University of Zurich[UZH]

Master's thesis
for obtaining the academic degree
Master of Arts
at the Faculty of Arts and Social Sciences of the University of Zurich

Unsupervised Translation Direction Detection

Author: Michelle Wastl
Matriculation Number: 16-727-398

Supervisor: Prof. Dr. Rico Sennrich

Institute of Computational Linguistics

Submission date: 01.12.2023

Abstract

This thesis describes an unsupervised approach to determine the translation direction for parallel texts. Traditional methods in this field rely on large amounts of homogeneous parallel data, which limits their applicability in real-world scenarios. This research shows how this caveat can be overcome by leveraging translation probabilities that are generated by neural machine translation (NMT) models for parallel texts. The effectiveness of this approach at sentence level is tested on a dataset that encompasses different language pairs across various resource levels and domains, and which include human, pre-neural, and neural machine translations. At document level the effectiveness is even more prominent. Furthermore, the approach is applied to texts from a real plagiarism case. The approach demonstrates comparable performance to existing methods while being less resource-intensive and showing more robustness.

Zusammenfassung

Die vorliegende Arbeit untersucht einen nicht-supervisierten Ansatz zur Bestimmung der Übersetzungsrichtung bei parallelen Texten. Traditionelle Methoden in diesem Bereich verwenden grosse Mengen an parallelen Texten, was ihre Anwendung in realen Situationen einschränkt. Diese Arbeit zeigt, wie dieses Problem mit Hilfe von Übersetzungswahrscheinlichkeiten neuronaler maschineller Übersetzungsmodelle (NMÜ) für parallele Texte überwunden werden kann. Die Wirksamkeit dieses Ansatzes auf Satzebene wird an einem Datensatz geprüft, der verschiedene Sprachpaare mit unterschiedlicher Menge an Ressourcen, sowie menschliche, vor-neuronale und neuronale maschinelle Übersetzungen über mehrere Domänen umfasst. Auf Dokumentebene ist die Wirksamkeit des Ansatzes noch ausgeprägter. Zusätzlich wird der Ansatz an Texten eines echten Plagiatfalls angewendet. Der Ansatz zeigt vergleichbare Wirksamkeit zu bestehenden Methoden, ist dabei jedoch weniger ressourcenintensiv und weist eine grössere Robustheit auf.

Acknowledgements

I would like to express my sincere gratitude and appreciation to my supervisor, Prof. Dr. Rico Sennrich, for his valuable guidance, patience, and continuous support throughout the course of my master's thesis.

I wish to extend my gratitude to Dr. Jannis Vamvas for introducing me to this topic, whose innovative idea has served as the foundation of this thesis, and whose valuable input was of great help to the process of this work.

Furthermore, I would like to thank Eric Szabó Félix and Selena Calleri for proof-reading my work. Their inputs improved its quality substantially.

Last but not least, I would like to thank my family for their constant support and understanding throughout all of my studies, and especially for the duration of this thesis.

Contents

List of Figures

List of Tables

List of Acronyms

BLEU	Bilingual Evaluation Understudy
CAT	Computer Assisted Translation
CL	Computational Linguistics
HT	Human Translation
MMT	Multilingual Machine Translation
MT	Machine Translation
NLI	Natural Language Inference
NLP	Natural Language Processing
NMT	Neural Machine Translation
Pre-NMT	Pre-Neural Machine Translation
QA	Question Answering
S	Source Language
s	Source: Source side of a translation
SVM	Support Vector Machine
T	Target Language
t	Target: Target side of a translation
WMT	Workshop in Machine Translation

List of Language Codes

bn	Bengali
cs	Czech
de	German
en	English
fr	French
ha	Hausa
hi	Hindi
ru	Russian
uk	Ukranian
xh	Xhosa
zh	Chinese
zu	Zulu

1 Introduction

Translation direction detection is the task of identifying the source side and the translation in a parallel text pair. It is a complex task that can prove challenging even to professional translators, but it is not impossible to solve. The source language leaves traces in the translation, which can be used to detect the translation direction manually as well as automatically.

The influence of the source text on its translation is a common phenomenon, which is called *interference* [Toury, 1980]. Interference belongs to a set of characteristics that are found in translations but not in original texts [Toury, 1980; Blum and Levenston, 1978; Blum-Kulka, 1986; Baker, 2019; Volansky et al., 2015]. These characteristics appear consistently and in such a large number that some of them are considered translation universals [Blum and Levenston, 1978] and characterize translations so distinctively that they have earned a name for their own language variety: *translationese* [Gellerstam, 1986]. Human translators are not the only ones susceptible to producing translationese. Automatically generated translations by machine translation (MT) systems exhibit similar characteristics, occasionally even to a larger degree [Riley et al., 2019].

Nonetheless, the characteristics of translationese can be subtle and are hardly recognized by the human eye if the translation is of good quality. This can pose problems in a variety of scenarios in research as well as in real-life scenarios. Accurately identifying these subtleties is not just a matter of academic interest; it has practical implications in fields such as legal investigations and academic integrity, as the following subsection will illustrate.

1.1 Motivation

The crucial role of the capability to determine translation direction can be exemplified with the following plagiarism allegation case [Zenthöfer, 2022a,b; Ebbinghaus, 2022]: In February 2022, two expert plagiarism investigators were commissioned to inspect a German coroner's 1987 dissertation for plagiarism. Multiple sections

from this dissertation seemed to be uncited translations of the English version of Romanian conference proceedings that were published in the German Democratic Republic (GDR) in 1983. Both investigators quickly came to the following conclusion: The dissertation was plagiarized.

The case became more complex when the coroner's alma mater investigated following this allegation. The supposed proceedings were not to be found in any university library around the world but mysteriously surfaced on a newly created account on a reseller website shortly before the investigators were commissioned. Additionally, the English version contained some prominent content errors, such as an awkwardly translated technical term that would not have been used by any expert on the subject, and contained inconsistencies that seemed laid out to fit the coroner's dissertation rather than to cohere to the actual proceedings. Furthermore, the font, pictures, and material of the physical copy were all dating back to the early 2000s rather than the 1970s or 80s [Zenthöfer, 2022a,b; Ebbinghaus, 2022].

All of this and more pointed towards the proceedings in reality being a modern-day forgery to frame the coroner. Meanwhile, the researchers involved have come to the conclusion that the overlapping parts between the two texts are translations from German into English rather than the other way around and that the plagiarism allegations arose from a carefully planned scheme with an incredible amount of effort that managed to fool even professional investigators, with substantial consequences for everyone involved [Zenthöfer, 2022a,b; Ebbinghaus, 2022].

Given the considerable investment of time and resources dedicated to detecting the translation direction in the plagiarism allegation case, one can clearly see the need for a system that can automatically detect the translation direction between parallel texts. Interestingly, such systems do already exist.

Previous research in the fields of computational linguistics (CL) and translation studies has shown that translated texts can be accurately identified automatically [Baroni and Bernardini, 2006; Ilisei et al., 2010; Koppel and Ordan, 2011; Rabinovich and Wintner, 2015; Sominsky and Wintner, 2019]. However, so far these systems have relied on traditional machine learning methods. These methods require a considerable amount of text that matches the domain of the texts that are to be identified. The conditions of resource availability and matching domain are rarely met in real-life scenarios, such that people have to resort to more tedious investigative labor, as the above-described plagiarism allegation case illustrates.

This pretext calls for a novel method to detect the translation direction without the need for excessive amounts of data. Hence, this thesis proposes an unsupervised

translation direction detection approach, which leverages translation probabilities generated by neural machine translation (NMT) models. This method leans on different streams of natural language processing (NLP) research, such as automatic translation evaluation and text similarity [Thompson and Post, 2020; Vamvas and Sennrich, 2022], where translation probabilities are used in tasks, which require high attention to stylistic detail between parallel texts – a trait that is beneficial for the translation direction detection task as well.

1.2 Hypothesis and Research Questions

The main hypothesis in this work is that NMT models produce higher probabilities when confronted with a sentence pair in the original translation direction than when confronted with the same sentence pair in the inverse direction. It is based on the intuition that by training an NMT model on pairs of original sentences and their translations for a translation task, it implicitly also learns that the characteristics of translationese are on the target side. As a result, it will exhibit higher confidence when confronted by a sentence pair that matches this stylistic asymmetry, and consequently, produce higher translation probabilities for this direction. Based on this assumption the translation direction with the higher probability will be chosen as the original one, by which one then can infer which side of the sentence pair is the original and which the translation.

The efficacy of this approach will be examined by using a diverse range of NMT models, which encompass diverse architectural (e.g.: bilingual vs. multilingual) and training data-related (e.g.: data augmentation) properties, to produce the probabilities. The evaluation will be conducted on a comprehensive dataset comprised of multiple years worth of test sets from the Conference on Machine Translation (WMT)[1]. The resulting data set encompasses multiple language pairs at different resource levels, across various domains. By including the reference translation as well as the system outputs, the dataset covers different translation strategies: human translations, pre-neural, and neural machine translations. To illustrate the real-life application capability, the approach is applied to excerpts from the dissertation and alleged conference proceedings from the above-described plagiarism allegation case.

To the best of my knowledge, the approach proposed in this work is novel and subjected to experimental validation for the first time. This being considered, the following research questions shall be answered within this thesis:

[1]The main annually reoccurring conference on machine translation and machine translation research, including shared tasks on different aspects of machine translation [Harison, 2023].

1. Can translation probabilities be used for translation direction detection?
2. Can potential biases be mitigated by normalizing the translation probabilities?
3. What other properties do translation probabilities display in terms of translation direction?

