compounds are a special type of nouns. The comparison with simple nouns sheds light on the question to what extent compounds are blind spots of a neural metric. The hypothesis is that a WWM-based metric reacts more sensitive towards mistakes in common nouns than an SWM-based metric. However, the sensitivity difference is expected to be less pronounced as for compounds. Since simple nouns are typically shorter and more frequent, they are likely to be split up into fewer sub-words or even consist of only one sub-word. In these cases, a SWM-based metric is expected to show a similar sensitivity as a WWM-based metric.

Named entities and numbers are expected to be identical to the entity or number in the reference translation. As each sub-word of the entity or number in the candidate should correspond to the respective sub-word in the reference, it is expected that SWM-based and WWM-based metrics exhibit a similar sensitivity.

Regarding the question of multilinguality, I expect that a monolingual model is better at handling nouns and compounds, as it can dedicate more capacity to the target language than a multilingual model. In contrast, multilinguality is expected to increase the sensitivity towards named entities as these are often transferred unchanged from the source to the target language. Since a multilingual model was also pre-trained on the source language, I assume that it has a better notion of words in that language and reacts more sensitive to changes in named entities. Regarding changes in numbers, I expect that multilinguality does not have a major impact on the model's sensitivity as numbers are the same in English and in German.

5 Results

Having outlined the experimental setup along with the used materials and methods, the results of the various experiments described in Chapter 4 are presented in the subsequent sections.

5.1 Automatic Evaluation of MBR Translation Quality

To ensure that MBR decoding is implemented correctly and that the newly trained metric models are suited as utility function, the MBR-decoded translations are evaluated in terms of various automatic metrics. To enable a meaningful interpretation of the quality scores, they are compared to the scores assigned to beam search translations on the one hand and to MBR-decoded outputs with ChrF and ChrF++ as utility functions on the other hand. The results of this analysis are presented in Table 12. Originally intended as a sanity check, this evaluation revealed several interesting phenomena.

	BLEU	ChrF++	BLEURT-20	COMET-20	GCOMET$_{\text{WWM}}$
Beam Search	**26.220**	**52.961**	**62.841**	**0.2601**	**-1.5052**
MBR ChrF++	20.941	48.330	52.114	-0.1084	-1.6887
MBR ChrF	20.390	48.156	51.825	-0.1225	-1.6966
MBR COMET-20	16.980	44.580	57.281	**0.2440**	-1.5599
MBR COMET$_{\text{Contrastive}}$	15.897	42.604	53.790	0.0007	-1.5729
MBR GCOMET$_{\text{WWM}}$	17.294	44.162	56.607	0.0591	**-1.3982**
MBR GCOMET$_{\text{SWM}}$	16.655	43.606	55.852	0.0387	-1.4847
MBR$_{100}$ GBLEURT$_{\text{WWM}}$	17.454	**44.868**	**57.308**	0.0549	-1.4768
MBR$_{99}$ GBLEURT$_{\text{WWM}}$	**17.468**	44.867	57.307	0.0544	-1.4774
MBR$_{100}$ GBLEURT$_{\text{SWM}}$	17.151	44.039	56.305	0.0074	-1.5233
MBR$_{99}$ GBLEURT$_{\text{SWM}}$	17.130	44.033	56.314	0.0070	-1.5235
MBR Metric Combination	17.249	44.665	**58.351**	0.2027	-1.4312

Table 12: Evaluation of the translation quality in terms of automatic metrics

It stands out from the results that according to all metrics except GCOMET$_{\text{WWM}}$ the beam search translations are of superior quality. The metrics clearly favour beam search translations over MBR-decoded outputs regardless of the used utility function.

Among the tested neural metrics, GBLEURT$_{\text{WWM}}$ performs best as utility function in

MBR decoding receiving the highest scores from three evaluation metrics (BLEU, ChrF++ and BLEURT-20). In contrast, COMET-20 and GCOMET$_{\mathrm{WWM}}$ prefer themselves as utility function which points to an overfitting effect (see below).

Regarding the question of whether the masking strategy in the pre-training affects the performance of a metric, the results indicate that WWM-based utility functions generate better translations than SWM-based utility functions. All five metrics used to evaluate the MBR-decoded outputs agree that the translations of GBLEURT$_{\mathrm{WWM}}$ and GCOMET$_{\mathrm{WWM}}$ are of higher quality than the translations of their SWM-based equivalents.

One of the defined goals of the automatic evaluation is to investigate the question of whether comparing a candidate to itself (MBR$_{100}$) or not (MBR$_{99}$) during MBR decoding affects the final output. The analysis shows that it *does* make a difference, even though not a big one. From the 1002 segments of the test set, only six were translated differently when GBLEURT$_{\mathrm{WWM}}$ was used with MBR$_{100}$ and MBR$_{99}$ respectively. With GBLEURT$_{\mathrm{SWM}}$ as utility function, the two MBR approaches led to a different outcome in only four cases. The results of the automatic evaluation indicate that including the comparison of a candidate to itself leads to slightly better translations.

Moreover, a phenomenon known as “metric overfitting” (Fernandes et al., 2022) can be observed. The two lexical overlap-based metrics BLEU and ChrF++ show a preference for MBR-decoded translations obtained with a utility function of the same type (ChrF and ChrF++), whereas the neural metrics favour translations generated with a neural metric as utility function. BLEURT-20 prefers GBLEURT as utility function. GCOMET$_{\mathrm{WWM}}$ favours itself along with GBLEURT$_{\mathrm{WWM}}$ and GCOMET$_{\mathrm{SWM}}$. COMET-20 shows an extremely strong preference for outputs obtained with itself as utility function, rating all other utility functions as clearly inferior.

To further investigate the phenomenon of metric overfitting, GCOMET$_{\mathrm{WWM}}$ and COMET-20 were combined in MBR decoding by averaging their scores. The last line of Table 12 presents the results of this novel approach. They indicate that combining two utility functions tends to improve the quality of the resulting translations. Especially BLEURT-20 clearly favours the outputs of the metric combination over all other MBR-decoded outputs. COMET-20, that exhibits a strong overfitting effect, assigns a surprisingly high score to the outputs obtained with the metric combination. Nonetheless, it still favours the translations obtained with itself as utility function. The same behaviour is observed for GCOMET$_{\mathrm{WWM}}$. Even though its overfitting effect is less pronounced than that of COMET-20, it still prefers itself as utility function, but assigns the second highest score to MBR translations obtained with a metric combination.

Focusing on the original goal of the automatic evaluation, the results demonstrate that except for COMET-20 the metrics agree that MBR decoding with the new metrics as utility functions generates outputs that are of reasonable quality. Their quality is comparable to that of MBR$_{\mathrm{COMET\text{-}20}}$-decoded translations and somewhat inferior to beam search. Overall both the MBR implementation and the new metrics passed the sanity check.

5.2 Semi-Automated Analysis of Compound Translations in MBR-Decoded Outputs

Table 13 presents the results of the semi-automated analysis of the quality of compound translations in the 1002 test sentences. Thereby, the neural utility functions are compared to the lexical metrics ChrF and ChrF++. Additionally, a comparison is made between the MBR-decoded outputs and beam search translations.

	Unknown Words	New Compounds	Mistranslated Compounds
Beam Search	**461**	**374**	**241**
MBR ChrF++	790	476	386
MBR ChrF	844	511	415
MBR COMET-20	855	553	447
MBR COMET$_{\text{Contrastive}}$	904	550	454
MBR GCOMET$_{\text{WWM}}$	752	494	369
MBR GCOMET$_{\text{SWM}}$	738	476	379
MBR GBLEURT$_{\text{WWM}}$	729	464	352
MBR GBLEURT$_{\text{SWM}}$	**686**	**438**	**335**

Table 13: Unknown words, newly formed compounds and mistranslated compounds found in the output sentences of different decoding strategies

As already noticed in 3.4, MBR decoding with COMET-20 results in a higher number of unknown words and hence of incorrect compounds in the output than MBR decoding with ChrF and ChrF++. This effect is even amplified when its smaller, source-free variant, COMET$_{\text{Contrastive}}$ is used as utility function.

In contrast, the monolingual GCOMET and GBLEURT metrics show improvements over COMET. They reduce the amount of unknown words in the outputs considerably. Consequently, their translations also contain fewer incorrect compounds. They reduce the number of ill-formed compounds by 16.5 % to 26.2 % compared to COMET$_{\text{Contrastive}}$. Moreover, GCOMET and GBLEURT outperform ChrF and ChrF++ in terms of unknown words and mistranslated compounds. Overall, GBLEURT performs better as utility function in MBR than GCOMET producing less unknown words and incorrect compounds.

When comparing the WWM-based metrics to the SWM-based metrics, no clear conclusion can be drawn. For both GCOMET and GBLEURT, the SWM variant generates less unknown words in the outputs than the respective WWM equivalent. However, the picture is different when looking at the incorrectly formed compounds. Translations decoded with GCOMET$_{\text{WWM}}$ contain less mistranslated compounds than those obtained with GCOMET$_{\text{SWM}}$. The opposite is the case for GBLEURT. Translations decoded with GBLEURT$_{\text{WWM}}$ exhibit considerably more mistranslated compounds than those obtained with $\text{GBLEURT}_{\text{SWM}}$.

Finally, the results highlight that beam search generates by far less unknown words and consequently fewer incorrect compounds than all variants of MBR decoding underpinning the findings from 3.4. The novel GCOMET and GBLEURT metrics alleviate the problem of unknown words and ill-formed compounds in translation outputs, but they do not reach the performance of beam search.

5.3 MBR-Based Sensitivity Analysis

Finally, a targeted sensitivity analysis using sampling-based MBR decoding was conducted to analyse and quantify the sensitivity of the various metrics towards errors in German compounds more reliably. The results are presented in Table 14.

	COMET-20	COMET-Contrastive	GCOMET-WWM	GCOMET-SWM	GBLEURT-WWM	GBLEURT-SWM
most frequent	0.132	0.098	**0.167**	0.139	**0.143**	0.132
most similar	0.086	0.091	**0.159**	0.135	**0.152**	0.121
compounds total	0.109	0.095	**0.163**	0.137	**0.148**	0.126
copy	1.429	1.148	**2.614**	2.387	**2.608**	1.871

The sensitivity scores are calculated as the average difference between the MBR score of the 1-best beam search output and the perturbed candidates.

Table 14: Sensitivity scores of different metrics towards errors in compounds

The results consistently indicate that metrics based on a language model that was pre-trained with WWM are clearly more sensitively to errors in compounds that SWM-based metrics. This finding points to the value of using WWM in pre-training.

As expected, $\text{COMET}_{\text{Contrastive}}$ is generally less sensitive to mistakes in compounds than its larger sibling COMET-20. However, it reacts more sensitive to substitutions of a compound with similar alternatives than COMET-20.

The sensitivity gains of GCOMET over $\text{COMET}_{\text{Contrastive}}$ are substantial. Both monolingual GCOMET models even outperform the much larger multilingual COMET-20 model in terms of compound sensitivity. The sensitivity scores of the GBLEURT models cannot be directly compared to the scores of the COMET-based models as they belong to a different metric family.

Regarding the two types of perturbations, the results reveal that all metrics except for $\text{GBLEURT}_{\text{WWM}}$ react more sensitive, if the compound is replaced with a frequent variant than if it is substituted with a similar variant.

The sensitivity analysis performed with regard to common nouns, named entities (NEs) and numbers sheds more light on the strengths and weaknesses of the different metric models. The analysis on the challenge set by Amrhein and Sennrich (2022) is of particular interest as not only entire words are replaced but also minor changes are applied to the target word or number. The results are summarised in Table 15.

	COMET-20	COMET-	GCOMET-	GCOMET-	GBLEURT-	GBLEURT-
		Contrastive	WWM	SWM	WWM	SWM
noun_{add}	0.257	0.172	0.385	**0.453**	**0.380**	0.331
noun_{del}	0.215	0.126	0.304	**0.342**	**0.291**	0.265
noun_{sub}	0.295	0.181	0.392	**0.456**	**0.388**	0.338
$\text{noun}_{\text{whole}}$	**0.509**	0.376	0.471	0.429	**0.528**	0.366
NE_{add}	0.108	0.071	0.148	**0.173**	0.203	**0.227**
NE_{del}	0.078	0.050	0.093	**0.097**	0.140	**0.151**
NE_{sub}	0.113	0.059	0.139	**0.157**	0.192	**0.211**
NE_{whole}	0.173	0.133	**0.192**	0.164	**0.354**	0.317
num_{add}	0.057	0.038	**0.120**	0.117	0.117	**0.120**
num_{del}	0.063	0.030	**0.067**	0.049	0.084	**0.091**
num_{sub}	0.019	0.016	**0.051**	0.048	0.053	**0.067**
$\text{num}_{\text{whole}}$	0.079	0.048	**0.128**	0.126	0.123	**0.126**
hallucination	2.055	**2.620**	2.192	1.840	**2.097**	1.606
copy	1.350	1.024	**2.578**	2.458	**2.466**	1.696

The sensitivity scores are calculated as the average difference between the MBR score of the 1-best beam search output and the perturbed candidates.

Table 15: Sensitivity scores of different metrics towards errors in common nouns, named entities and numbers

Most results from the sensitivity analysis of compounds are confirmed. Throughout all analysed cases, $\text{COMET}_{\text{Contrastive}}$ is less sensitive to mistakes than the larger COMET-20. Both models are clearly outperformed by GCOMET with two exceptions. COMET-20 is surprisingly sensitive to substitutions of entire nouns and $\text{COMET}_{\text{Contrastive}}$ reacts very sensitively to hallucinations.

The contrastive examples by Amrhein and Sennrich (2022) allow for some interesting insights into the effects of WWM and SWM on the performance of metric models. While $\text{GCOMET}_{\text{WWM}}$ is more sensitive to the replacement of entire nouns and NEs, $\text{GCOMET}_{\text{SWM}}$ exhibits an increased sensitivity towards minor changes in nouns and NEs. Interestingly, this effect cannot be observed for numbers where $\text{GCOMET}_{\text{WWM}}$ consistently shows a higher sensitivity.

It is noteworthy that GBLEURT behaves differently. For NEs, the same effect can be observed as for GCOMET: $\text{GBLEURT}_{\text{WWM}}$ is more sensitive to replacements of an entire NE, while $\text{GBLEURT}_{\text{SWM}}$ reacts more sensitively to minor changes. However, with regard to nouns, $\text{GBLEURT}_{\text{WWM}}$ consistently exhibits a higher sensitivity to all types of changes than $\text{GBLEURT}_{\text{SWM}}$. In contrast, regarding numbers, it is $\text{GBLEURT}_{\text{SWM}}$ that is consistently more sensitive to mistakes.

Moreover, the results show that all metrics except for GCOMET$_{\text{SWM}}$ are more sensitive to the replacement of an entire word or number than to adding, deleting or substituting a single character.

Further, the analysis reveals that multilingual models are clearly more sensitive to hallucinations than to copied segments, whereas monolingual models react stronger to copies than to hallucinations.

When comparing the sensitivities of a given metric towards different linguistic phenomena, various conclusions can be drawn. Firstly, all metrics are consistently more sensitive to errors in common nouns than to mistakes in compounds, named entities or numbers. However, for the GBLEURT models, the sensitivity gap betweeen common nouns, named enitites and compounds are not as pronounced as for the GCOMET models.

Secondly, when comparing the sensitivities towards compounds and NEs, two opposing tendencies are observed. COMET-20, COMET$_{\text{Contrastive}}$ and GCOMET$_{\text{WWM}}$ generally react more sensitive to errors in compounds than to minor changes in NEs. However, replacing an entire NE causes a stronger response. In contrast, GCOMET$_{\text{SWM}}$, GBLEURT$_{\text{WWM}}$ and GBLEURT$_{\text{SWM}}$ show a greater sensitivity towards changes in NEs than in compounds.

Thirdly, the sensitivity of all metrics towards changes in numbers is very low. The metrics are by far less sensitive to errors in numbers than to mistakes in NEs or compounds.

6 Discussion

In this chapter, the most relevant insights from the conducted experiments will be discussed, analysed and interpreted. Moreover, the formulated hypotheses are reflected upon, analysing to what extent they have proven to be accurate or inaccurate.

6.1 Automatic Evaluation of MBR Translation Quality

The quality evaluation in terms of established automatic metrics served to analyse the effect of MBR_{100} compared to MBR_{99}, when a neural metric is used as utility function. The analysis revealed that including or excluding a copy of the candidate in the support set indeed influences the outcome of MBR decoding. This finding underlines that neural metrics sometimes behave unpredictably. They do not always assign the same score when assessing two identical segments. Rather, the assigned score varies from segment to segment.

In practice, this effect might be negligible as the cases where different candidates are chosen as best translation are relatively rare and might not be relevant when MBR decoding is applied to a large set of segments. The differences in the final translation quality are only minimal as the results in Table 12 show. Nonetheless, it is important to keep the uncontrollable behaviour of neural metrics in mind as they become more and more widespread in evaluation and MBR decoding.

In addition, the results indicate that MBR_{100} leads to slightly better translations. This is somewhat surprising as MBR_{99} is methodologically more correct. One might even expect that comparing a given segment to itself and assigning a score that depends on the segment might distort the results. Why including this comparison nevertheless improves the outcome, remains unclear.

Moreover, the results demonstrate the WWM in the pre-training of the language model underlying a certain metric has a positive impact on the metrics performance. The beneficial effect of WWM goes beyond compound translation and seems to be of a general nature. This finding is further supported by the results in Table 10 which demonstrate that WWM-based metrics reach a higher Pearson correlation in system-level rankings than their SWM-based counterparts.

Further, the results of the automatic evaluation suggest that $\text{GBLEURT}_{\text{WWM}}$ is best suited as utility function in MBR decoding as it overall generates translations of the highest quality. According to four of the five applied evaluation metrics, GBLEURT$_{\mathrm{WWM}}$ even outperforms COMET-20 as utility function. This finding is surprising as COMET-20 is considered to be state-of-the-art (Kocmi et al., 2021). It might be an indication that COMET-20 does not only have difficulties with compounds, but also with other peculiarities of the German language. However, this supposition requires further analysis (cf. 6.3).

The good performance of GBLEURT$_{\mathrm{WWM}}$ regarding translation quality is in line with its high Pearson correlation in system-level rankings (cf. Table 10). Nonetheless, as discussed in Chapter 2, quality assessments obtained with automatic metrics are not always reliable and should be taken with a grain of salt. Hence, further analysis is needed to confirm that GBLEURT$_{\mathrm{WWM}}$ is indeed the best neural metric model scrutinised in this thesis. Its seemingly good overall performance does not guarantee that it performs similarly well regarding the handling of specific linguistic phenomena. This question is further discussed in 6.2 and 6.3.

Furthermore, four of the five evaluation metrics favour beam search translations over any MBR-decoded outputs. The reasons for this clear preference might be two-fold. On the one hand, a manual, though not systematic, inspection of the translations arrives at the conclusion that the beam search outputs are generally of a higher quality than the MBR-decoded translations. On the other hand, the observed preference of neural evaluation metrics might be amplified by a bias in the training procedure, as they are typically trained on beam search translations. The fact that beam search seeks to match the surface form of the reference (Freitag et al., 2022a), might even enhance this effect, since neural metrics still rely on surface overlap (Amrhein et al., 2022). Especially in the case of COMET-20 such a bias is well possible, as it exhibits a strong preference for beam search, while clearly disliking MBR-decoded translations generated with any utility function other than itself. In contrast to the other neural metrics, GCOMET$_{\mathrm{WWM}}$ does not favour beam search, even though it was trained on the same data sets. Rather, other factors seem to control its behaviour (see below).

Finally, the results of the automatic evaluation corroborate the problem of “metric overfitting”. Fernandes et al. (2022) observed that when a fine-tuned metric is used as utility function and hence optimised for in MBR decoding, this metric does no longer reliably evaluate the resulting translations. Rather, it favours translations generated with itself as utility function regardless of their actual quality.

Interestingly, the overfitting effect can also be observed with non-neural, lexical metrics. When excluding beam search outputs from the analysis, ChrF++ assigns the highest scores to translations obtained with itself as utility function and the second highest scores to those generated with the closely related utility function ChrF. The overfitting effect even exists across different metrics belonging to the same group. BLEU prefers MBR outputs obtained with the lexical metrics ChrF++ and ChrF that pursue the same objective of maximising the surface overlap between the translations and the reference. BLEURT-20, on the other hand, considers translations produced with GBLEURT$_{\text{WWM}}$ as best. Even though the two metrics are completely independent, they share certain similarities: 1) They both rely on a BERT model and 2) their regression layer was trained in the same way. Possibly, these similarities are already enough for BLEURT-20 to show a slight bias towards translations obtained with GBLEURT$_{\text{WWM}}$. On the other hand, it is possible that the GBLEURT$_{\text{WWM}}$ translations are indeed the best. As BLEURT-20 is the evaluation metric that is most dissimilar from and most independent of the metrics used as utility functions, it produces the most reliable scores that coincide best with the manual inspection: Beam search clearly produced the best translations while the quality of the MBR-decoded outputs with different utility functions is very similar. Generally, WWM-based utility functions generate slightly better translations, while the outputs of the lexical utility functions ChrF and ChrF++ are of slightly lower quality.

The overfitting effect is most pronounced for COMET-20 that strongly favours beam search outputs and translations obtained with itself as utility function, while considering the outputs with all other utility functions as poor translations. However, a manual inspection of the translations identifies only minor quality differences between the outputs of the various utility functions. Hence, the low scores assigned by COMET-20 are not justified. The observed overfitting by far exceeds the one reported by Fernandes et al. (2022). This extreme discrepancy between the assigned scores and the actual translation quality is concerning as COMET-20 is currently considered the state-of-the-art evaluation metric and is recommended as default metric (Kocmi et al., 2021). To better understand its behaviour and to investigate whether other models of the COMET family behave similarly, the automatic evaluation is additionally run with GCOMET$_{\text{WWM}}$. As expected, GCOMET$_{\text{WWM}}$ shows a relatively strong overfitting effect as well. However, this effect is less pronounced as for COMET-20.

The quality scores assigned by GCOMET$_{\text{WWM}}$ shed light on further noteworthy issues. Surprisingly, in contrast to all other metrics, GCOMET$_{\text{WWM}}$ does not assign the highest score to beam search translations. Instead, it prefers MBR-decoded translations obtained with a utility function with which it shares common features. Apart from itself, GCOMET$_{\text{WWM}}$ considers GBLEURT$_{\text{WWM}}$ as the best utility function. This metric is built upon exactly the same language model, GBERT$_{\text{WWM}}$. In the ranking, GCOMET$_{\text{SWM}}$ follows that relies on a different, yet similar language model, GBERT$_{\text{SWM}}$, and whose regression layer is trained in the same fashion as that of GCOMET$_{\text{WWM}}$.

Hence, it can be concluded that it is mainly the underlying language model that drives the metric's decisions and preferences. Apparently, the embeddings provided by the language model contain crucial information on which the metric bases its decisions. How the regression layer is trained and what features are extracted influences the model's preferences as well. A similar effect is observed for COMET-20: It prefers the GCOMET models, whose regression layer is trained in a different, yet similar way, over their GBLEURT equivalents.