1.3 Thesis Structure

In this first chapter, the topic translation direction detection is briefly introduced and the motivation for research in this specific field is exemplified. Furthermore, the leading hypothesis and research questions for this work are presented. Chapter 2 establishes a background in translation studies, provides the terminology for the following work, and puts it into a linguistically motivated frame. The relevant computational concepts for this field are introduced in an additional section within that chapter. In Chapter 3, the findings of previous work on the topic of translation direction detection are presented and connections between this thesis and related work are established. The methodical specificities are described in Chapter 4, where details on the approach, models, and data that are used in this work are provided. Chapter 5 presents the outcome of this work's research, which will be discussed in Chapter 6. Finally, the work is concluded in Chapter 7.

2 Background

As Section 1 has indicated, not being able to identify the translation direction can have detrimental real-life consequences. However, the motivation for the work described in this thesis is not restricted to solving novel-worthy criminal schemes. This work is ultimately performed from a research perspective in the field of CL. More precisely, it shows how computational methods from the field of NLP can be used on a problem that is rooted in the linguistic theories of translation studies. Therefore, an introduction into these topics is necessary to establish a common background. The first section of this chapter briefly introduces the fundamental concept of MT and gives a short overview of common techniques and highlights those which are relevant to understanding the methods used in this thesis. After the technical background, this chapter delves deeper into the linguistic theory behind translations with a brief historical overview of translation studies that focus on translationese as a specific register. The findings of the latter in terms of the characteristics of translationese are outlined. And lastly, the scientific relevance of this topic is emphasized by summarizing the consequences of translationese in the areas of NLP, more specifically MT, and the interdisciplinary field of forensic linguistics.

2.1 Neural Machine Translation

This first section is dedicated to exposing the fundamental concepts of NMT systems, which are not only the current state-of-the-art in MT but also lie at the core of the methodical part of this thesis. Initially, a brief overview of the core concept of MT is described and is tied to key concepts of a classic bilingual MT system. Afterward, insight into the more recent multilingual machine translation (MMT) systems will be provided, including the entailing advantages, disadvantages, and new possibilities that they bring along.

2.1.1 Bilingual Machine Translation

The task of an NMT system is to produce the most probable translation (T^*) in a specific target language given an input (S) in a specific source language [Bahdanau et al., 2016; Sennrich, 2022]:

$$T^* = \underset{T}{\operatorname{argmax}}\, P(T|S) \tag{2.1}$$

To achieve this, a neural model, consisting of two components – an encoder and a decoder –, is trained on a parallel corpus of the chosen language pair to maximize the conditional probability of the sentence pairs. [Bahdanau et al., 2016; Gehring et al., 2017; Vaswani et al., 2017].

Bilingual MT is the most basic form MT can take, by incorporating a single language pair and being able to translate in one translation direction only. Training is usually done on vast amounts of parallel data of one language pair, requiring millions of sentences in each language to achieve usable results. As a consequence, NMT systems only perform well on language pairs, where large amounts of data are available, and perform poorly on lesser-resourced language pairs [Mohammadshahi et al., 2022].

Efforts are made to find a solution to improve performance for low-resource languages. Examples include back-translation, where monolingual data from the target side is automatically translated into the source side and the resulting parallel corpus is used as additional training data [Sennrich et al., 2016; Edunov et al., 2018]; unsupervised NMT, which are translation systems trained without parallel data [Artetxe et al., 2017; Lample et al., 2018; Garcia et al., 2020; Ko et al., 2021]; multi-task learning [Domhan and Hieber, 2017], or, more recently, MMT [Firat et al., 2016; Johnson et al., 2017; Zhang et al., 2020; Thompson and Post, 2020; Fan et al., 2021; Tang et al., 2021; Goyal et al., 2021; Mohammadshahi et al., 2022], which will be discussed in more depth in the following subsection.

2.1.2 Multilingual Machine Translation

MMT models, unlike their single-pair counterparts, are capable of handling multiple language directions within a single model. This substantially reduces operational costs during training and deployment in production systems [Arivazhagan et al., 2019]. Furthermore, due to joint training techniques that transfer knowledge from high-resource to lesser-resourced languages, a positive effect on the translation per-

formance of low-resource language pairs is found [Zoph et al., 2016; Nguyen and Chiang, 2017]. This goes to the extent that translations are enabled for language pairs, for which no parallel data was seen during training: a process called zero-shot translation [Aharoni et al., 2019; Arivazhagan et al., 2019]. A side effect of enabling zero-shot translations is that one can now "translate" from the source language back into the source language, effectively performing paraphrasing [Thompson and Post, 2020], a topic, which is further discussed in Section 3.2.

A notable downside of MMT models is their large size. If they are supposed to reach the translation performance of their single-pair counterparts, the model capacity has to be increased substantially [Aharoni et al., 2019; Arivazhagan et al., 2019; Zhang and Toral, 2019], which in turn calls for larger multilingual datasets – making the MT problem in this setting even more resource-intensive. One consequence is that oftentimes specialized hardware is required to even use the models. To bypass this problem to some extent, some MMT models focus on translating only from and to English. Since this does not resolve the issue, efforts go into finding novel data mining strategies, leveraging existing data augmentation techniques, and model scaling to create MMT systems that include language pairs without having to center around English [Fan et al., 2021].

One example, of a model scaling technique is knowledge distillation. A smaller, more compact model (student model) can be trained with the output or intermediate representations of the larger initial model (teacher model). Although this technique requires a large model to begin with, it is quite popular, since it decreases the time and memory consumption of the student model compared to its teacher, while at the same time keeping the performance on par. This enables broader public accessibility. [Hinton et al., 2015; Kim and Rush, 2016; Hu et al., 2018; Akula et al., 2022; Mohammadshahi et al., 2022].

This section illustrated how the translation problem is modeled, and which techniques in NMT can be used to solve it by encompassing both single-pair systems and the expansive capabilities of multilingual models. It has also highlighted the high costs at which state-of-the-art MT systems come and which efforts have been made to reduce them. Given the increased presence of MT systems and their accessibility, efforts have been made to expand the use of MT systems to other tasks, which is also the aim of this thesis.

2.2 Translationese

After establishing the technical background, this section focuses on theories from translation studies. Beginning with terminological explanations and how they came about in the field of translation studies, this section leads to the definition of the problem that is to be solved within this thesis. In the following subsection, the results of translation studies for translationese are briefly summarized to emphasize the sensitivity that is expected of a system to automatically identify translation direction. And lastly, the consequences of translationese in different fields are discussed to highlight the need of such a system.

2.2.1 Terminology, History and Problem Definition

Translationese is a term originally coined by Gellerstam in 1986 to describe the distinctive linguistic characteristics that typically appear in translated texts, distinguishing them from texts originally written in that language. Gellerstam finds that this is a statistically detectable phenomenon, which suggests that translation is not simply a matter of swapping words from one language to another but involves complex interactions between the source and target languages and leaves identifiable marks on the translated work, introducing an asymmetry between the two types of texts [Volansky et al., 2015].

The concept of translationese evolved into the broader field of translation studies, where it became part of a theoretical framework known as "translation universals" or "laws of translation." Gellerstam's initial observations were expanded by scholars such as Toury [1980, 2012], who proposed two fundamental laws: the law of interference, which refers to the remnants of the source text within the translation, and the law of growing standardization, which is the tendency to conform the translation to the norms and idioms of the target language [Volansky et al., 2015].

Throughout the 1990s, the research on translationese gained momentum as Baker [1993] advocated for the use of comparable corpora — collections of translated texts set against non-translated texts of similar genre and time frames to identify these characteristics empirically. She proposed that certain features are universal in the translated text, beyond the influence of specific language systems. Chesterman et al. [2004] later refined the categorization of universals into two types: S-universals (source-text related) and T-universals (target-text related), necessitating different types of corpora for study. Parallel texts for S-universals and comparable texts for T-universals.

The continuing study of translation universals with corpus linguistic methods led to an increase in resources, such that the research in this field became increasingly empirical and computational [Volansky et al., 2015]. Not only have these resources been used to reveal the characteristics of translated texts, but by 2006, the first automatic text classifier was built to distinguish original from translated texts [Baroni and Bernardini, 2006]. The system is based on the previously discovered translation universals, which in this context are called by the computationally connoted term *features* [Volansky et al., 2015]. It is worth noting that this first system was based on a comparable corpus, which, as Section 3 will illustrate, most of the research has been focused on [Baroni and Bernardini, 2006; Van Halteren, 2008; Kurokawa et al., 2009; Ilisei et al., 2010; Ilisei and Inkpen, 2011; Koppel and Ordan, 2011], with only a minority dedicated to parallel corpora [Sominsky and Wintner, 2019].

In the realm of automatic text classification, the choice of corpora changes the task at hand: If a comparable corpus is used, the task is usually defined as translation detection. In translation detection, the task is to identify translations among original texts, usually, in a monolingual setting. If a parallel corpus is at the core of the study, on the other hand, the task is framed as translation *direction* detection: The same text is available in one or more languages and the classifier's task is to identify from which source language into which other target language(s) the translation has been performed. The work described in this thesis is framed as a *translation direction detection* task on parallel corpora.

2.2.2 Characteristics of Different Types of Translations

As outlined in Subsection 2.2.1, the differences between original and translated have been studied extensively. However, in reality, we are confronted by more than simply one type of translation. A myriad of translation strategies exist, and all of them influence the translations in terms of the degree of translationese and asymmetry between source and target text. The simplest distinction can be based on whether the translation was produced manually by a human or automatically via a machine translation system. Automatically generated translations can be grouped further based on the model's architecture type: NMT or statistical machine translation (SMT). The following points summarize previous research's findings on the three types of human, statistical, and neural translations:

- **Human Translation (HT)** The characteristics found in human translations are interference, simplification, explicitation, and normalization. Interference has already been mentioned above, as it is the most obvious sign for transla-

tionese. It is the phenomenon that the source language influences the target language in the translation in its word order [Toury, 1980]. Simplification is the phenomenon that translations tend to be simpler than the source text [Blum and Levenston, 1978; Baker, 1993; Laviosa, 2021]. Explicitation refers to explicitly spelling out information that was implicit in the source [Baker, 1993], while normalization refers to the phenomenon that translators show a preference for grammaticality rather than proximity to the source [Baker, 1993; Toury, 2012].