Finally, COMET$_{\text{Contrastive}}$ is deemed to produce the poorest translations. Surprisingly, also COMET-20 arrives at this conclusion, even though COMET$_{\text{Contrastive}}$ is the metric that is most similar to it. It seems that MBR translations of COMET$_{\text{Contrastive}}$ are of lower quality than that of other utility functions. It is clearly outperformed by GCOMET and GBLEURT that are very comparable to it in size and training procedure. Further, GBLEURT$_{\text{WWM}}$ even outperforms the much larger COMET-20 model according to four evaluation metrics, while GCOMET$_{\text{WWM}}$ shows a similar performance as COMET-20. These results suggest that monolingual utility functions are superior to multilingual ones, especially when based on WWM. The question of monolinguality versus multilinguality is further discussed in 6.2 and 6.3.

All these findings highlight the difficulty and simultaneously the importance of finding a reliable automatic evaluation method. The analysis uncovered that the effects of metric overfitting are more subtle and severe than assumed, as overfitting does not only happen for a given metric, but was rather observed across metrics that share certain similarities, such as building on the same or a similar language model, having a similar architecture or relying on surface similarities to calculate a score.

This raises the question of how to reliably assess the translation quality of MBR-decoded outputs. Certainly, it is crucial that future research is aware of the interplay between building blocks of related metrics. To circumvent or at least reduce distorting effects in the evaluation of MBR outputs, it is best to choose a metric that is as distinct as possible from the deployed utility function. That is, a metric that builds on a language model that uses a different architecture and different training corpora than the underlying language model of the utility function. Further, the regression layer of the evaluation metric should be trained in a different fashion, extracting other features, than that of the utility function. Ideally, the evaluation metric and the utility function are trained on different data set. Today, this is usually not the case as human assessment data is scarce. Therefore, most neural metrics are trained on the WMT Metrics Shared Task data. However, using only a handful of data sets to train various metric models might introduce biases and further distorting effects in the evaluation procedure.

One possibility to alleviate the overfitting and its distorting effects might be to combine different metrics in MBR decoding. In addition, this strategy might also overcome the unpredictable choices of neural metrics, as a candidate is only chosen as best option if it receives high scores from both metrics. The results in Table 12 indicate that combining the scores of two metrics in MBR decoding indeed improves the translation quality. Especially BLEURT-20 identifies a clear quality improvement. A manual inspection confirms that the translations chosen by the combination of two metrics are generally more adequate than those selected by a single metric.

Moreover, this novel approach seems to alleviate the overfitting effect. COMET-20 assigns a surprisingly high score to the translations obtained with the metric combination. This score is much more appropriate than the very low scores assigned to all other MBR outputs.

Similarly, $\text{GCOMET}_{\text{WWM}}$ considers the translations of the metric combination as the second best after those obtained with itself as utility function. Hence, the two metrics captured the actual translation quality much more accurately than before. However, one should keep in mind that the MBR decoding optimised towards the two metrics that were later used for evaluation. Hence, the relatively high scores are probably still owed to an overfitting effect, even though an alleviated one. Nonetheless, the combination of two metrics in MBR decoding, and possibly also in the evaluation of translation quality, is a promising approach for future work as it improves translation quality and alleviates the problem of metric overfitting at least to some degree.

6.2 Semi-Automated Analysis of Compound Translations in MBR-Decoded Outputs

The semi-automated analysis of translation outputs revealed that COMET-20 and its smaller variant $\text{COMET}_{\text{Contrastive}}$ perform especially poor regarding German compounds. They select candidate translations with particularly many unknown words and incorrectly formed compounds. The two COMET metrics are not only outperformed by GCOMET and GBLEURT, but also by the lexical metrics ChrF and ChrF++ that are generally deemed to be less accurate than fine-tuned neural metrics (Kocmi et al., 2021; Fernandes et al., 2022; Freitag et al., 2021a,b). This is a strong indication that German compounds are indeed a blind spot of COMET.

The clear superiority of GCOMET and GBLEURT points to an advantage of monolingual models over multilingual ones regarding the translation of German compounds. However, as the semi-automated analysis suffers from some limitations (cf. 3.3.3), the results presented here are not entirely reliable. Therefore, the role of multilinguality is further investigated in the sensitivity analysis in 6.3.

Regarding the effect of the masking strategy, the expectation was that metrics based on WWM are more sensitive to incorrect compositions of word parts leading to a reduced number of mistranslated compounds in their MBR translations. However, this is only partially the case. Surprisingly, the translations obtained with $\text{GCOMET}_{\text{SWM}}$ and $\text{GBLEURT}_{\text{SWM}}$ as utility functions contain clearly fewer unknown words and hence fewer new compounds than those of their WWM-based counterparts. Regarding the correctness of these compounds, the results are ambiguous. In the case of GCOMET, the WWM-based variant generates fewer incorrect compounds, while for GBLEURT, the SWM-based model performs better.

This outcome is unexpected. Nevertheless, it does not necessarily mean that WWM-based metrics are unable to improve compound translation. Rather,the limitations of the chosen approach may have influenced the results. Hence, it is possible that $\text{GBLEURT}_{\text{SWM}}$ prefers translation candidates where only one constituent of a compound was translated. Such a translation consists only of a common, probably known noun instead of a compound and does not appear in the statistics. It is also possible that GBLEURT$_{\mathrm{SWM}}$ favours other linguistic constructions over composition, e.g. genitive constructions, adjective + noun etc. In these cases the words in the translation are typically known, not appearing in the statistics presented in Table 13. Thus, further analysis is required to draw an informed conclusion on the effect of WWM versus SWM (cf. 6.3).

Furthermore, the results suggest that GBLEURT-decoded translations generally contain fewer unknown words and hence fewer incorrect compounds than GCOMET-decoded translations. This is in line with the results from the automatic evaluation that concedes an advantage to GBLEURT over GCOMET. However, due to the limitations of the semi-automated analysis one should not jump to conclusions about the superiority of GBLEURT. Rather, one should only conclude that the architecture of a metric model might influence its sensitivity towards specific linguistic phenomena.

6.3 MBR-Based Sensitivity Analysis

The sensitivity analysis sheds light on various intriguing issues. Firstly, the results regarding the sensitivity of the different metrics towards German compounds will be discussed. The second part of this section addresses the metrics' sensitivity towards common nouns, named entities and numbers.

The sensitivity analysis allows insights into the core question of this thesis. The two models that are built on top of a language model pretrained with Whole Word Masking are substantially more sensitive to both analysed error types in compounds than their SWM-based counterparts. Hence, the hypothesis is confirmed. WWM in the pre-training enforces the model's ability to attend to an entire word and fosters its understanding of the relations between sub-words. Thus, its sensitivity towards wrong compositions of different word parts is increased. In conclusion, Whole Word Masking in pre-training can indeed alleviate certain blind spots of a fine-tuned metric.

Moreover, the results in Table 14 illustrate that both COMET-20 and its smaller sibling COMET$_{\mathrm{Contrastive}}$ exhibit a very low sensitivity to German compounds. They are substantially outperformed by GCOMET$_{\mathrm{WWM}}$ *and* GCOMET$_{\mathrm{SWM}}$. As GCOMET$_{\mathrm{SWM}}$ is pre-trained with Sub-Word Masking, one expects that the much larger COMET-20 would outperform it. That this is not the case demonstrates that German compounds are indeed a blind spot of COMET-20.

This finding further suggests that the low compound sensitivity of COMET-20 is not only a matter of the masking strategy, but rather an issue of the underlying language model. In contrast to GBERT, that forms the basis of GCOMET, the underlying language model of COMET-20, XLM-RoBERTa$_{\mathrm{Large}}$, is multilingual. The results indicate that the multilinguality has a negative impact on the sensitivity towards nominal composition in German. Regarding language specific linguistic phenomena, a monolingual language model might have an advantage over a multilingual one, as it can devote more capacity to idiosyncratic peculiarities of a given language, whereas a multilingual model has to generalize well across various languages.

Moreover, the results demonstrate that monolingual models are more sensitive to copies of the source sentence. This is not surprising, as they expect exclusively German input, while the multilingual models are accustomed to different languages. Nonetheless, the multilingual metrics clearly distinguish between different languages and discard candidate translations that are copies of the source. Interestingly, the WWM-based models are even more sensitive towards copies than their SWM-based equivalents. The increased awareness of the relation between various sub-words might be beneficial to distinguish between different languages. However, this is only a supposition that requires further research.

Finally, the perturbation type affects the sensitivity of a given model. Except for GBLEURT$_{\text{WWM}}$, the metrics are more sensitive to replacements of a compound with a frequent alternative than with a similar alternative. This is the expected outcome, as it is more difficult to differentiate between similar strings or meanings than between frequent alternatives. Why GBLEURT$_{\text{WWM}}$ behaves differently remains unclear. However, it should be noted that the distinction between the most frequent and the most similar alternatives is not always clear-cut. It is to some degree a matter of subjectivity which alternatives are regarded as the most similar ones. In addition, in some cases, the most similar alternatives are also among the most frequent ones. Hence, the distinct behavior of GBLEURT$_{\text{WWM}}$ does not necessarily hint at a meaningful difference and might rather occur haphazardly.

Further insights are gained from the sensitivity analysis of nouns, named entities and numbers. It stands out from the results in Table 15 that COMET$_{\text{Contrastive}}$ as well as the much larger COMET-20 are clearly outperformed by the two GCOMET models in most of the tested settings. This further supports the supposition that multilinguality has a detrimental impact on the metric's sensitivity towards specific linguistic phenomena. Interestingly, monolingual models are not only superior in handling language-specific phenomena, such as compounds, but also in regard to language-independent phenomena, such as numbers that can be copied from the English source to the German translation. Many named entities can be copied from the source as well (e.g. names of persons), while others need to be translated (e.g. certain names of cities or countries). That multilingual metrics are little sensitive to discrepancies between the hypothesis, the reference and the source is surprising, especially in the case of COMET-20. In contrast to the other metrics, COMET-20 additionally receives the source segment as input. Hence, it can assess each hypothesis twice comparing the numbers and NEs against the reference *and* the source. That it nonetheless fails to detect such discrepancies reliably, is astonishing and underpins that it suffers from unexpected blind spots.

When investigating the effect of WWM, the results are inconsistent. In the case of

GCOMET, the SWM-based model reacts more sensitive to minor changes in nouns and NEs, while the WWM-based model is more sensitive to replacements of entire words. This behaviour coincides with the observations made by Dai et al. (2022). In their experiments on grammatical error correction in Chinese, they found that SWM-based models perform better when a single character had to be corrected, whereas WWM-based models showed the stronger performance when two or more characters had to be replaced. However, this effect is not observed for the other metrics that are all more sensitive to replacements of entire words than to perturbations of single characters. Hence, the behaviour of GCOMET is probably a coincidence rather than a systematic characteristic of WWM- versus SWM-based metrics.

Overall, no clear winner can be identified. Both masking strategies perform similarly, one having the edge on the other in some cases, the other being superior in other cases depending on the architecture of the metric and the phenomenon under study.

As discussed above, multilingual models are less sensitive to copies of the source than monolingual models. However, they react strongly to hallucinations, with $\text{COMET}_{\text{Contrastive}}$ showing the greatest sensitivity. For GCOMET and GBLEURT, the WWM-based models show a stronger reaction than the SWM-based ones. Hence, both multilinguality as well as WWM seem to foster the detection of hallucinations.

The comparison of the metrics' sensitivity towards different linguistic phenomena reveals that compounds as well as NEs and numbers are very clearly blind spots of *all* neural metrics studied in this thesis. Their sensitivity towards said phenomena is substantially lower than their sensitivity towards common nouns. Monolingual models that were pre-trained with WWM can alleviate this problem to some degree, but they do not solve it. Nonetheless, the sensitivity gains may already lead to important improvements in the metric's performance.

That the sensitivity gap between nouns on the one hand and NEs and compounds on the other hand is so pronounced, is to some extent surprising as compounds and NEs are special kinds of nouns. However, their morphological peculiarities seem to be sufficient to trigger certain failures in the metrics. Said sensitivity gap is larger for the GCOMET models than for GBLEURT. This is an indication that not only the masking strategy, but also the architecture of the regression layer has an impact on the sensitivity.

Finally, numbers are the most pronounced blind spots of *all* analysed metrics. The metrics are particularly insensitive to deletions and substitutions of single digits, even though these operations modify the meaning of a number significantly. The effect of these operations is unambiguously more severe for numbers than for words. Why fine-tuned metrics have such great difficulties in handling numbers remains unclear. Perhaps, the training data of the metrics or the underlying language models contain samples with mismatching numbers leading to a blind spot in the metric. Future research is needed to identify the reasons of the blind spots.

6.4 Synthesis

Having discussed the results of the individual experiments, the most important findings are synthesised across the various examinations. This section focuses on the major insights that are discussed form a more holistic point of view.

6.4.1 Whole Word Masking versus Sub-Word Masking

Regarding the core question of this thesis, several insights were gained across the various experiments. First of all, the hypothesis that WWM increases the metric's sensitivity towards mistakes in compounds was confirmed. However, it should be noted that an even larger sensitivity gain is achieved when using a monolingual language model instead of a multilingual one (cf. 6.4.2). Nonetheless, WWM further boosts the sensitivity towards compounds.

With respect to other linguistic phenomena, such as common nouns, named entities and numbers, both masking strategies perform similarly. However, the results demonstrate that metrics relying on WWM are generally more sensitive to replacements of entire words or numbers than metrics based on SWM. Further, WWM enhances the sensitivity towards copied segments and hallucinations.

When used as utility function in MBR decoding, WWM does *not* reduce the amount of unknown words in the generated translations. On the contrary, SWM is more beneficial in decreasing the number of unknown output words. Regardless of this effect, the automatic evaluation arrives at the conclusion that translations obtained with WWM-based metrics are overall of higher quality. Especially the three neural evaluation metrics detect a quality difference between translations of WWM-based and SWM-based utility functions.

Taken together, these findings suggest that WWM-based models might not produce fewer errors, but *less severe* errors. They react more sensitively to major mistakes, like wrong words, copies and hallucinations, that affect the meaning of a sentence strongly and severely hinder the understandability of a translation. In contrast, they punish minor errors, like replacements of single characters, less, as these types of errors have less severe consequences for the intelligibility of a sentence. Hence, they treat mistakes with more human-like priorities than SWM-based metrics.

In conclusion, the results indicate that WWM has a positive impact on the overall performance of a given metric, making it behave more expectedly. Therefore, I recommend to base future metrics on language models that were pre-trained with WWM.

6.4.2 Monolinguality versus Multilinguality

The experiments revealed that not only the masking strategy influences the sensitivity of a metric, but also the nature of the underlying language model. The monolingual German metrics consistently outperformed the multilingual metrics across all conducted experiments. They are not only more sensitive towards mistakes in nouns and compounds, but also to errors in named entities and numbers, where multilinguality is expected to be an advantage. Particularly in the case of COMET-20, the low sensitivity is surprising, as it has access to additional information from the source segment. That it nevertheless fails at identifying wrong numbers and NEs, underpins that COMET-20 has blind spots that cannot be easily explained.

In addition, monolingual metrics also output fewer unknown words when used as utility function and their MBR-decoded translations are generally of a higher quality than those of multilingual utility functions. GBLEURT and GCOMET clearly outperform $\text{COMET}_{\text{Contrastive}}$ in terms of translation quality. $\text{GBLEURT}_{\text{WWM}}$ even outperforms COMET-20, that is much larger and hence expected to show a superior performance, while $\text{GLBUERT}_{\text{SWM}}$ and the two GCOMET models achieve a very similar quality as COMET-20, when deployed as utility function.

An explanation for the superiority of monolingual metrics is the often cited *curse of multilinguality* (Conneau and Lample, 2019; Conneau et al., 2020; Pu et al., 2021). In multilingual models, various languages share the capacity of the model. While multilingual models with only few languages usually outperform monolingual models, adding too many languages is detrimental to the performance. After a certain point, a capacity bottleneck is reached, where the available capacity per language is no longer large enough. As a result, the performance decreases. As XLM-RoBERTa, that underlies COMET-20 and $\text{COMET}_{\text{Contrastive}}$, is a massively multilingual model with 100 languages, this bottleneck seems to be reached.

Moreover, multilingual models typically improve the performance on low-resource languages, while deteriorating the performance on high-resource languages (Conneau et al., 2020). As German is a high-resource language, the multilinguality of XLM-RoBERTa is a disadvantage with harmful effects on the model's performance. This is in line with the finding by Amrhein et al. (2022) that multilingual embeddings can have a negative impact on the performance of the metric when evaluating MT output.

The findings highlight the value of monolingual models, especially for high-resource languages. In contrast to multilingual models, monolingual models can fully concentrate on a certain language. As a result, they are more familiar with its peculiarities, capture more subtle nuances and thus are more sensitive to idiosyncratic phenomena.

6.4.3 Blind Spots of Fine-Tuned Neural Metrics

The present thesis demonstrated that fine-tuned neural metrics have various concerning blind spots, including named entities, numbers and compounds. The findings illustrate that for a high-resource language like German, monolingual models that are pre-trained with Whole Word Masking can alleviate the problem, but they cannot entirely solve it. The blind spots still persist. This behaviour of the metrics underpins the insight of Amrhein and Sennrich (2022): The unexpected blind spots are inherent to the metric model and cannot be easily removed.

However, it is still unclear where these biases stem from. For future work, it might be worth the effort to investigate the training data for possible biases, as language models are known to catch up such biases in their embeddings which affects the outcome of downstream tasks (Chan et al., 2020). The underlying language models of the studied metrics, XLM-RoBERTa and GBERT, were both trained on portions of the Common Crawl corpus. Hence, it is well possible that the blind spots are a consequence of noisy training data. However, whether this is the case or not has to be elucidated in future research.

The biases may also stem from the training data used to fine-tune the regression layer, i.e. the Direct Assessment data from the WMT Metrics Shared Tasks 2017 - 2019. However, the experiments conducted in this thesis revealed that the underlying language model has a much greater impact on the behaviour of a fine-tuned metric than the regression layer. Therefore, it is more likely that the problematic biases stem from the embeddings.

6.4.4 Evaluation of Fine-Tuned Neural Metrics

The experiments conducted in this thesis underpinned that GCOMET and GBLEURT show a superior performance to COMET-20 and $\text{COMET}_{\text{Constrastive}}$, both in terms of sensitivity and regarding the quality of their MBR-decoded translations.

In sharp contrast to these results, $\text{COMET}_{\text{Contrastive}}$ outperformed GCOMET and GBLEURT in the official WMT 2020 Metrics evaluation task of ranking MT systems according to their translation quality (cf. Table 10). When outlier MT systems were removed from the analysis, $\text{COMET}_{\text{Contrastive}}$ was only outperformed by $\text{GBLEURT}_{\text{WWM}}$ and the much larger COMET-20 model that reached the highest Pearson correlation with human judgements in both settings.

These results are in line with the observations by Sellam et al. (2020b) that multilingual models were more accurate than monolingual ones in the official evaluation of the WMT 2020 Metrics Shared Task. They attribute the superior performance of multilingual models to the larger amount of fine-tuning data that they saw during training. However, the experiments conducted for this thesis demonstrate that the MBR-decoded translations with $\text{COMET}_{\text{Contrastive}}$ are of lower quality than those obtained with GCOMET or GBLEURT. Further, both $\text{COMET}_{\text{Contrastive}}$ and COMET-20 are particularly insensitive to various linguistic phenomena. These findings highlight that system-level ranking of MT systems is not a good indicator of the quality and performance of a metric, as it only measures a very general ability of a metric ignoring its capability to handle particular linguistic phenomena.

Hence, the results of the WMT 2020 Metrics evaluation task are misleading and can result in catastrophic outcomes of experiments with winning metrics. For example, Freitag et al. (2022a) investigated an MBR-like re-ranking of a candidate list with COMET-QE. According to the WMT Metrics Shared Tasks, COMET-QE was among the winning systems achieving a high Pearson correlation in the system-level ranking task. However, the candidate re-ranking with COMET-QE resulted in low quality translations. In a similar vein, Amrhein and Sennrich (2022) as well as this thesis uncovered that winning metrics of the WMT Metrics evaluation suffer from certain blind spots and do not necessarily produce high quality translations when used in MBR decoding.

A good and accurate metric evaluation should uncover blind spots and measure more subtle capabilities than simple system-level ranking. It would be desirable that the WMT Metrics Shared Task would conduct a more fine-grained evaluation to reveal weaknesses of neural metrics as soon as possible. Such an evaluation should address the linguistic competence of metrics more directly. One way to achieve this would be to evaluate the metrics on a challenge set containing particular linguistic phenomena and targeted perturbations of candidates as proposed by Amrhein et al. (2022).

6.4.5 Metric Overfitting

The experiments shed light on the issue of metric overfitting. While both neural and non-neural metrics prefer translations that were decoded with the utility function that is most similar to themselves, the effect is considerably more pronounced for fine-tuned neural metrics than for lexical metrics.

The analysis revealed that it is above all the underlying language model that determines the preferences of a given neural metric. Using GCOMET$_{\text{WWM}}$ in the evaluation, illustrated that it is biased towards MBR-decoded translations generated with GBLEURT$_{\text{WWM}}$ that is built on the same underlying language model. Further, the architecture of the regression layer also plays a role. The bias of GCOMET$_{\text{WWM}}$ towards translations produced with GCOMET$_{\text{SWM}}$ as utility function is almost as strong as the one towards GBLEURT$_{\text{WWM}}$. However, it should be noted that the underlying language model of GCOMET$_{\text{SWM}}$ is very similar to that of GCOMET$_{\text{WWM}}$ differing only in the masking strategy used in pre-training. Nonetheless, the evaluation scores obtained with COMET-20 show that the architecture of the regression layer has an impact on the preferences of a neural metric. COMET-20 favours MBR-decoded translations generated with GCOMET over those decoded with GBLEURT.

These observations underscore that the evaluation of MBR-decoded translations in terms of automatic metrics is intricate and often leads to unreliable results. In the future, it will be crucial to choose evaluation metrics that are as distinct as possible from the utility function used in MBR decoding. The evaluation metric should rely on a language model that has a dissimilar architecture and is trained on different data sets than the language model of the utility function. Further, the regression layers of the evaluation metric and the utility function should differ as much as possible. Only then reliable evaluation results can be obtained.

6.4.6 Combining Metrics in MBR Decoding

This thesis experimented with a new approach of combining two utility functions during MBR decoding. A major downside of neural metrics are their occasional unpredictable failures in quality assessment. As the failures are unpredictable, different metrics are likely to fail on different segments. Hence, by combining two metrics in MBR decoding, each segment is checked twice before a decision is made as to which candidate is best.

The experiments demonstrated that this double check indeed leads to better and more accurate decisions. The translations obtained with a metric combination as utility function are of higher quality both in terms of a manual inspection as well as of BLEURT-20 and ChrF++ scores. Moreover, the results showed that the combination of two metrics alleviates the metric overfitting problem, even though it does not entirely solve it.