- **Statistical Machine Translation (SMT)** Since MT systems are trained on parallel texts that consist of human translations on the target side, it is no surprise that MT also shows signs of translationese. Those signs, however, differ from those observed in HT. In general, SMT tends to oversimplify the translation when compared to HT as well as NMT [Bizzoni et al., 2020]. It also exhibits more lexical variety and has a higher lexical density than NMT. This means that SMT translations have a higher proportion of content words (nouns, verbs, adjectives, adverbs) and are less repetitive [Toral, 2019].

- **Neural Machine Translation (NMT)** Research has shown that NMT translations exhibit more interference than (HT) by following the source word order more strictly and being more monotone [Toral and Sánchez-Cartagena, 2017; Burlot and Yvon, 2019; Zhou et al., 2019; Voita et al., 2021]. Compared to SMT, NMT systems produce more changes in word order, however, the reorderings are closer to the reference translation than the ones of SMT systems. Additionally, the output from NMT systems is generally more fluent [Toral and Sánchez-Cartagena, 2017; Toral, 2019; Bizzoni et al., 2020], which aligns with the normalization phenomenon in HT. Further findings have shown that model distillation can make translationese even more prominent [Riley et al., 2019; Akula et al., 2022].

In summary, the linguistic characteristics of translationese can be very subtle and nuanced for all three categories. However, when the focus is shifted towards MT, these subtleties become more pronounced – but with clear differences between SMT and NMT.

In this thesis, one of the aims is to delve deeper into how these differences of translationese are reflected in NMT probabilities and, ultimately, how they affect the translation direction detection. Hence, the test dataset in this work is split into these three categories to provide a comprehensive comparison: HT, NMT, and pre-

NMT[1]. Translations that would belong into an intermediate category, for example because they have been produced using machine translation systems and human post-editing[2], or the distinction between native and non-native[3] text are disregarded in this work for simplicity's sake.

2.2.3 Consequences for Natural Language Processing Research and Applications

So far, this section has primarily been focused on the description of translationese and tracing its history in research. Now that this foundation has been formed, it is worth exploring the underlying motivations behind identifying the translation direction, or translationese in a broader sense. This subsection will explore those motivations in the realm of NLP research and its applications – stressing its significance beyond the theoretical confines of translation studies.

Machine Translation: Most evidently, MT is a field in NLP, where translationese has direct as well as indirect impact. Early work in the field of translation (direction) detection was motivated to a large extent by improving SMT systems. These systems greatly benefited from parallel training data, in which the translation direction aligns with their translation direction [Kurokawa et al., 2009].

For NMT, there have been indications that systems, which are trained on source original data, outperform systems trained on data with reverse translation direction or mixed settings if evaluated with source original test data by using an automatic reference-based metric like BLEU [Papineni et al., 2002; Sominsky and Wintner, 2019; Bogoychev and Sennrich, 2019]. If the translations are evaluated manually, however, the system that was trained on source original parallel text only was rated worst, especially in terms of fluency [Bogoychev and Sennrich, 2019]. For MT evaluation a consensus has been reached that the inclusion of reverse-created test data confounds the results of automatic evaluation metrics [Toral et al., 2018; Läubli et al., 2018; Freitag et al., 2019; Zhang and Toral, 2019; Graham et al., 2020]. This led to the discarding of reverse-created test data in all WMT test sets from 2019 onwards [Barrault et al., 2019].

Additionally, a more recent study has brought to attention that mismatches in

[1]Includes all non-neural systems, such as SMT and rule-based systems, before the transition to NMT in the field

[2]For more information on the characteristics of translations that have been automatically produced and manually post-edited see: Toral [2019]

[3]For more information on this distinction see: Rabinovich et al. [2016]

directionality can be even more fine-grained, suggesting that mismatches between the test data and the model, training data and the test data, and between the data (both training and test set) and the model all have different effects on the MT performance and should be treated with care [Ni et al., 2022].

Cross-Lingual Benchmarks and Applications: Other areas of NLP that are affected by translationese are those that require multilingual datasets either to train or to test their systems. Multilingual datasets can be either created by collecting and annotating different data in different languages separately [Clark et al., 2020], or by collecting one dataset in a single language and translating it into other languages afterward to create parallel texts [Artetxe et al., 2019].

While the majority of these datasets are designed for assessment purposes, the goal of cross-lingual transfer learning is to utilize the extensive datasets available in one language, usually English, to develop multilingual models. These models are created to effectively generalize across various languages. This is achieved by incorporating translations either in the training phase or within the test set.

Examples, where cross-lingual transfer learning is used, are natural language inference (NLI) and question answering (QA). There, research has found that performance loss that was previously assigned to language transfer was rather due to the mismatch between original and translated texts, hence their systems' performances are assumed to be underestimated by current benchmarks [Artetxe et al., 2020].

Forensic Linguistics: All of the aforementioned subjects have real-life applications, that to some extent influence people's daily lives. However, the field that carries the most dramatic consequences in real life, is arguably forensic linguistics, where linguistic techniques are applied to legal cases, such as the plagiarism allegation case, outlined in Section 1, as well as to research [Olsson and Luchjenbroers, 2014]. With the world's increasing digitalization the number of computational approaches to forensic linguistics has risen steeply as well. Early approaches focused on corpus linguistic analyses and treating corpora as large bodies of linguistic evidence, but methods are becoming more computationally sophisticated [Olsson and Luchjenbroers, 2014; Sousa-Silva, 2018].

The main foci of computational forensic linguistics lie in tasks such as authorship analysis, profiling, stylometry, and plagiarism detection, which are occupied with identifying the author of a text. When dealing with multilingual texts or translations, understanding the direction of translation can provide insights into the author's linguistic background and potentially their identity – a topic called cross-

lingual authorship identification [Sousa-Silva, 2018].

As for plagiarism detection, translation takes an important role since translation-based plagiarism poses one of the biggest challenges in the field [Maurer et al., 2006], as it was illustrated in Chapter 1 with the plagiarism allegation case. The detection of translation-based plagiarism oftentimes includes manual linguistic analyses as well as comparison translations from multiple different machine translations, which are taken into account to identify the translation direction [Sousa-Silva, 2018]. At this point, it is also worth mentioning that forensic linguists oftentimes have to rely on small "DIY corpora", since most publicly available corpora do not match the different styles that are encountered in the texts that are subject of investigation [Sousa-Silva, 2018].

To summarize, in the field of MT, the influence of translation direction is evident, both in terms of system training and evaluation, with studies highlighting the complexity and varied effects of translationese on NMT system performance and the evaluation thereof. Similarly, for cross-lingual benchmarks and applications like NLI and QA, the role of translationese in creating multilingual datasets is significant, affecting the accuracy and reliability of these systems. With forensic linguistics, those challenges seem to culminate in a field with profound real-life impact, where the detection and understanding of translation direction is a most valuable asset that is either difficult to obtain through manual labor or not obtainable at all due to a lack of sufficient data. Hence, in all three of those fields translation direction detection would be an asset, particularly if it does not require large volumes of text data

3 Related Work

The related work can be divided into two parts by subject and method: previous research on translation direction detection and previous work with translation probabilities. Both areas will be outlined below, not only highlighting the novelty of this thesis but also demonstrating how the former research area could overcome caveats by using methodologies of the latter.

3.1 Translation (Direction) Detection

Within the scope of translation detection, numerous successful systems have been documented and published over the past decade and a half. The task was formulated in different ways, such as the classification of original and translated texts in monolingual settings, detection of a source language in a multilingual setting, or as determining the translation direction in a bilingual setting – as it is done within this work as well. This section is intended to provide a comprehensive overview of the various systems that have been employed to solve those tasks, their performance, and an analysis of their advantages and drawbacks. Given that most of the work has focused on *supervised* detection, a summary of this area will be presented initially. Afterwards, the more limited yet still significant endeavors for *unsupervised* detection will be described.

3.1.1 Supervised Methods

In 2006 Baroni and Bernardini introduced an early method for translation detection to differentiate original texts from translations by employing an ensemble of support vector machines (SVMs) [Joachims, 1998, 1999]. Their experiments were conducted on a monolingual, comparable corpus of Italian geopolitical news articles at document level (here: article). Drawing on translation studies and the characteristics of translationese (see Section 2.2), they used linguistic features such as the distribution of function verbs, personal pronouns, and adverbs. This approach achieved an

accuracy of well above 80%, with follow-up experiments demonstrating the system's reliability to surpass even the human judgment of professional translators.

In subsequent years, similar methods were adopted and the results improved, showing the methods' capabilities in various languages and language pairs [Van Halteren, 2008; Kurokawa et al., 2009; Ilisei et al., 2010; Ilisei and Inkpen, 2011; Koppel and Ordan, 2011], including experiments in Hebrew, a lesser-resourced language [Avner et al., 2016]. The main differences between the systems were mostly tied to the choice of linguistic features. Examples of linguistic features that have proven useful for this task were part of speech tags or lemmas [Baroni and Bernardini, 2006; Van Halteren, 2008; Kurokawa et al., 2009], function word frequencies [Koppel and Ordan, 2011], character-level features [Popescu, 2011; Avner et al., 2016], surface and lexical features [Ilisei et al., 2010; Volansky et al., 2015], syntactic features [Ilisei et al., 2010; Rubino et al., 2016] and morpheme-based features [Avner et al., 2016; Volansky et al., 2015], to name a few.

More notable work has been published by Sominsky and Wintner [2019], who accomplished overcoming the restriction of having to use large text chunks to obtain reliable results for translation direction detection. They achieve accuracy scores of over 80% at phrase and sentence level in a single-domain setting, while their best-performing system reaches a 72% accuracy for one language pair (French↔English) in a multi-domain setting with three domains. These results are achieved by transitioning from traditional machine learning to neural systems – while still relying on linguistic features. Their research contributed novel insights regarding the effectiveness of the task on short text segments and by formulating the problem of translation direction detection on parallel text instead of translation detection on comparable corpora.