Nonetheless, the combination of two (or more) metrics in MBR decoding is a very promising approach for future work. The results indicate that the different strengths and weaknesses of distinct metrics cancel each other out, leading to more stable outcomes. Reliable quality scores are not only essential in MBR decoding, but also in the evaluation of MT outputs. Hence, combining metrics with different characteristics, architectures, strengths and weaknesses might also be beneficial to that field of research and application.

6.4.7 GCOMET versus GBLEURT

When comparing GCOMET and GBLEURT, the experiments revealed some differences between the two model types. For example, GBLEURT is clearly more sensitive to mistakes in NEs than to errors in compounds, whereas GCOMET exhibits a similar sensitivity to the two phenomena. GBLEURT reacts considerably more sensitive to replacements of entire nouns and NEs than to modifications of single characters. For GCOMET, the sensitivity gap between minor and major changes is less pronounced.

When considering all experiments conclusively, GBLEURT$_{\text{WWM}}$ seems to be the best of the four German metrics. MBR-decoded translations with GBLEURT$_{\text{WWM}}$ as utility function receive higher evaluation scores and contain fewer unknown words than those with GCOMET$_{\text{WWM}}$. However, the gains of GLBEURT$_{\text{WWM}}$ over GCOMET$_{\text{WWM}}$ are only minor. It cannot be definitely awarded as the best metric.

The important takeaway here is that the architecture of the regression layer affects the general performance of a metric and influences its behavior towards certain linguistic phenomena. However, the influence of the regression layer architecture is relatively small compared to the impact of the language model. It is above all the underlying language model that controls the behaviour of a neural metric.

6.4.8 MBR$_{100}$ versus MBR$_{99}$

This thesis also examined, if and to what extent including or excluding the comparison of a given candidate to itself as support segment during MBR decoding influences the outcome, i.e. whether other candidates are chosen as best translations in the two settings.

The analysis revealed that MBR$_{100}$ and MBR$_{99}$ can indeed lead to different translations. However, less than 1 % of the translated segments are affected. Hence, the question of whether to use MBR$_{100}$ or MBR$_{99}$ can be neglected in practice. Nonetheless, it is an important insight that fine-tuned neural metrics do not always assign the same score to two identical segments. Rather, the score depends on the segment itself.

7 Conclusion

This thesis was dedicated to the weaknesses of trained neural MT metrics. In particular, it elucidated the sensitivity of neural metrics towards German compounds. To identify blind spots and to gain insights about the sensitivity of different metrics, sampling-based MBR decoding with various utility functions was used, following the approach by Amrhein and Sennrich (2022). In sampling-based MBR decoding, the candidate translations are typically of a lower quality than beam search translations, on which most MT metrics are trained. The unusual errors in the MBR samples challenge the neural metrics and cause them to reveal their weaknesses.

The first part of the thesis concentrated on a case study on COMET-20. The MBR translations obtained with COMET-20 as utility function were analysed for mistranslated compounds and compared against beam search translations and MBR-decoded outputs generated with ChrF and ChrF++. The analysis revealed that the large amount of nonsensical compounds in $\text{MBR}_{\text{COMET-20}}$ outputs are not just an artifact of the MBR candidate pool, since the translations obtained with MBR_{ChrF} and $\text{MBR}_{\text{ChrF++}}$ contained considerably fewer incorrect compounds. Rather, the results highlight that German compounds are a particular weakness of COMET-20.

Having identified this blind spot of COMET-20, the second part of the thesis was devoted to the question whether a language model, that was pre-trained with Whole Word Masking instead of Sub-Word Masking, could enhance the sensitivity of a neural metric towards German compounds.

For that purpose, two new types of metrics were trained, GCOMET and GBLEURT. Both are built on the monolingual German GBERT model, but they differ in their architectures of the regression layer. Both metric types are trained in two flavours: one builds on GBERT pre-trained with Sub-Word Masking, the other one on GBERT pre-trained with Whole Word Masking. In addition, to assess the effect of a multilingual language model compared to a monolingual one, $\text{COMET}_{\text{Contrastive}}$ was trained, that is based on the multilingual $\text{XLM-RoBERTa}_{\text{Base}}$.

To shed light on the research questions, a series of experiments was run, in which the five new metrics were deployed as utility functions in MBR decoding. In the first experiment, the MBR translations were evaluated in terms of various automatic MT metrics. The second experiment analysed the amount of nonsensical compounds in the MBR outputs of the different utility functions. Finally, an MBR-based sensitivity analysis, as proposed by Amrhein and Sennrich (2022), was conducted, measuring the reaction of the metrics towards targeted changes in nouns, compounds, named entities and numbers.

The results of all these experiments confirmed the hypothesis that Whole Word Masking in the pre-training of the underlying language model renders the metric more sensitive to compounds. However, the increased sensitivity towards mistakes in compounds does not lead to a reduction of unknown words in the MBR translations of WWM-based utility functions. On the contrary, SWM-based utility functions perform better. Moreover, WWM does not consistently enhance the sensitivity of a metric to other linguistic phenomena like nouns, named entities and numbers. Hence, the blind spots still persist. WWM alleviates the weaknesses, but it does not fully remove them.

Despite these inconsistent effects of WWM, various automatic MT evaluation metrics agree that WWM-based utility functions generate better translations than SWM-based ones. Hence, the results indicate that WWM-based metrics do not make fewer mistakes, but *less severe* mistakes. They are more sensitive to errors that seriously hinder the comprehensibility of a translation, while punishing smaller mistakes, like perturbations of single characters, less. Thus, they assess errors with more human-like priorities. Given the overall positive impact of WWM on the performance of a neural metric, I recommend to base future metrics on language models that were pre-trained with WWM.

Further, the experiments illustrated that for a high-resource language like German, monolingual metrics overall tend to generate translations of higher quality and clearly outperform multilingual metrics at handling specific linguistic phenomena. Massively multilingual models seem to be detrimental to the performance of a metric and exacerbate its weaknesses, at least in the high-resource scenario. As the capacity is divided between a large amount of languages, a multilingual model is incapable of capturing idiosyncratic nuances. Hence, for future work on high-resource languages, it is advisable to rely on monolingual metrics for both MT evaluation as well as MBR decoding.

Further, the experiments demonstrated that the system-level ranking of MT systems is not an appropriate measure to assess the quality of an MT metric. The multilingual metrics that achieved the highest Pearson correlation with human judgements in this task, showed a poor performance when used as utility function in MBR decoding and are particularly insensitive to language-specific phenomena. This finding is concerning as the system-level ranking of MT systems is currently the official evaluation task of the WMT Metrics Shared Task. To avoid that the NLP community optimises towards metrics with severe blind spots, better evaluation methods are needed as soon as possible.

Moreover, this thesis investigated the effect of different MBR implementations. Specifically, it elucidated whether including or excluding the comparison of a candidate segment to itself as support segment influences the final outcome. The experiment demonstrated that neural metrics do not always assign the same score when comparing two identical segments. Rather, the scores vary from segment to segment. Hence, including or excluding the comparison of identical segments affects the outcome of MBR decoding in an unforeseeable way.

Finally, the thesis addressed the issue of metric overfitting that was observed in the automatic evaluation of MBR-decoded translations. The effect was especially strong for neural metrics and even occurred across metrics of the same family. Thereby, it is mainly the underlying language model that determines the preferences of neural metrics. Neural metrics clearly favour other metrics that build on the same language model. Additionally, they show a slight preference for metrics with a similar regression layer.

To alleviate this problem, I experimented with combining two different utility functions during MBR decoding. This combination turned out to be beneficial to the overall performance, improving translation quality and alleviating the overfitting effect. The two metrics seem to compensate for each other's weaknesses leading to more reliable results with fewer unpredictable failures.

7.1 Future Work

The various insights gained in this thesis also raised new questions. First of all, the experiments illustrated that the weaknesses and blind spots are inherent to a neural metric. Replacing the underlying language model of a metric with a monolingual model pre-trained with WWM enhances the sensitivity of the metric to certain linguistic phenomena and alleviates its weaknesses. However, it does not fully erase the metric's blind spots. The same holds true for retraining the metric on synthetic data (cf. Amrhein and Sennrich, 2022). Hence, it remains unclear where these blind spots stem from and how they can be removed. Future work should be dedicated to identify the reasons for the observed weaknesses. One possibility is that the blind spots are owed to biases in the training data. Both XLM-RoBERTa as well as GBERT are trained on portions of the Common Crawl corpus. Hence, it might be worth the effort to scrutinise this data set for possible biases.

While working on this thesis, I encountered evidence for additional weaknesses of COMET-20 that were already mentioned in Amrhein and Sennrich (2022). In particular, COMET-20 seems to be insensitive towards gender and polarity errors. Both error types can drastically change the meaning of a sentence. Hence, future research should elucidate these problems.

Further, the combination of two metrics yielded promising results. Hence, this approach should be studied more extensively, as it could be beneficial to various fields of application. On the one hand, it has the potential to improve the quality of MBR-decoded outputs. On the other hand, it might also be used in the evaluation of MT translations. The combination of different metrics might provide more reliable scores and alleviate the problem of metric overfitting.

Moreover, the experiments demonstrated that the evaluation of MBR translations is an intricate issue, since distorting overfitting effects can even occur across metrics that share certain similarities. As MBR decoding becomes more prominent and widespread, it is of essential importance to find methods to evaluate the translation quality reliably. One approach might be the above mentioned combination of different metrics. Another solution might be to use metrics that are as distinct as possible from the utility function in terms of the underlying language model, the masking strategy, the regression layer and the training data.

Finally, the results of this thesis indicate that system-level ranking of MT systems, the current evaluation method of the WMT Metrics Shared Task, does not produce reliable estimates of a metric's quality. As the NLP community heavily relies on the winning metrics of this shared task to guide its decisions, it is of crucial importance to find more reliable evaluation methods. A good evaluation procedure should take into account the ability of a metric to deal with specific linguistic phenomena. One possibility would be to conduct the evaluation on different challenge sets as proposed by Amrhein et al. (2022).

References

Farhad Akhbardeh, Arkady Arkhangorodsky, Magdalena Biesialska, Ondřej Bojar, Rajen Chatterjee, Vishrav Chaudhary, Marta R. Costa-jussa, Cristina España-Bonet, Angela Fan, Christian Federmann, et al. Findings of the 2021 conference on machine translation (WMT21). In *Proceedings of the Sixth Conference on Machine Translation*, pages 1–88. Association for Computational Linguistics, 2021. URL `https://aclanthology.org/2021.wmt-1.1`.

Chantal Amrhein and Rico 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 Twelfth International Joint Conference on Natural Language Processing*, pages 1125–1141. Association for Computational Linguistics, 2022. URL `https://aclanthology.org/2022.aacl-main.83`.

Chantal Amrhein, Nikita Moghe, and Liane Guillou. ACES: Translation accuracy challenge sets for evaluating machine translation metrics. In *Proceedings of the Seventh Conference on Machine Translation*, pages 479–513. Association for Computational Linguistics, 2022. URL `https://aclanthology.org/2022.wmt-1.44`.

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

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

Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam T. Kalai. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. *Advances in Neural Information Processing Systems*, 29:4349–4357, 2016. URL `https://papers.nips.cc/paper_files/paper/2016/hash/a486cd07e4ac3d270571622f4f316ec5-Abstract.html`.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al.

Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901, 2020. URL https://papers.nips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html.

Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334): 183–186, 2017. URL https://www.science.org/doi/10.1126/science.aal4230.

Branden Chan, Stefan Schweter, and Timo Möller. German's next language model. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6788–6796. International Committee on Computational Linguistics, 2020. URL https://aclanthology.org/2020.coling-main.598.

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

Hyung Won Chung, Thibault Fevry, Henry Tsai, Melvin Johnson, and Sebastian Ruder. Rethinking embedding coupling in pre-trained language models. In *Ninth International Conference on Learning Representations*. OpenReview.net, 2021. URL https://openreview.net/forum?id=xpFFI_NtgpW.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. ELECTRA: Pre-training text encoders as discriminators rather than generators. In *Eighth International Conference on Learning Representations*. OpenReview.net, 2020. URL https://openreview.net/forum?id=r1xMH1BtvB.

Alexis Conneau and Guillaume Lample. Cross-lingual language model pretraining. *Advances in Neural Information Processing Systems*, 32:7059–7069, 2019. URL https://papers.nips.cc/paper_files/paper/2019/hash/c04c19c2c2474dbf5f7ac4372c5b9af1-Abstract.html.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451. Association for Computational Linguistics, 2020. URL https://aclanthology.org/2020.acl-main.747.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, and Ziqing Yang. Pre-training with whole word masking for Chinese BERT. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3504–3514, 2021. URL https://ieeexplore.ieee.org/document/9599397.

Yong Dai, Linyang Li, Cong Zhou, Zhangyin Feng, Enbo Zhao, Xipeng Qiu, Piji Li, and Duyu Tang. “Is whole word masking always better for Chinese BERT?”: Probing on Chinese grammatical error correction. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1–8. Association for Computational Linguistics, 2022. URL https://aclanthology.org/2022.findings-acl.1.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina 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*, pages 4171–4186. Association for Computational Linguistics, 2019. URL https://aclanthology.org/N19-1423.

Bryan Eikema and Wilker Aziz. Is MAP decoding all you need? The inadequacy of the mode in neural machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4506–4520. International Committee on Computational Linguistics, 2020. URL https://aclanthology.org/2020.coling-main.398.

Bryan Eikema and Wilker Aziz. Sampling-based approximations to minimum Bayes risk decoding for neural machine translation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10978–10993. Association for Computational Linguistics, 2022. URL https://aclanthology.org/2022.emnlp-main.754.

Patrick Fernandes, António Farinhas, Ricardo Rei, José G. C. de Souza, Perez Ogayo, Graham Neubig, and Andre Martins. Quality-aware decoding for neural machine translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1396–1412. Association for Computational Linguistics, 2022. URL https://aclanthology.org/2022.naacl-main.100.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474, 2021a. URL https://aclanthology.org/2021.tacl-1.87.

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

Markus Freitag, David Grangier, Qijun Tan, and Bowen Liang. High quality rather than high model probability: Minimum Bayes risk decoding with neural metrics.

Transactions of the Association for Computational Linguistics, 10:811–825, 2022a. URL https://aclanthology.org/2022.tacl-1.47.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André 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*, pages 46–68. Association for Computational Linguistics, 2022b. URL https://aclanthology.org/2022.wmt-1.2.

Vaibhava Goel and William J Byrne. Minimum Bayes-risk automatic speech recognition. *Computer Speech & Language*, 14(2):115–135, 2000. URL https://www.sciencedirect.com/science/article/abs/pii/S0885230800901384.

Alex Graves. Sequence transduction with recurrent neural networks. *arXiv:1211.3711*, 2012. URL http://arxiv.org/abs/1211.3711.

Fabio Kepler, Jonay Trénous, Marcos Treviso, Miguel Vera, and André F. T. Martins. OpenKiwi: An open source framework for quality estimation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 117–122. Association for Computational Linguistics, 2019. URL https://aclanthology.org/P19-3020.

Huda Khayrallah and Philipp Koehn. On the impact of various types of noise on neural machine translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 74–83. Association for Computational Linguistics, 2018. URL https://aclanthology.org/W18-2709.

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

Tom Kocmi, Christian Federmann, Roman Grundkiewicz, Marcin Junczys-Dowmunt, Hitokazu Matsushita, and Arul Menezes. To ship or not to ship: An extensive evaluation of automatic metrics for machine translation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 478–494. Association for Computational Linguistics, 2021. URL https://aclanthology.org/2021.wmt-1.57.

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

Taku Kudo and John Richardson. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System*

Demonstrations, pages 66–71. Association for Computational Linguistics, 2018. URL https://aclanthology.org/D18-2012.

Aviral Kumar and Sunita Sarawagi. Calibration of encoder decoder models for neural machine translation. *arXiv:1903.00802*, 2019. URL https://arxiv.org/abs/1903.00802.

Shankar Kumar and William Byrne. Minimum Bayes-risk decoding for statistical machine translation. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 169–176. Association for Computational Linguistics, 2004. URL https://aclanthology.org/N04-1022.

Alon Lavie and Michael J Denkowski. The METEOR metric for automatic evaluation of machine translation. *Machine translation*, 23:105–115, 2009. URL https://doi.org/10.1007/s10590-009-9059-4.

Katherine Lee, Orhan Firat, Ashish Agarwal, Clara Fannjiang, and David Sussillo. Hallucinations in neural machine translation, 2019. URL https://openreview.net/forum?id=SkxJ-309FQ.

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81. Association for Computational Linguistics, 2004. URL https://aclanthology.org/W04-1013.

Chao Liu, Cui Zhu, and Wenjun Zhu. Chinese named entity recognition based on BERT with whole word masking. In *Proceedings of the 2020 Sixth International Conference on Computing and Artificial Intelligence*, pages 311–316. Association for Computing Machinery, 2020a. URL https://doi.org/10.1145/3404555.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized BERT pretraining approach. In *Eighth International Conference on Learning Representations*. OpenReview.net, 2020b. URL https://openreview.net/forum?id=SyxS0T4tvS.

Chi-kiu Lo. YiSi - a unified semantic MT quality evaluation and estimation metric for languages with different levels of available resources. In *Proceedings of the Fourth Conference on Machine Translation*, pages 507–513. Association for Computational Linguistics, 2019. URL https://aclanthology.org/W19-5358.

Qingsong Ma, Ondřej Bojar, and Yvette Graham. Results of the WMT18 metrics shared task: Both characters and embeddings achieve good performance. In *Proceedings of the Third Conference on Machine Translation*, pages 671–688. Association for Computational Linguistics, 2018. URL https://aclanthology.org/W18-6450.

Qingsong Ma, Johnny Wei, Ondřej Bojar, and Yvette 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*, pages 62–90. Association for Computational Linguistics, 2019. URL https://aclanthology.org/W19-5302.

Benjamin Marie, Atsushi Fujita, and Raphael Rubino. Scientific credibility of machine translation research: A meta-evaluation of 769 papers. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the Eleventh International Joint Conference on Natural Language Processing*, pages 7297–7306. Association for Computational Linguistics, 2021. URL https://aclanthology.org/2021.acl-long.566.

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

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

Mathias Müller and Rico Sennrich. Understanding the properties of minimum Bayes risk decoding in neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the Eleventh International Joint Conference on Natural Language Processing*, pages 259–272. Association for Computational Linguistics, 2021. URL https://aclanthology.org/2021.acl-long.22.

Mathias Müller, Annette Rios, and Rico Sennrich. Domain robustness in neural machine translation. In *Proceedings of the 14th Conference of the Association for Machine Translation in the Americas*, pages 151–164. Association for Machine Translation in the Americas, 2020. URL https://aclanthology.org/2020.amta-research.14.

Malte Ostendorff, Till Blume, and Saskia Ostendorff. Towards an open platform for legal information. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020*, pages 385–388. Association for Computing Machinery, 2020. URL https://doi.org/10.1145/3383583.3398616.

Myle Ott, Michael Auli, David Grangier, and Marc'Aurelio Ranzato. Analyzing uncertainty in neural machine translation. In *Proceedings of the 35th International Conference on Machine Learning*, pages 3956–3965. Proceedings of Machine Learning Research, 2018. URL https://proceedings.mlr.press/v80/ott18a.html.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318. Association for Computational Linguistics, 2002. URL `https://aclanthology.org/P02-1040`.

Maja Popović. chrF: Character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395. Association for Computational Linguistics, 2015. URL `https://aclanthology.org/W15-3049`.

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

Amy Pu, Hyung Won Chung, Ankur Parikh, Sebastian Gehrmann, and Thibault Sellam. Learning compact metrics for MT. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 751–762. Association for Computational Linguistics, 2021. URL `https://aclanthology.org/2021.emnlp-main.58`.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. *OpenAI Blog*, 2018. URL `https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf`.

Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. Sequence level training with recurrent neural networks. *arXiv:1511.06732*, 2015. URL `https://arxiv.org/abs/1511.06732`.

Ricardo Rei, Craig Stewart, Ana C. Farinha, and Alon Lavie. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 2685–2702. Association for Computational Linguistics, 2020a. URL `https://aclanthology.org/2020.emnlp-main.213`.

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

Ricardo Rei, Ana C. Farinha, Chrysoula Zerva, Daan van Stigt, Craig Stewart, Pedro Ramos, Taisiya Glushkova, André F. T. Martins, and Alon Lavie. Are references really needed? Unbabel-IST 2021 submission for the metrics shared task. In *Proceedings of the Sixth Conference on Machine Translation*, pages 1030–1040. Association for Computational Linguistics, 2021. URL `https://aclanthology.org/2021.wmt-1.111`.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C. Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation*, pages 578–585. Association for Computational Linguistics, 2022. URL `https://aclanthology.org/2022.wmt-1.52`.

Helmut Schmid, Arne Fitschen, and Ulrich Heid. SMOR: A German computational morphology covering derivation, composition and inflection. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation*. European Language Resources Association, 2004. URL `http://www.lrec-conf.org/proceedings/lrec2004/pdf/468.pdf`.

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

Thibault Sellam, Amy Pu, Hyung Won Chung, Sebastian Gehrmann, Qijun Tan, Markus Freitag, Dipanjan Das, and Ankur Parikh. Learning to evaluate translation beyond English: BLEURT submissions to the WMT metrics 2020 shared task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 921–927. Association for Computational Linguistics, 2020b. URL `https://aclanthology.org/2020.wmt-1.102`.

Rico Sennrich and Beat Kunz. Zmorge: A German morphological lexicon extracted from Wiktionary. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, pages 1063–1067. European Language Resources Association, 2014. URL `http://www.lrec-conf.org/proceedings/lrec2014/pdf/116_Paper.pdf`.

Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 1715–1725. Association for Computational Linguistics, 2016. URL `https://aclanthology.org/P16-1162`.

Rico Sennrich, Orhan Firat, Kyunghyun Cho, Alexandra Birch, Barry Haddow, Julian Hitschler, Marcin Junczys-Dowmunt, Samuel Läubli, Antonio Valerio Miceli Barone, Jozef Mokry, and Maria Nădejde. Nematus: A toolkit for neural machine translation. In *Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 65–68. Association for Computational Linguistics, 2017. URL `https://aclanthology.org/E17-3017`.

Hiroki Shimanaka, Tomoyuki Kajiwara, and Mamoru Komachi. RUSE: Regressor using sentence embeddings for automatic machine translation evaluation. In *Proceedings of the Third Conference on Machine Translation*, pages 751–758. Association for Computational Linguistics, 2018. URL `https://aclanthology.org/W18-6456`.

Raphael Shu and Hideki Nakayama. Later-stage minimum Bayes-risk decoding for neural machine translation. *arXiv:1704.03169*, 2017. URL `https://arxiv.org/abs/1704.03169`.

Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. A study of translation edit rate with targeted human annotation. In *Proceedings of the Seventh Conference of the Association for Machine Translation in the Americas*. Association for Machine Translation in the Americas, 2006. URL `https://aclanthology.org/2006.amta-papers.25`.

Pavel Sountsov and Sunita Sarawagi. Length bias in encoder decoder models and a case for global conditioning. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1516–1525. Association for Computational Linguistics, 2016. URL `https://aclanthology.org/D16-1158`.