Although supervised classification with linguistically motivated feature engineering proves to be effective in the task of translation detection in the work described over the last few paragraphs, recent work introduces new approaches. Pylypenko et al. [2021] describe the transition to systems that rely on representation learning (embeddings) instead of linguistic features and that BERT-based models outperform the traditional SVM classifiers. One motivation behind this transition, besides the scientific interest to explore state-of-the-art methods, lies with manually designed features perhaps being "partial and non-exhaustive in a sense that they are based on our linguistic intuitions, and thus may not be guaranteed to capture all discriminative characteristics of the input data seen during training", as stated by Pylypenko et al. [2021, p. 8596]. While these newer models deliver better results (Accuracy: 84%-90%), they operate at paragraph level, making them more innovative but less

effective on smaller text chunks compared to the work described in Sominsky and Wintner [2019] [Pylypenko et al., 2021].

Despite the success of supervised translation (direction) detection systems, there are notable limitations to consider. Firstly, the supervised approach of the text classification task requires a lot of labeled data to train systems that achieve the results described above. Consequently, the systems are restricted to cases where large labeled corpora are readily available or can be collected before building the system. Secondly, once the systems have been built and are ready for application, they reach the reported accuracies only in single-domain settings. When the number of domains was increased the accuracy has shown to drop. Lastly, each of the systems mentioned in the work above has been tested on a fairly narrow set of languages. This highlights the necessity for further research regarding the applicability to a wider linguistic landscape and the adaptability of a translation (direction) detection system to multiple domains. Efforts to address the resource limitations of supervised systems have led to the development of unsupervised systems, which will be discussed in the following subsection.

3.1.2 Unsupervised Methods

Notwithstanding their success, efforts to address the resource limitations of supervised systems have led to the development of unsupervised systems. The most notable successes are the work of Rabinovich and Wintner [2015] and Nisioi [2015].

In 2015, Rabinovich and Wintner propose a method for unsupervised detection of translationese based on clustering algorithms. In their work, they use K-Means clustering [Lloyd, 1982] with a principal component analysis (PCA) [Jolliffe, 2003] with a set of linguistic features that seem to capture the differences between original texts and translationese for French→English. With a subsequent majority voting of the individual feature's results, they achieve accuracies on in-domain datasets high above 80% and 90%. They state that in a mixed-domain setting, a simple clustering is not enough, since the domain signal is too dominant and overshadows the signal of translationese. To overcome this hurdle, they introduce a two-phase clustering method, in which they first cluster the texts into domains and then cluster these clusters in a second step into either translations or original texts. With this method, they outperform previous attempts of supervised mixed-domain experiments for translation direction detection by reaching accuracy scores up to 90%. These experiments, however, were conducted on only 2-3 domains, where the example with three domains reaches only 67%.

A second unsupervised method was proposed in 2015 for the task of authorship attribution. Nisioi [2015], in his work on unsupervised translation detection, describes how distinguishing between original and translated literary work can succeed with a corpus by the multilingual author Vladimir Nabokov. Similar to the work described above, Nisioi [2015] utilizes a clustering algorithm. However, instead of using K-Means, he employs a generalization of Ward's method [Ward, 1963], which starts with one cluster per document and, consecutively, merges two clusters based on a minimum distance. Same as above they use lexical features as described in Volansky et al. [2015], in addition to features that have been tied to authorship attribution [Juola, 2008; Koppel et al., 2009]. He first collects the Russian-English corpus of the author's work, which has largely been translated by the author himself – his earlier work from Russian to English and the other way around for his later work. The corpus is parallel on document-level (here: novel). Then, Nisioi [2015] applies his method to both languages, treating them as two corpora, and argues that if the approach results in similar results in both languages, he can be confident that the results are based on distinctions between translations/translator and original/author. He works on entire documents as well as on chunks of 2000 tokens, achieving F1-scores over 90% in all cases but English on chunk level, where they report a drop to 60-80%.

However, both of these systems have their limitations. Even though the amount of data decreases drastically in clustering compared to a supervised setting, because the need for training data dissolved, Rabinovich and Wintner [2015] describe that their system works best with an increasing number of text chunks. These text chunks consist of at least 250 and up to 2000 tokens, reaching satisfactory results at 1000-token chunks. The same tendencies are seen in the work of Nisioi [2015]. His systems perform on full-length novels, while cutting them into smaller chunks significantly decreases their performance.

It is also worth noting that clustering algorithms such as K-means are highly susceptible to unbalanced data. This means that if the proportion of translations and original text clusters are not the same, the results might be skewed. Even though Rabinovich and Wintner [2015] mention experiments, in which the dataset was not balanced, it seems that they only tested a 2:1 ratio. Furthermore, the two clusters obtained by this method still need to be labeled as either a cluster of translations or original texts, which requires either manual inspection of the texts in each cluster or a labeling algorithm [Rabinovich and Wintner, 2015].

Additionally, both methods rely on manually extracted linguistic features, which require a thorough occupation of the linguistic properties of the texts and introduce a labor intensity that may not scale well. This method does not leverage the more advanced and automated feature extraction techniques that have become the norm in the years following their publication and have shown to be more effective in a multitude of tasks, as it has been described in Section 3.1.1.

A last point worth mentioning is that even though the two-phase clustering setup for mixed-domain data achieves good results, it is shown in the paper that it decreases drastically if three domains are mixed. Presumably, these accuracies would decline even further if more domains were involved. Additionally, the two-phase method requires the number of domains in the dataset to be known beforehand – which is again unlikely to be known in real-life scenarios. The latter point does not arise in Nisioi [2015], however, an effect is seen with the influence of authorship in their results which is comparable to the domain overshadowing effect in Rabinovich and Wintner [2015].

In summary, the unsupervised methods for translation detection introduced in Rabinovich and Wintner [2015] and Nisioi [2015] prove to be effective but face practical challenges for real-world application. Their dependency on large text chunks, susceptibility to data imbalance, the necessity for manual labeling, and reliance on feature extraction limit their effectiveness outside controlled experimental settings. Furthermore, while their approaches show effectiveness in specific scenarios, they do not fully address the complexity of real-life datasets, which are often more varied in terms of domain and shorter in text length. These constraints underline the need for more robust, flexible, and scalable approaches to accommodate the complex and diverse nature of translation detection in practical, real-world scenarios, which could be achieved by adjusting the problem formulation to a translation *direction* detection and choosing so far unexplored methods for unsupervised classification.

3.2 Translation Probabilities

Methodically, the work in this thesis relies on translation probabilities. Translation probabilities are a by-product of NMT systems, that have proven to be valuable data points in settings where parallel texts are compared. An overview of which tasks have been solved by using translation probabilities is given in the following subsection.

3.2.1 Filtering of Noisy Parallel Corpora

Junczys-Dowmunt [2018] proposed the use of translation probabilities to filter noisy parallel data. Noisy parallel corpus filtering is a task, in which misaligned or poorly translated parallel texts are removed from a parallel corpus. The author described the approach of modeling bilingual adequacy between parallel texts with cross-entropy scores that were generated by two NMT models that have inverse translation directions and were trained on the same data in inverse directions. For each sentence (x, y) he produced a single score $f(x, y)$ as the product of partial scores $f_i(x, y)$:

$$f(x, y) = \prod f_i(x, y) \tag{3.1}$$

Thereby generating the values $H_A(y|x)$ and $H_B(x|y)$ for the conditional cross-entropy of the probability distributions $P_A(\cdot|\cdot)$ and $P_B(\cdot|\cdot)$, where x corresponds to the source and y to the target for models A (trained in the original direction) and B (trained in the inverse direction). Once the conditional cross-entropy scores were generated, Junczys-Dowmunt [2018] calculated a score "to find maximal symmetric agreement (minimal absolute difference) of dissimilar distributions (two translation models over inverse translation directions) trained on the same data (same parallel corpus)", as written by Junczys-Dowmunt [2018, p. 3]. Since this calculation produces only positive values with 0 being the best possible score, he negated and exponentiated them to turn them into values between 0 and 1, where 1 is the best possible score:

$$adq(x, y) = exp(-(|H_A(y|x) - H_B(x|y)| + \frac{1}{2}(H_A(y|x) + H_B(x|y)))) \tag{3.2}$$

Using this score in addition to a language filter and a cross-entropy difference filtering [Moore and Lewis, 2010], Junczys-Dowmunt [2018] achieved successful results in the WMT18 shared task on parallel corpus filtering and thereby proved the translation scores' comparison capabilities.

3.2.2 Automatic Machine Translation Evaluation and Zero-Shot Paraphrasing

Another branch of text similarity research focuses on the use of bilingual corpora for paraphrase extraction, where paraphrase probabilities (formulated as translation probabilities) of phrases ($P(y|x)$) and their inverse counterparts ($P(x|y)$) are used

to rank paraphrases for a given phrase [Callison-Burch et al., 2006; Mallinson et al., 2017]. In 2020, Thompson and Post brought translation probabilities back into the field of MT by proposing a metric for the automatic evaluation of machine translations that is based on these paraphrase probabilities of parallel text.

Thompson and Post [2020] started by building a system for paraphrasing. Essentially, the system was an MMT model trained on parallel text. However, they treated paraphrasing in this paper as a zero-shot translation task – translating from source language to source language (see Section 2.1.2).

Once their paraphrasing system was trained, Thompson and Post [2020] utilized it to estimate the probability of outputs from MT systems based on their corresponding human references. These probabilities could then be interpreted as the performance of an MT system on a given sentence pair. Similarly to Junczys-Dowmunt [2018], they first generated partial scores that were at token level and needed to be combined for a sequence-level score. They consider two the following methods to do so [Thompson and Post, 2020]:

$$G(y|x) = \sum_{t=1}^{|y|} \log p(y_t|y_{i<t}, x) \tag{3.3}$$

$$H(y|x) = \frac{1}{|y|} G(y|x) \tag{3.4}$$

Equation 3.3 describes the aggregation of token-level probabilities into one sequence probability by summing up the individual values, while Equation 3.4 normalizes these values by averaging by sequence length. For the case of reference-based evaluation, y denotes the system output that is to be evaluated while x denotes its human reference. Thompson and Post [2020] explored the scores for the other direction as well (y: human reference; x: system output) and averaged the scores of both directions to penalize missing information on either side. For their quality estimation metric, they considered the system output as y and the source sentence as x.

Thompson and Post's experiments show that their MMT system can be used as a lexically/syntactically unbiased, multilingual paraphraser, and its probabilities (as seen in Equation 3.4) can be interpreted as MT quality estimation metrics, achieving state of the art results in both areas. This highlights the usefulness of NMT systems outside of the realm of basic machine translation on one hand, and the information richness of the estimated probabilities by such systems on the other.

3.2.3 Text Similarity Measures

The two previously described methods can be described as text similarity measures. More recently, Vamvas and Sennrich [2022] analyzed different translation-based similarities and they proposed a new measure as well as corresponding normalization techniques.

Vamvas and Sennrich [2022] compared three different translation-based similarity measures: Firstly, direct translation probability in a bilingual setting (see Equation 3.5, left), which was based on Junczys-Dowmunt [2018] and Thompson and Post [2020]. Secondly, the probability of pivot translation (Equation 3.6, left), where the paraphrastic similarity was estimated by translation to a pivot language [Mallinson et al., 2017]. And thirdly, translation cross-likelihood, for which a translation into any language was generated and an estimation was made of how likely it was that the generated sequence was a translation of the source sentence (see Equation 3.7, left). Each of these probabilities was normalized before being used as a similarity measure [Vamvas and Sennrich, 2022]:

$$P_{direct}(y|x) = p(y|x) \rightarrow \text{NMTSCORE-direct}(y|x) = \frac{p(y|x)}{p(y|y)} \tag{3.5}$$

$$P_{pivot}(y|x) = p(y|x') \rightarrow \text{NMTSCORE-pivot}(y|x) = \frac{p(y|x')}{p(y|y')} \tag{3.6}$$

$$\text{Cross-likelihood}(y|x) = p(x'|y) \rightarrow \text{NMTSCORE-cross}(y|x) = \frac{p(x'|y)}{p(x'|x)} \tag{3.7}$$

The authors call the first two normalization strategies that are seen on the right side of Equations 3.5 and 3.5 *reconstruction normalization*, since $p(y|y)$ is the probability that the sentence remains identical when zero-shot paraphrasing is performed, as described in Section 3.2.2. The third normalization strategy (Equation 3.7, right) follows a normalization technique by Mallinson et al. [2017] [Vamvas and Sennrich, 2022].

The scores are tested on a variety of languages and language pairs, achieving high accuracies in multilingual paraphrase identification. An ablation study also shows that the high accuracies are to a large extent due to the probability normalization – especially for the first two scores. Altogether, this once again shows "the usefulness of NMT translation probabilities for similarity tasks that require high attention to detail" Vamvas and Sennrich [2022, p. 206].

In conclusion, previous work on translation probabilities has shown that NMT systems can be used for other tasks than mere translation. They have shown significant results in various applications, including filtering noisy parallel corpora, evaluating machine translation output, and other text similarity measures. These results suggest that the probabilities, generated by NMT systems, are not only meaningful but also rich in information about single sentences and their translations or paraphrases. This includes aspects like word order, which has proven useful in diverse settings where parallel texts need comparison or evaluation. Given these characteristics, it seems promising to employ translation-probability-based methodologies in the translation direction detection task, which would not only diminish the amount of data that is needed to solve the task but also eliminate the necessity for training a model for this specific task.

4 Methods

4.1 Unsupervised Translation Direction Detection

The main task in this work is to identify the translation direction between a language S and a language T given a parallel sentence pair and, consequently, establish which side is the original and which the translation. This is to be achieved by comparing the conditional translation probability $P(t|s)$ of a sentence pair, produced by an NMT model $M_{S \to T}$ in one translation direction $S \to T$, with the conditional translation probability $P(s|t)$ in its inverse direction $T \to S$, produced by a model $M_{T \to S}$. The higher probability is assumed to be assigned to the sentence pair, for which the original matches the source s and t is the translation. Figure 1 illustrates this process.

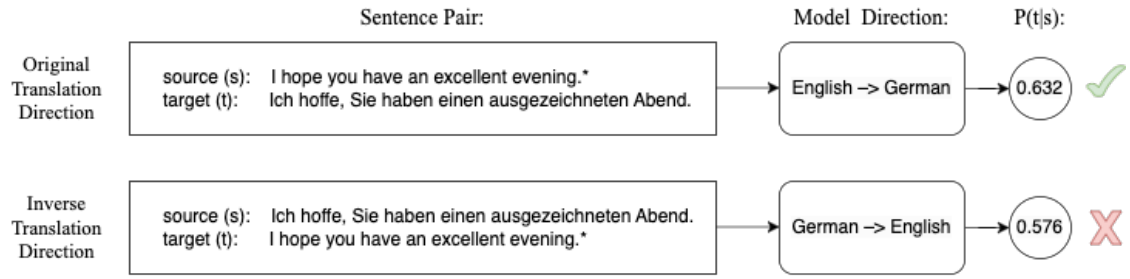

Figure 1: Illustration of the translation direction detection workflow and decision process. The original sentence is marked with an asterisk (*).

4.1.1 Translation Probabilities for Translation Direction Detection

NMT models are usually trained to minimize cross-entropy on a training set [Junczys-Dowmunt, 2018]. The cross-entropy One can use NMT models to generate the cross-entropy scores of the target given the source sentence by providing a source and a target. However, these scores are first generated partially, such that each token in the sentence is assigned a score. This aligns with Equations 3.4 and 3.3 before they are combined to sentence-level scores [Junczys-Dowmunt, 2018; Thompson and Post, 2020]. To aggregate the partial scores, the approach by Thompson and Post [2020] was followed by averaging the scores:

$$H(t|s) = \frac{1}{|y|} \sum_{t=1}^{|y|} \log p(y_t|y_{i<t}, x) \tag{4.1}$$

However, this score is not yet interpretable as a conditional probability. To achieve this, the approach follows Junczys-Dowmunt [2018], by negating and exponentiating the score to obtain values between 0 and 1:

$$P(t|s) = exp(-H(t|s)) \tag{4.2}$$

This process was performed once in the translation direction $S \rightarrow T$ and once in the inverse direction $T \rightarrow S$ using the corresponding NMT models.

Since this approach is based on the hypothesis that translation probabilities obtained from source original sentence pairs are higher, $P(t|s)$ and $P(s|t)$ were compared to see which conditional probability is higher to perform the binary decision for the translation direction detection:

$$\text{Original Translation Direction} = \begin{cases} S \rightarrow T, & \text{if } P(t|s) > P(s|t) \\ T \rightarrow S, & \text{otherwise} \end{cases}$$

4.1.2 Translation Probability Normalization

Previous approaches suggested that raw translation probabilities are not ideal to base a decision on and normalizing the probabilities first should be considered [Mallinson et al., 2017; Vamvas and Sennrich, 2022]. Hence, the normalization approach for direct probabilities in Vamvas and Sennrich [2022] was followed to normalize the conditional probabilities that were obtained as described in the pre-

vious subsection:

$$P_{norm}(t|s) = \frac{P(t|s)}{P(t|t)} \tag{4.3}$$

And its inverse:

$$P_{norm}(s|t) = \frac{P(s|t)}{P(s|s)} \tag{4.4}$$

In the context of this work, the zero-shot translation probabilities were explored as a possibility to balance potential biases of the NMT model towards one translation direction in the sentence pair and to improve the overall classification performance. Since zero-shot probabilities can only be generated by MMT models [Johnson et al., 2017], the normalization technique was only applied to those experiments, for which MMT models were used to generate translation probabilities. The single-pair models were therefore omitted in the normalization experiments.

4.1.3 From Sentence to Document Level

The initial experiments were conducted at sentence level. The datasets used in this work provide information on document affiliation, such that document-level results can be inferred based on the results at sentence level. This was done by aggregating the predicted sentence-level results of a document and choosing the document-level label based on a majority vote on the predicted sentence-level labels (hard voting) [Brownlee, 2021].

4.2 Models

The selection of the NMT models that were used to generate the probabilities in the experiments was based on multiple conditions: First, the models had to be publicly available to use them in the experiments for this work as well as in a potential reproduction thereof. Second, the models should cover as many languages and directions from the collected dataset as possible – including low-resource language pairs – to test the robustness and universal applicability of the approach. The straightforward choice to meet this condition was to utilize MMT models. However, a range of architectures, data settings, and model sizes should be covered to allow room for exploration and comparison. For that reason, bilingual single-pair models

were also included in the selection. For the bilingual models, one model had to be available for each translation direction. The final selection of models consists of the MMT models NLLB (in different sizes), M2M100, SMaLL100, and 14 single-pair Opus MT models.

4.2.1 Opus-MT

The Opus-MT models are a set of mostly bilingual NMT models [Tiedemann and Thottingal, 2020]. For each language pair in the dataset two bilingual Opus models were chosen – one for each translation direction. The aim was to choose models that are equal (or as similar as possible) in terms of architecture, size, and training dataset for both translation directions. Unfortunately, not every Opus model is accompanied by a research paper explaining its architecture, such that most of the information on the models stems from their respective model cards and a general paper [Tiedemann and Thottingal, 2020], which describes the whole Opus-MT project. Based on that information all of the Opus models that were chosen for this work were released between 2019 and 2020. They were implemented using the MarianMT framework [Junczys-Dowmunt et al., 2018], were based on the standard transformer architecture, and have been trained on the parallel OPUS corpus [Tiedemann, 2012; Tiedemann and Thottingal, 2020], and, to my understanding, none of the chosen models adopted back-translation as a data augmentation method. Although they are the oldest models in this selection, their performance is strong for high-resource language pairs, but exhibit weakness in low-resource settings [Tiedemann and de Gibert, 2023].