Lucia Specia, Frédéric Blain, Marina Fomicheva, Chrysoula Zerva, Zhenhao Li, Vishrav Chaudhary, and André F. T. Martins. Findings of the WMT 2021 shared task on quality estimation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 684–725. Association for Computational Linguistics, 2021. URL `https://aclanthology.org/2021.wmt-1.71`.

Felix Stahlberg and Bill Byrne. On NMT search errors and model errors: Cat got your tongue? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the Ninth International Joint Conference on Natural Language Processing*, pages 3356–3362. Association for Computational Linguistics, 2019. URL `https://aclanthology.org/D19-1331`.

Felix Stahlberg, Adrià de Gispert, Eva Hasler, and Bill Byrne. Neural machine translation by minimising the Bayes-risk with respect to syntactic translation lattices. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 362–368. Association for Computational Linguistics, 2017. URL `https://aclanthology.org/E17-2058`.

Miloš Stanojević and Khalil Sima'an. BEER 1.1: ILLC UvA submission to metrics and tuning task. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 396–401. Association for Computational Linguistics, 2015. URL `https://aclanthology.org/W15-3050`.

Pedro Javier Ortiz Suárez, Benoît Sagot, and Laurent Romary. Asynchronous pipeline for processing huge corpora on medium to low resource infrastructures. In *Proceedings of the Seventh Workshop on the Challenges in the Management of Large Corpora*, pages 9–16. Leibniz-Institut für Deutsche Sprache, 2019. URL `https://ids-pub.bsz-bw.de/frontdoor/index/index/docId/9021`.

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. Mitigating gender bias in natural language processing: Literature review. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1630–1640. Association for Computational Linguistics, 2019. URL https://aclanthology.org/P19-1159.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. *Advances in Neural Information Processing Systems*, 27:3104–3112, 2014. URL https://papers.nips.cc/paper_files/paper/2014/hash/a14ac55a4f27472c5d894ec1c3c743d2-Abstract.html.

Brian Thompson and Matt Post. Automatic machine translation evaluation in many languages via zero-shot paraphrasing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 90–121. Association for Computational Linguistics, 2020. URL https://aclanthology.org/2020.emnlp-main.8.

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

Roy Tromble, Shankar Kumar, Franz Och, and Wolfgang Macherey. Lattice minimum Bayes-risk decoding for statistical machine translation. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 620–629. Association for Computational Linguistics, 2008. URL https://aclanthology.org/D08-1065.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30:6000–6010, 2017. URL https://papers.nips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Weiyue Wang, Jan-Thorsten Peter, Hendrik Rosendahl, and Hermann Ney. CharacTER: Translation edit rate on character level. In *Proceedings of the First Conference on Machine Translation*, pages 505–510. Association for Computational Linguistics, 2016. URL https://aclanthology.org/W16-2342.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv:1609.08144*, 2016. URL https://arxiv.org/abs/1609.08144.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations*. OpenReview.net, 2020. URL https://openreview.net/forum?id=SkeHuCVFDr.

Junru Zhou, Zhuosheng Zhang, Hai Zhao, and Shuailiang Zhang. LIMIT-BERT : Linguistics informed multi-task BERT. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4450–4461. Association for Computational Linguistics, 2020. URL `https://aclanthology.org/2020.findings-emnlp.399`.

Curriculum Vitae

Sarah Elisabeth Kiener
Gotthardstrasse 30
6410 Goldau
sarahelisabeth.kiener@uzh.ch

Education

Since 2018	Master of Arts in Computational Linguistics & Language Technology University of Zurich
2019	FIDE Certificate as Instructor for Language and Integration Courses
2017	SVEB-1 & EUROLTA Certificates in Language Teaching to Adults Migros Klubschule Lucerne
2013 – 2016	Master of Arts in Linguistics University of Berne
2013 – 2014	Erasmus Exchange Scholarship Complutense University of Madrid, Spain
2009 – 2013	Bachelor of Arts in Linguistics University of Berne

Work Experience

Since 2023	Software Engineering Intern Textshuttle, Zurich
2020 – 2023	Research Assistant in the Digital Linguistics Group Department of Computational Linguistics, University of Zurich
2020 – 2023	Student Assistant in the Research Project ”Diffusion of Complex Linguistic Structures in Social Media“ Institute of Romance Studies, University of Zurich

2019 – 2021	Tutor for ”Introduction to Computational Linguistics I & II“ Department of Computational Linguistics, University of Zurich
2016 – 2019	German Teaching to Adults and Support of Asylum Seekers Caritas Switzerland, Canton of Schwyz
2016	Internship in Support of Asylum Seekers Caritas Switzerland, Canton of Schwyz
2014	Internship in Computational Linguistics UNDL Foundation, Geneva
2012 – 2013	Internship in Terminology Federal Chancellery of Switzerland, Terminology Section, Berne

A Tables

A.1 Hyperparamters of GBLEURT

Table 16 shows the hyperparameters that were used to train GBLEURT. For hyperparameters that are not specified in the table, the default values were used (cf. Sellam et al., 2020a).

Hyperparameter	$\text{GBLEURT}_{\text{SWM}}$	$\text{GBLEURT}_{\text{WWM}}$
Encoder	$\text{GBERT}_{\text{SWM}}$	$\text{GBERT}_{\text{WWM}}$
Optimizer	AdamW	AdamW
Batch Size	8	8
Gradient Accum. Steps	4	4
Learning Rate	1e-5	1e-5
Training Steps	20 000	20 000
Evaluation Steps	500	500
Logging Steps	500	500
Max. Length	512 Tokens	512 Tokens
Padding	Max. Length	Max. Length
Truncation	False	False
Early Stopping	True	True
Best Checkpoint	After 1 000 training steps	After 3 500 training steps

Table 16: Hyperparameters of the GBLEURT models

A.2 Hyperparameters of GCOMET and $\text{COMET}_{\text{Contrastive}}$

Table 17 summarizes the most important hyperparameters that were used to train GCOMET and $\text{COMET}_{\text{Contrastive}}$. For hyperparameters that are not specified in the table, the default values were used (cf. Rei et al., 2021).

Hyperparameter	GCOMET$_{\text{SWM}}$	GCOMET$_{\text{WWM}}$	COMET$_{\text{Contrastive}}$
Encoder	GBERT$_{\text{SWM}}$	GBERT$_{\text{WWM}}$	XLM-R$_{\text{Base}}$
Optimizer	AdamW	AdamW	AdamW
Batch Size	4	4	4
Gradient Accum. Steps	4	4	4
Encoder Learning Rate	1e-5	1e-5	1e-5
Learning Rate	3.1e-5	3.1e-5	3.1e-5
Layerwise Decay	0.95	0.95	0.95
Nr. Frozen Epochs	0.3	0.3	0.3
Keep Embeddings Frozen	True	True	True
Pooling	Average	Average	Average
Layer	Mix	Mix	Mix
Dropout	0.15	0.15	0.15
Hidden Sizes	[2048, 1024]	[2048, 1024]	[2048, 1024]
Epochs	1	1	1
Max. Length	512 Tokens	512 Tokens	512 Tokens
Padding	Longest	Longest	Longest
Truncation	True	True	True
Early Stopping	True	True	True
Best Checkpoint	After 11 130 training steps	After 11 130 training steps	After 18 550 training steps

Table 17: Hyperparameters of the GCOMET models

B Errors in Compounds

The manual exploration of MBR$_{\text{COMET-20}}$-decoded outputs (cf. 3.2) revealed that COMET-20 is not sensitive enough to different kinds of errors in German compounds. Even though the correct translation is often among the hypotheses in the candidate pool, COMET-20 often fails to identify it and instead chooses a candidate with an incorrect translation of the compound.

Several error types appear frequently in the MBR-decoded translations chosen by COMET-20. These types of errors are described in more detail in the following sections and illustrated with a few examples from the data. The list serves to shed light on the most frequently observed error types and to exemplify the various kinds of difficulties that COMET-20 exhibits when dealing with German compounds. However, the list is not exhaustive and more kinds of mistakes can be found in the MBR$_{\text{COMET-20}}$-decoded outputs.

B.1 Mistranslation of the Second Compound Component

Nonsensical compounds in which the first constituent is translated accurately, while subsequent components are not, are frequently encountered in the MBR$_{\text{COMET-20}}$-decoded output. Thus, COMET-20 seems to be insensitive to the semantic relationship between the two (or more) parts of a compound. If the first part is translated adequately, COMET-20 seems to consider the compound as sufficiently similar to the source and the support sentences. The examples in Table 18 illustrate this type of error. The ID in the table corresponds to the number of the sentence in the test set that contains the compound, whereby the first sentence is assigned the ID 0.

In some cases, the translation of the second component is entirely unrelated to the source sentence, e.g., in *Geburts-antrags-weise* that corresponds to *birth-proposal-manner* instead of *birth-day-present*, or *Garten-lichter*, meaning *garden-lights* instead of *garden-gimmick*. Sometimes, the second part is not even an existing German word, such as in *Schraub-schwendern* or *Magen-rewellen*.

However, the second part often exhibits a certain surface similarity to the reference translation, whereby it shares several characters with the correct translation, e.g., *Hunde-m**ark**t* and *Hunde-p**ark***, *Werkzeug-**k**öpf**e*** and *Werkzeug-**k**äst**e**n*, *Garten-**g**ebä**ck**en* and *Garten-**G**immi**ck**s*, *Dorf-**be**sitz**er*** and *Dorf-**be**wohn**er***, or *Straßen-**b**ü**n**d**n**iss**e*** and *Straßen-**b**ah**nen***. In rare cases, this surface similarity can be found between the translation and

ID	MBR COMET-20	Reference	Source
3	Hundemarkt	Hundepark	dog park
7	Video-Fotografie	Videoaufnahme	video footage
31	Elitembolleute	Elitesoldaten	elite soldiers
71	Schraubschwendern	Schraubenziehern	screwdrivers
75	Werkzeugköpfe	Werkzeugkästen	toolboxes
244	Gartengebäcken	Garten-Gimmicks	gardening gimmicks
245	Gartenlichter	Garten-Gimmicks	garden gimmick
471	Dorfbesitzer	Dorfbewohner	villagers
501	Herzverlust	Herzstillstand	cardiac arrest
504	Presseansprache	Pressemitteilung	press release
511	Magenrewellen	Bauchkrämpfe	stomach cramps
559	Geburtsantragsweise	Geburtstagsgeschenk	birthday present
816	Straßenbündnisse	Straßenbahnen	trams

Table 18: Examples of German compounds in which the first component is adequate, while subsequent components are wrong

the source, as in *Video-**Fotografie*** and *viedo-**footage***.

It is important to note that this kind of similarity exists purely on the superficial string-level, not on the semantic level. Semantic similarities between the inadequately translated second component of the compound and the reference or source word are rarely observed in this error type. An example might be *Presse-ansprache* and *Presse-mitteilung* where a certain semantic similarity between *Ansprache* (speech, address) and *Mitteilung* (message, notice, announcement) is given.

B.2 Mistranslation of the First Compound Component

The inverse error type, where the first constituent is incorrect, while the subsequent components are correct, is observed as well, even though with lower frequency. A few examples are shown in Table 19.

As mentioned in the previous section, the incorrectly translated compound part might exhibit a similar surface form as the correct translation, such as ***Mitt**ag-nacht* and ***Mitt**er-nacht* or ***L**ösch-wagen* and ***L**iefer-wägen* that share at least the same onset. In rare cases, the mistranslation is triggered by a word in the source sentence that appears outside the compound. An example is sentence 256 where the adjective *sky blue* occurs a few words before *fruit trees*. The word ***sky*** has probably caused the inadequate translation ***Himmel**-bäume*.