4.2.2 Many-to-Many Multilingual Model (M2M-100)

The Many-to-Many Multilingual Model (M2M-100) is the earliest MMT model in this work's selection, having been released in 2020. It covers 100 languages, thereby, enabling translations for all directions in the dataset of this work, providing a more controlled setting than by using the bilingual Opus models, and additionally, enabling zero-shot translations. The model architecture is based on the transformer architecture that combines parameter sharing for all directions with parameter sharing for a specific language group. To collect training data, a data mining strategy was applied, which exploits language similarity to avoid mining in all directions, focusing on non-English-centric directions. Additionally, they leverage back-translation to improve the translation performance for zero-shot and low-resource language pairs that score between 2 and 10 BLEU points. For the experiments in this work, the

418 million parameter version of the model was chosen. At the time of its release, it performed comparable to bilingual systems for high-resource language pairs and competitively for low-resource language pairs [Fan et al., 2021; Tiedemann and de Gibert, 2023].

4.2.3 Shallow Multilingual Machine Translation Model (SMaLL-100)

The Shallow Multilingual Machine Translation Model (SMaLL-100) is a distilled version of the M2M-100 12 billion parameter model, coming at a size of 200-600 million parameters, covering the same 100 languages. It was released in 2022. The same architecture as in M2M-100 had been employed, using a subset of its training data, which was uniformly sampled for all language pairs. Therefore, again a focus on improving performance for low-resource language pairs was set. Its translation performance exceeds the one of M2M-100-418M, being more comparable to its 1.2B relative while matching the size of the former [Mohammadshahi et al., 2022].

4.2.4 No Language Left Behind (NLLB)

The No Language Left Behind (NLLB) models are MMT models from 2022, which come in various versions, making them ideal candidates to explore the impact that certain differences, such as size or distillation of the NMT model can have. They cover 200 languages and also enable zero-shot translations [Akula et al., 2022]. Similarly to their predecessors, they utilize the same shared model capacity for multiple language pairs. While Akula et al. [2022] implement a Sparsely Gated Mixture of Experts (MoE) model architecture that activates only a subset of model parameters per input, compared to a dense transformer, which activates all model parameters per input, only the transformer models were chosen for this work because the MoE model with a size of 54.5 billion parameters is too large for the hardware that was used to conduct the experiments here [Akula et al., 2022]. As MMT models, each of the chosen NLLB models also covered all of the languages in the dataset, once again providing a more controlled setting than in a bilingual case. The following NLLB model versions were chosen:

NLLB-1.3B: NLLB-1.3B is a 1.3 billion parameter model. It is not only the largest model chosen from the NLLB-family, but also the largest model overall in this work. The model was trained on data obtained by diversified data mining techniques and included back-translated data obtained by NMT and SMT models on a large scale,

covering more than 300 directions [Akula et al., 2022]. As the largest and newest undistilled model used in this work, it exhibits the highest translation performance overall [Tiedemann and de Gibert, 2023].

NLLB-distilled-1.3B: This model version is a 1.3 billion parameter model that has been distilled from the 54.4B MoE model. It has been chosen as a model to compare to NLLB-1.3B [Akula et al., 2022]. In terms of translation, it performs comparable to its undistilled counterpart [Tiedemann and de Gibert, 2023].

NLLB-distilled-600M: The smallest model of the NLLB selection is a 600 million parameter model, which has also been distilled from the 54.4B MoE. It has been chosen as a direct comparison to the larger NLLB-distilled-1.3B, to explore the effect of model size in distilled models [Akula et al., 2022]. It has the lowest translation performance in the set of NLLB models, but still outperforms the older model families [Tiedemann and de Gibert, 2023].

4.3 Dataset

Similarly to the selection of NMT models, the test dataset for the experiments described in this work had to meet several conditions: First, the data had to be parallel and sentence-aligned. Second, the original translation direction had to be known. Third, the translations in the dataset should cover multiple translation strategies including HT and MT translations. These also had to be marked as such. Furthermore, a variety of language pairs – including low-resource language pairs – translation directions, and domains had to be covered to test the robustness of the approach.

In this context of identifying the translation direction, the WMT test datasets presented an exemplary fit. Not only does each sentence come with an aligned set of one source sentence and (at least) one professionally translated human reference translation, but also with corresponding translations from multiple MT systems, since WMT releases the general shared task (former news task) participating systems' outputs alongside the test data. The competitive nature of the WMT shared tasks ensures that the system output translations are representative of state-of-the-art machine translation techniques.

Four different years' worth of WMT test datasets were included in this work. By doing so the set includes not only translations generated by NMT but also by SMT and rule-based MT systems. According to that, the collected test set for the experiments was split into three categories based on the translation strategy: HT data

Direction	Num. Sents	Num. Docs	HT	NMT	Pre-NMT
cs→en	2957	185	4415	14621	10493
cs→uk	1930	1037	1930	23160	-
de→en	4032	568	6016	25943	13491
de→fr	1984	271	1984	9920	-
en→cs	2957	200	4415	14621	10493
en→de	4032	430	6016	25943	13491
en→ru	4733	356	4733	30613	8988
en→uk	3308	297	3308	30334	-
en→zh	3850	344	5724	50254	-
fr→de	1984	281	1984	9920	-
ru→en	4733	372	4733	30613	8988
uk→cs	1930	648	1930	23160	-
uk→en	3308	967	3308	30334	-
zh→en	3850	435	5724	50254	-
Total	45588	6391	56220	369690	65944

Table 1: Overview of the translations used for the high-resource test dataset.

(reference translations), NMT data (translations by NMT systems), and pre-NMT data (translations by systems that are statistical- or rule-based before the field's transition to NMT).

The language pairs for the main test set were chosen based on whether all the models in Section 4.2 support it, such that a direct comparison between the models' ability to identify the translation direction is possible. This test set encompasses a total of 14 directions, 7 languages[1], 7 language pairs, 11 domains, and 3 scripts[2]. An overview of the detailed statistics of the collected test set in terms of number of sentences, documents and translations for the main experiment is provided in Table 1. This dataset was balanced at the level of sentences per language pair, such that each translation direction is represented by as many examples as its inverse counterpart. In the following sections of the work, this test set will referred to as the *high-resource test set*. This label is somewhat of a simplification since the classification of resource availability into low-, mid-, and high-resource depends on the source. While Akhbardeh et al. [2021] label all pairs as high-resource, this is not the case for Mohammadshahi et al. [2022], who label cs, uk and zh as medium-

[1]Czech (cs), German (de), English (en), French (fr), Russian (ru), Ukranian (uk), Chinese (zh).
[2]Latin, Cyrillic, Chinese.

Direction	Num. Sents	Num. Docs	HT	NMT
bn→hi	503	145	503	4527
hi→bn	503	138	503	4527
xh→zu	503	145	503	2515
zu→xh	503	138	503	2515
en→ha	997	65	997	14955
ha→en	997	100	997	14955
Total	4006	732	4006	43994

Table 2: Overview of translations used for the low-resource test dataset.

resource languages.

Furthermore, an additional, smaller test set consisting of low-resource directions[3] was collected. It consists of 5 new languages[3], 3 language pairs, and 3 scripts[4]. The overview for the low-resource test set is given in Table 2. This dataset was also balanced within language pair. It was used for follow-up exploration with the best-performing systems for the high-resource language pairs.

In addition to the WMT data, a supplementary test set was curated, comprising excerpts from translations related to the plagiarism allegation case described in Chapter 1. This dataset introduces an element of real-world complexity, as it involves translations that might not adhere strictly to professional standards nor does it have to fit into either of the aforementioned categories of translation strategy. Such a dataset provides a unique opportunity to test the robustness and adaptability of the translation direction detection system in a scenario that extends beyond controlled environments, thereby presenting a more holistic evaluation of its capabilities.

The following subsections delve deeper into the characteristics of each subset and why it was included in the test dataset in this work. They will provide comprehensive details about the language pairs, domains, and resource settings encompassed within these subsets.

[3]This again depends on the source. In Mohammadshahi et al. [2022] zu, xh, and ha are classified as low, while the rest is classified as medium. In Akula et al. [2022] hi, xh, and zu are classified as high, while bn and ha are classified as low. en is not a low-resource language but in combination with ha it forms a low-resource language pair.

[3]Bengali (bn), Hindi (hi), Hausa (ha), Xhosa (xh), Zulu (zu)

[4]Latin, Devanagari, Bengali

4.3.1 WMT16: Pre-Neural and Early Neural System Outputs

In 2016, a significant shift in the machine translation landscape became evident as NMT systems began to replace SMT systems. This transition is notably illustrated in the findings of the WMT16 shared task, where NMT systems consistently outperformed pre-NMT strategies [Bojar et al., 2016]. Given this context, it is fitting to incorporate system outputs from the WMT16 shared task into this test set, capturing a blend of predominantly later pre-NMT outputs and early NMT outputs. Moreover, reference translations have also been included to form the HT category in this test set.

The parallel data is organized for sentence level, sorted by their associated document, and annotated in order to discern whether the source sentence is the original or the translation. Only source-original parallel sentences were selected for this work, the rest were discarded. It is worth noting that only sentences from the news task were used. Consequently, the data sourced from WMT16 originates exclusively from one domain, namely newspaper articles. The reference translations were produced by professional translators [Bojar et al., 2016].

4.3.2 WMT22: State of the Art System Outputs and Different Domains

At the beginning of this study, the WMT22 test dataset was the most recent WMT test data release. By 2022, all of the participating systems in the WMT shared task have transitioned to NMT, which is why that year's system outputs form a well-sized corpus of state-of-the-art machine translations. Hence, it was chosen to form – together with the WMT16 data – the basis for this study. Additionally, by 2022 the standard news task has evolved into the general machine translation task, such that it encompasses not only the news domain but also the domains social, e-commerce, and conversational that were collected as follows [Kocmi et al., 2022]:

- **news:** Content sourced from online news websites.
- **social** Comprises public Reddit discussions, maintaining individual posts as distinct documents; for languages with limited Reddit content, alternate sources like social media pages for Chinese and Zen blog platform for Russian were used.
- **e-commerce** Involves product descriptions provided by various companies.
- **conversational** For languages like English, German, French, and Chinese,

the data include agent-customer dialogues, with each message treated as a separate document and only using messages written in the original language. The resulting documents were oftentimes short.

The data were provided in a similar manner as for WMT16 with aligned sentences but with document-level annotations. However, for the WMT22 test set, there was no need for filtering samples that are source-original because they aimed to collect a test set that fulfills this condition to begin with, being aware of the effects of translationese in test data sets [Freitag et al., 2019; Läubli et al., 2020; Graham et al., 2020; Kocmi et al., 2022]. Translations by various professional translation agencies were used as reference translations. Only the language pairs zh↔en, de↔en, uk↔en, and cs↔en received translations from the same agency and were checked by a second translator. The language pairs zh↔en, cs↔en, and de↔en received a second reference in each direction from different translators. Furthermore, cs↔en has a third reference, because the first reference was deemed to be of low quality. Hence, an additional with grammar tools corrected version was added [Kocmi et al., 2022].

This means that there might be differences in quality between the different reference translations over all the language pairs. However, in the case of this work, this is a welcome characteristic, since it gives an indication about the robustness of the translation detection over different qualities of HT.

4.3.3 WMT23: More Recent State of the Art System Outputs

By the later stages of this study, the WMT23 test data set and system outputs were released. The reasons to include this were twofold. For one, the NLLB models (see: Subsection 4.2.4) were released after the release of the WMT22 test data, which opens the possibility that this data was part of the NLLB training set. Hence, if a model were to perform exceptionally well on an earlier subset, but not on this one, it would be a sign of the model overfitting on the earlier data. Secondly, a field like NLP research evolves at a rapid pace, an experiment over several months implies on the fly adaptation in order to contribute to the existing body of research. As for the time this is being written, the findings of WMT23 have not been released yet. Hence, little details on how the data were assembled can be provided here at the moment besides the metadata given in XML files that contained the translations. The XML files were structured in the same manner as for the previous years, giving the same information about the associated document and domain. The domain labels among the chosen directions were the following, while the explanations are

based on the interpretation of the name and data, and therefore, might not be completely accurate:

- **mastodon:** Texts from the micro-blogging platform Mastodon.
- **clipboard:** Details are unclear, although some of the text seems to be written dialogues.
- **speech:** Likely to be transcriptions of spoken language.
- **voice:** It is not clear how this domain differs from the *speech* domain.
- **games:** Most likely reviews for games.
- **manuals:** Manuals for a variety of things, such as games and technical hardware.
- **userreview, reviews, user_review** These were listed as three separate domains, however, due to their similarity, it is assumed that they represent the same domain. User reviews for online orders.

4.3.4 WMT21: Low-Resource Languages

A subset of the WMT21 data and their corresponding system output were added to broaden the test set's resource settings, in order to test the best-performing systems' robustness on low-resource languages. WMT21 was chosen because that year's news translation task added especially low-resourced languages into their general task – namely xh↔zu, hi↔bn, and ha. Although these language pairs were part of the general translation task, the source texts of the language pairs xh↔zu and hi↔bn were – in contrast to the other language pairs, which were extracted from online news sites – part of the FLORES-101 benchmark [Goyal et al., 2021] and were extracted from Wikipedia [Akhbardeh et al., 2021]. In Table 2 more in-depth statistics on the low-resource test dataset can be found.

4.3.5 Real-World Example: Plagiarism Allegation Incident

As discussed in Subsection 2.2.3, determining the translation direction without having to rely on a training set would be a valuable asset in forensic linguistics. Therefore, in addition to evaluating the methods presented in this work on translations generated in a controlled setting, evaluating them on real-world data provides valuable insights. Owing to the excerpts of the plagiarism allegation case being publicly

available[5], they were used in this analysis. The excerpts form a corpus of 86 parallel sentences in German and English. Current research confirms that the English version of the text is translated [Zenthöfer, 2022b; Ebbinghaus, 2022]. However, it is still open for discussion whether the text was manually or automatically translated, or automatically translated and manually edited[6]. For this specific language pair (de↔en), the optimal system found in this work was applied to determine its alignment with current research findings and to gather indications about the translation strategies used.

4.4 Evaluation Metrics

Since the translation direction task in this work is formulated as a binary classification and the classes in the dataset are balanced at sentence level, accuracy was used as the main evaluation metric, calculated as follows [Czakon, 2023]:

$$ACC = \frac{tp + tn}{tp + fp + tn + fn} \tag{4.5}$$

In the case of the experiments described in this work, true positives tp are the instances that were correctly classified in one direction of a language pair, true negatives tn are the instances that were correctly classified into the other direction of the same language pair, and the false negatives fn and false positives fp were the instances that were incorrectly classified for each respective direction.

Furthermore, in Subsection 5.2.1 bias towards one translation direction within a language pair is presented. The bias scores B are calculated as follows:

$$B = abs(50 - B_{prc}) \tag{4.6}$$

Since the datasets are balanced within a language pair, in the ideal case, B_{prc} is 50%, which would mean that the classification predicted each direction an equal number of times. To make this measure more intuitively interpretable, the percentages B_{prc} have been converted to the absolute differences to 50%. Hence, the higher the number in the table, the higher the bias, with 0 being the best possible and 50 being the worst possible score.

[5]Gathered, aligned, and kindly provided by Dr. Jannis Vamvas

[6]For more information on the characteristics of manually edited machine translations see Toral [2019]

4.5 Tools, Software and Libraries

During the execution of the experimental phase of this thesis, a range of tools were employed that are listed here for a holistic description of the experimental environment[7]. The HuggingFace Transformers library [Wolf et al., 2020] was pivotal and was utilized for all model implementations. However, it is worth noting that for the SMaLL-100 model using an additional script was necessary [Mohammadshahi et al., 2022]. PyTorch [Paszke et al., 2019] served as the primary framework for the usage of all models. In addition, the Opus models required the SentencePiece [Kudo and Richardson, 2018] and SacreMoses libraries, whereas SMaLL-100 also required SentencePiece.

[7] All of the data and code for data preparation, experiments, and result analysis can be found in the following repository:
`https://github.com/miwytt/unsupervised-translation-direction-detection`

5 Results

5.1 Unsupervised Translation Direction Detection

The first section of this chapter focuses on the results of the initial experiment, providing insight into the general translation direction detection capability of this approach using various models. The focus lies on presenting the results for different translation strategies but also for different resource settings. The chapter begins with a presentation of high-resource language pairs before moving on to the task of classifying low-resource language pairs.

5.1.1 High-Resource Language Pairs

	Opus	NLLB-dist-600M	NLLB-dist-1.3B	NLLB-1.3B	M2M-100-418M	SMaLL-100	Avg.
cs↔en	74.22	67.12	67.08	66.78	74.38	72.24	70.30
de↔fr	60.77	67.40	67.77	67.98	69.51	72.93	67.73
en↔ru	69.42	68.30	64.97	66.23	72.11	73.34	69.06
de↔en	**77.04**	**73.85**	71.82	**72.55**	74.95	74.93	**74.19**
en↔uk	67.93	65.22	66.05	65.11	**75.39**	75.53	69.20
cs↔uk	73.61	73.37	**72.06**	72.20	74.53	76.17	73.66
en↔zh	65.58	56.10	55.19	55.15	73.79	**76.84**	63.78
Macro-Avg.	69.80	67.34	66.42	66.57	73.52	74.57	69.70

Table 3: Accuracy (%) per language pair for NMT at sentence level. The scores marked in bold are the best scores for each model, whereas the underlined scores are the best scores for each language pair.

The results of the main experiment are listed in Table 3 for NMT, Table 4 for HT, and Table 5 for pre-NMT. They show the accuracy for all language pairs in the dataset for each of the tested models. The highest results overall can be observed for NMT, with all of the results reaching above-chance accuracy – showing that translation direction can indeed be detected using the approach described in this work. However, there are some substantial differences that need to be addressed.

One observation that can be made from Table 3 is that the different NMT models exhibit differences in performance for NMT data. In this case, a difference of 8.15% can be observed between the average accuracies of the best-performing model, SMaLL-100, and the worst-performing model, NLLB-dist-1.3B. SMaLL-100, as the best-performing model in this setting, outperforms the other models in all but two language pairs, cs↔en and de↔en, for which the Opus model pairs and M2M-100 perform better – Opus reaching the single best result overall with 77.04%.

More pronounced differences can be seen when the results between the language pairs are compared. In this case, the difference between averages reaches 10.41% (best: de↔en and worst: en↔zh) and ranges between 17.75% (NLLB-dist-600M; again best: de↔en and worst: en↔zh) and 4.6% (SMaLL-100; best: en↔zh and worst: cs↔en) for the individual models' results.

Moving on to Table 4 for HT: First, one can observe an overall drop in performance compared to the NMT data. Nonetheless, the results still show a performance above chance in most cases, with average accuracy scores ranging between 57.89% (NLLB-dist-1.3B) and 66.33% (SMaLL-100), displaying a difference of 8.44%. The best scores were obtained for the language pairs de↔fr reaching 73.05% (best overall result for HT) with NLLB-dist-1.3B.

While the performance gap between the models is less prominent here than for the NMT data, the performance gap between the different language pairs grows with the largest difference between average performances of language pairs being 13.37% (best: de↔fr and worst: en↔ru). The differences between the language pairs within the results from one model range from 9.57% (Opus; best: cs↔uk and worst: de↔en) to 23.38% (NLLB-1.3B; best: de↔fr and worst: en↔zh).