ID	MBR COMET-20	Reference	Source
33	Wachstumskampf	Nahkampf	close-quarter fighting
74	Löschwagen	Lieferwägen	vans
118	Schlauchmitteln	Arzneimitteln	drugs
256	Himmelbäume	Obstbäumen	fruit trees
564	Mittagnacht	Mitternacht	midnight
663	Götterhühnern	Brathähnchen	roast chicken

Table 19: Examples of German compounds in which the first component is incorrect, while subsequent components are adequate

However, in most cases the incorrect first part of the compound is unrelated to any other words in the source sentence and does not show any surface similarity to the correct target compound as can be seen in various examples in Table 19

Related to the two error types discussed so far, a third, albeit rare type of error appears in the data. On seldom occasion, the parts of a compounds are swapped, whereby one part might be translated inadequately. One such example is given in Table 20.

ID	MBR COMET-20	Reference	Source
69	Schneeschrauben	Schraubenziehern	screwdrivers

Table 20: Examples of German compounds in which the two parts are swapped

B.3 Polysemous Words

Another challenge which the translation model and the metric used as utility function face are polysemous words that form part of a compound. These words possess multiple senses that are often associated with distinct domains. Sometimes, the MT model and the metric fail to recognize the correct domain and, as a result, select an inappropriate meaning for the word in question. Examples for polysemous compound components for which the wrong meaning was chosen are listed in Table 21.

Interestingly, the domain of sports seems to be especially challenging for the COMET-20 metric. The last two examples, *goal difference* and *warm-up session*, illustrate that COMET-20 seems prefers the more general meaning of polysemous words. The English word *goal* can be a synonym of *aim, target, objective.* In this general sense of the word, *Ziel* is a valid German translation. However, in the specialized domain of sports, in which *goal* refers to a physical structure or target as well as to a score or point, it must be translated as *Tor.* COMET-20 fails to select the domain-specific meaning, even though the correct translation is present in the candidate set. The same applies for *warm-up session*, where

ID	MBR COMET	Reference	Source
135	Pflegeleiterin	Interimstrainer	caretaker manager
242	Broadway-Sterne	Broadway-Stars	Broadway stars
408	Zieldifferenz	Tordifferenz	goal difference
802	Aufwärmungssitzungen	Aufwärmübung	warm-up session

Table 21: Examples of German compounds in which an inadequate meaning for a polysemous component was chosen

session, in a general sense of the word, similar to *meeting, conference*, can be translated as *Sitzung*. However, in the sports domain, the correct translation would be *Übung* or *Training*.

An interesting examples is *caretaker manager*. The compound part *caretaker* is associated with the health domain, in which it certainly occurs most often. It is noteworthy that it is the MT model that is ignorant of the sports term *caretaker manager* and fails to translate it correctly. Nonetheless, the MT model produces several candidates that are more adequate and that agree in gender to the denoted person, such as *Betreuungsleiter*. However, COMET-20 does not select one of these slightly more accurate translations.

Finally, the example of *Broadway stars* illustrates that COMET-20 does not only struggle with the sports domain, but also with other domains. It prefers the general meaning of *stars*, that refers to the astronomical object and corresponds to the German word *Sterne*. However, in the context of *Broadway*, the word *stars* clearly refers to a celebrities and should be translated as *Stars*. Again, the correct translation is present in the candidate pool, but not chosen by COMET-20.

B.4 Gender Mistakes

In German, compounds that refer to a person, such as job titles or family relationships, agree in gender with that person. Occasionally, the $\text{MBR}_{\text{COMET-20}}$ output contains examples where this is not the case as shown in Table 22.

ID	MBR COMET	Reference	Source
51	Alkoholschwester	alkoholischen Bruder	alcoholic brother
135	Pflegeleiterin	Interimstrainer	caretaker manager
352	Pressekreisekretär	Pressesprecherin	press secretary
353	Untersuchungssekretär	Pressesprecherin	press secretary

Table 22: Examples of German compounds that do not agree in gender with the person they refer to

Correctly translating the gender of English job titles into German is a challenge for the MT model and the utility function, as most English job titles are gender-neutral, whereas in German they agree in gender with the person they denote. Hence, the MT model and the metric have to identify the referent in the context and adjust the gender of the job title accordingly. It is therefore understandable that the metric sometimes fails to select a hypothesis with the correct gender from the candidate pool.

Nonetheless, the example from sentence 353 is very interesting. COMET-20 selected a hypothesis in which the gender of the referent was identified correctly. The word *Frau* (Ms.) is added before the referent's name, even though this title is not present in the source as illustrated below:

src: The Associated Press has accused White House press secretary Kayleigh McEnany [...]

ref: Associated Press hat der Pressesprecherin des Weißen Hauses Kayleigh McEnany vorgeworfen, [...]

hyp: Die Associate Press hat dem Untersuchungssekretär für das Weiße Haus, **Frau** Kayleigh McEnany, vorgeworfen,[...]

Although the referent's gender is known, COMET-20 selects a hypothesis where the gender of job title does not agree with that of the referent. Hence, it is not sensitive enough to gender mistakes in compounds.

A striking example for COMET's insensitivity towards gender mistakes in compounds is the translation of *alcoholic brother* with *Alkoholschwester* (*alcoholic sister*). Even though the English word *brother* clearly denotes a masculine person and was translated correctly as *Bruder* in the vast majority of candidates, COMET-20 chooses its female counterpart.

While COMET-20 is not sensitive enough towards gender in compounds, gender stereotypes may also come into play. *Pflegeleiterin* might be an example for such a gender bias. As outlined above, the MT model as well as COMET-20 do attribute the term *caretaker manager* to the health domain instead of the sports domain. In our health systems, caretakers are typically women. Various studies have demonstrated that language models learn such gender biases from the training data (Bolukbasi et al., 2016; Caliskan et al., 2017; Sun et al., 2019). Hence, it is possible that it was the learned gender stereotype that triggered COMET-20 to choose a hypothesis where the gender in the compound does not correspond to the gender of the referent.

Apart from gender mistakes in compounds, the $\text{MBR}_{\text{COMET-20}}$ output contains various examples where the grammatical gender of German nouns is incorrect. In these cases, the gender of the article and adjectives does not agree with the gender of the noun. Hence, grammatical gender might be another weakness of COMET-20. However, this assumption is based on a preliminary manual exploration of the data. To draw an informed conclusion, a systematic investigation is needed. I leave this question to future work.

B.5 Erroneous Analysis of Concept Boundaries

In several examples in the MBR$_{\text{COMET-20}}$ decoded output, the boundaries of concepts were misanalyzed. In these cases, two neighbouring words are mistakenly analyzed as a single concept and translated to German with a compound. Errors of this kind mostly concern two linguistic structures: 1) adjective + noun, 2) present participle + noun. The to words are merged into a single concept and translated with a compound. The examples in Table 23 illustrate this phenomenon.

ID	MBR COMET	Reference	Source
320	Eishöhen	eisigen Höhen	icy heights
654	Frühstücksanfrage	[und sie] ihr Frühstück einfordern	demanding breakfast
807	Großabfall	starken Rückgänge	large declines
843	Strahlzentrum	strahlendes Zentrum	beaming center

Table 23: Examples of German compounds with an erroneous analysis of concept boundaries

It is important to note that COMET-20 is only partially to blame for this misinterpretation of concept boundaries. In three out of the four listed cases, the majority of candidates contained a wrong translation of the phrase in question. The phrase *beaming center* was correctly translated in only one candidate. Hence, it is not only COMET-20 that occasionally struggles with analyzing the word and concept boundaries correctly, but also the MT model. Nonetheless, COMET-20 seems to have greater difficulties with handling the mentioned linguistic constructions than the MT model. In the case of *large declines*, most candidates contain an adequate translation, but COMET-20 chooses a wrong one. Similarly, while many candidates use the incorrect compound *Eishöhen* for *icy heights*, there are many other candidates offering a correct translation, but COMET-20 is unable to select one of them.

B.6 English Loanwords and False Friends

German incorporated certain English words as loanwords. If these words form part of a compound, COMET-20 tends to prefer candidates which try to translate the English word instead of just copying it into German as a loanword. However, in many cases, the attempt to avoid loanwords results in wrong or at least not very idiosyncratic translations as shown in Table 24. The observed preference of COMET-20 for replacing loanwords with inadequate German terms might be caused to some degree by the MT model which produces many candidates that try to find a translation for the English loanword.

In a few instances, COMET-20 is susceptible to false friends. An example is shown in Table 25. The word *stab* exists in English as well as in German. However, the meaning of

ID	MBR COMET	Reference	Source
242	Broadway-Sterne	Broadway-Stars	Broadway stars
244	Gartengebäcken	Garten-Gimmicks	gardening gimmicks
245	Gartenlichter	Garten-Gimmicks	garden gimmick
676	Twitter-Posten	Twitter-Post	Twitter post
865	Pfeffersprühmittel	Pfefferspray	pepper spray
987	Sozialmedienunternehmen	Social-Media-Unternehmen	Social media companies
1000	Gesellschaftsmedienkonten	Konten in den sozialen Medien	social media accounts

Table 24: Examples of German compounds avoiding loanword components

the two words is completely different. The German word *Stab* translates to English as *bar, stick*. COMET-20 does not recognize this false friend and selects a candidate that copies the English compound component into the German compound producing a nonsensical translation.

ID	MBR COMET	Reference	Source
55	Stabanzeichen	Stichwunden	stab marks

Table 25: Example of the mistranslation of a false friend

B.6.1 Repeated Compound Parts and Perturbation of Single Characters

Finally, the MBR$_{\text{COMET-20}}$-decoded output contains compounds in which a component was repeated several times as shown in the upper part of Table 26. In other compounds, a single character was replaced, deleted or added as illustrated in the lower part of Table 26. These error types are not exclusively related to compounds and are frequently encountered in MT output. However, as the examples below demonstrate, they also occur within compounds.

ID	MBR COMET	Reference	Source
289	Festlandsland	Festland	mainland
666	Geburtstagestag	Geburtstag	birthday
0	Hundpark	Hundepark	dog park
21	Pfeifferspray	Pfefferspray	pepper spray

Table 26: Examples of repeated compound parts and perturbations of single characters

B.7 Other Error Types

In the previous sections, the most frequently observed error types were analysed and exemplified in detail. This list is not exhaustive and serves to provide the reader with an overview of common errors associated with compounds. Many more error types related to compounds are encountered in the MBR-decoded output. However, they occur less frequently than the aforementioned types and it is beyond the scope of this work to analyze each of them in detail. In the present section, only the most interesting ones are briefly discussed.

In a some instances, COMET-20 selects a candidate that produces a compound where none is needed. In these cases, neither the English source word nor the adequate German translations are compounds. These compounds are usually nonsensical and mostly do not share any component or semantic content with the correct translation. Examples are given in the upper part of Table 27. An exception is the compound *Spiegelstellen* which incorporates the target word as component.

ID	MBR COMET	Reference	Source
70	Fließzone	Gehege	enclosure
72	Spiegelstellen	Spiegel	mirrors
686	Siegerschlag	Schießerei	shooting
72	Windburn-Witten	Scheibenwischer	windscreen wipers

Table 27: Other examples of nonsensical German compounds

In some instances, certain constituents of the compound do not exist as lexemes in the German language. Two examples were already listed in Table 18: *Schraub-**schwendern*** and *Magen-**rewellen***. An example where the MT output has nothing in common with the reference translation is shown in Table 27. *Windburn-Witten* shares a substring with the source, but not with the reference. Nonetheless, it can be considered as a compound, given the hyphen between the components on the one hand and the fact that the source contains a compound with a similar surface form on the other.

Finally, compounds also appear in hallucinations, i.e. in translations that have nothing in common with the source. Compounds generated in hallucinations can be correctly formed or nonsensical, but in either case they are entirely unrelated to the meaning of the source sentence.