	Opus	NLLB-dist-600M	NLLB-dist-1.3B	NLLB-1.3B	M2M-100-418M	SMaLL-100	Avg.
cs↔en	62.14	57.34	57.74	57.52	64.46	64.05	60.54
de↔fr	62.53	**70.56**	**73.04**	**72.99**	68.65	**71.42**	**69.87**
en↔ru	59.14	54.53	52.34	53.50	59.22	60.24	56.50
de↔en	57.01	55.29	53.37	54.07	62.17	61.83	57.29
en↔uk	56.54	54.56	54.62	55.17	**71.72**	71.33	60.66
cs↔uk	**66.58**	64.33	64.40	65.05	66.40	67.41	65.70
en↔zh	57.27	50.25	49.74	49.61	66.48	68.05	56.90
Macro-Avg.	60.17	58.12	57.89	58.27	65.59	66.33	61.06

Table 4: Accuracy (%) per language pair for HT at sentence level.

	Opus	NLLB-dist-600M	NLLB-dist-1.3B	NLLB-1.3B	M2M-100-418M	SMaLL-100	Avg.
cs↔en	**37.05**	**30.73**	**25.49**	**26.62**	38.72	38.98	**32.93**
en↔ru	24.94	17.81	13.68	14.80	32.24	32.95	22.74
de↔en	33.32	26.71	21.07	22.58	**41.85**	**40.99**	31.09
Macro-Avg.	31.77	25.08	20.08	21.33	37.60	37.64	28.92

Table 5: Accuracy (%) per language pair for pre-NMT at sentence level.

Here, SMaLL-100 reaches top accuracy only for three language pairs, tying with M2M-100. Nonetheless, SMaLL-100 outperforms M2M-100 in terms of average accuracy by reaching 66.33%. Although the result is not as strong for HT as it is for NMT, the best-performing model for HT is again SMaLL-100.

Finally, the results for the pre-NMT outputs show a substantial drop in performance for all models and language pairs, reaching well below-chance results in all cases. The overall best result is reached for de↔en with 41.85% by using M2M-100, while the lowest scores are reached by NLLB-dist-1.3B for en↔ru. On average, the language pair that was most discernible in terms of translation direction is cs↔en. And, SMaLL-100 continues to be the best-performing model as in previous datasets.

In summary, the results show clear differences in the classification power of NMT models between the three categories. While the results for automatically generated translations by NMT systems lie well above chance, and the results for manually translated sentences seem discernible with the approach as well, the last table of results indicates that the method fails for translations that were generated using pre-NMT systems.

5.1.2 Low-Resource Language Pairs

Given the performance in the main experiment, the SMaLL-100 and M2M-100 were chosen to explore their performance on low-resource languages. SMaLL-100 was the straightforward choice for this follow-up experiment as the best-performing system in the main experiment. However, due to their similar performance on HT (see: Table 4) and their relatedness, both were tested in a low-resource setting. Table 6 shows the results for NMT, while Table 7 depicts the results for HT.

For NMT in this resource setting, M2M-100 outperforms its distilled counterpart

	M2M-100-418M	SMaLL-100	Avg.
en↔ha	50.82	49.98	50.40
bn↔hi	**66.11**	53.02	59.57
xh↔zu	62.72	**57.02**	**59.87**
Macro-Avg.	59.88	53.34	56.61

Table 6: Accuracy (%) per language pair for NMT in a low-resource setting.

noticeably by reaching an average of 59.88% and performing with an accuracy of over 60% for bn↔hi and xh↔zu, where SMaLL-100 only reaches little above chance accuracy. For en↔ha no detection power is observable on this dataset using either model.

For HT results (Table 7), SMaLL-100 regains its place as the best-performing system, however, only with a small margin to M2M-100 (0.6%). Similarly to the results in Subsetion 5.1.1 on high-resource HT, the translation direction of the low-resource HT were more difficult to detect. The task on low-resource language pairs brings the accuracies down to 50%, indicating that the systems fail to detect the translation direction in this setting.

	M2M-100-418M	SMaLL-100	Avg.
en↔ha	49.90	49.90	49.90
bn↔hi	49.30	50.70	50.00
xh↔zu	**50.99**	**51.39**	**51.19**
Macro-Avg.	50.06	50.66	50.36

Table 7: Accuracy (%) per language pair for HT in a low-resource setting.

To summarize, the results for low-resource language pairs are lower than in a high-resource setting. Nonetheless, accuracy scores can be reached that are above chance by using M2M-100 for NMT data. The results for HT data stay at around 50%.

5.2 Bias and Normalization

The observation that the translation-probability-based translation direction detection approach results differ between language pairs, translation strategy, and resource level poses the question if there are difference between translation directions as well. The exploration in this section attempts to quantify the potential discrepancy with a bias score. The goal is to determine whether a bias towards a translation direction exists. Furthermore, the results of the attempt to mitigate potential biases for high-resource languages and improve overall results using the normalization strategy, which has been described in Subsection 4.1.2, are shown.

5.2.1 Bias before Normalization

To measure the potential bias towards one translation direction, bias has been quantified as described in Subsection 4.4. Tables 8, 9 and 10 depict those bias scores.

Table 8 shows the bias scores for the NMT dataset. One can observe that on average SMaLL-100 has the lowest signs of bias, achieving scores close to 0 (en↔uk), which indicates almost no bias towards a certain translation direction of the pair. The NLLB models, on the other hand, exhibit the highest signs – NLLB-dist-600M scoring the worst with a score of above 40 for en↔zh. Notable is also the high range of scores that the NLLB models exhibit for the different language pairs, ranging from almost no bias to nearly completely biased.

As for the language pairs, it seems that en↔ru, with an average score 4.01, is least susceptible to the bias, whereas en↔zh with 29.73 is most vulnerable.

	Opus	NLLB-dist-600M	NLLB-dist-1.3B	NLLB-1.3B	M2M100-418M	SMaLL-100	Avg.
en↔zh	23.07	40.97	42.03	41.51	20.12	10.69	29.73
en↔uk	12.48	28.12	23.45	26.89	3.50	2.19	16.10
cs↔uk	**1.96**	10.04	7.42	8.70	8.37	**0.09**	6.10
en↔ru	9.65	**0.67**	**5.82**	**1.50**	4.93	1.47	**4.01**
de↔en	3.44	10.34	6.14	8.54	8.40	5.69	7.09
cs↔en	6.17	23.31	18.45	21.38	**0.33**	6.10	12.62
de↔fr	29.21	18.69	14.60	15.22	21.39	13.61	18.79
Macro Avg.	12.28	18.88	16.84	17.68	9.58	**5.69**	13.49

Table 8: Bias for NMT data before normalization.

	Opus	NLLB-dist-600M	NLLB-dist-1.3B	NLLB-1.3B	M2M100-418M	SMaLL-100	Avg.
en↔zh	21.11	41.35	41.67	41.14	22.08	12.46	29.97
en↔uk	7.45	25.98	18.48	22.24	**0.17**	**1.56**	12.65
cs↔uk	7.93	18.01	16.42	16.81	12.56	2.75	12.41
en↔ru	9.66	**0.59**	**4.35**	**1.45**	12.15	7.35	5.93
de↔en	**0.21**	10.46	4.53	6.98	5.09	2.50	**4.96**
cs↔en	4.53	22.92	16.28	19.01	6.34	1.95	11.84
de↔fr	30.87	19.57	15.72	16.03	21.37	14.01	19.60
Macro Avg.	11.68	19.84	16.78	17.67	11.39	6.08	13.91

Table 9: Bias for HT data before normalization.

The scores for HT in Table 9 display a very similar pattern to the scores for the NMT dataset, reaching an overall bias score of almost 14%. The models with the highest and lowest scores align with the above-described results as well. The language pair with the on average lowest bias score is in this case de↔en, while en↔zh continues to reach the highest bias score.

Finally, the bias scores for the pre-NMT dataset continue to align with the previous two sets of results in terms of the model with SMaLL-100 reaching the lowest bias scores, while NLLB-dist-600M reaches the highest bias score – thus being the most vulnerable to bias overall. The average scores at the level of the language pair are slightly lower.

In summary, this subsection has shown that the models are susceptible to bias to a translation direction within a language pair. The models differ substantially in terms of bias exhibition and do so consistently for all translation strategies.

	Opus	NLLB-dist-600M	NLLB-dist-1.3B	NLLB-1.3B	M2M100-418M	SMaLL-100	Avg.
en↔ru	**2.29**	**3.34**	**1.55**	**3.32**	2.55	**0.37**	**2.24**
de↔en	6.77	12.61	8.94	11.02	11.40	7.52	9.71
cs↔en	6.73	19.95	14.83	16.88	**0.94**	4.69	10.67
Macro Avg.	5.26	11.97	8.44	10.41	4.96	4.19	7.54

Table 10: Bias for pre-NMT data before normalization.

5.2.2 Bias and Accuracy after Normalization

In order to explore the effects of probability normalization on the bias, a subset of the dataset and models were chosen for a follow-up experiment. Due to the normalization technique utilizing zero-shot translation probabilities, only MMT-model-based probabilities were possible to be normalized. Normalization was explored for the worst model in terms of bias scores (NLLB-dist-600M), as were the four language pairs most affected by the bias of the model since those results had the largest potential for correction. Additionally, the results for the same language pairs by M2M-100 and SMaLL-100 were chosen as comparative data, such that the normalization effect is not only further explored on different model sizes and architectures, but also on results that were indicating different levels of bias. The original bias score as well as the original translation direction detection accuracy (left) are presented in comparison to the normalized scores (right) in Table 11, 12 and 13.

Table 11 summarizes the results for the NMT dataset. For NLLB-dist-600M an improvement can be seen in three out of the four tested language pairs. The bias score for en↔zh indicates the largest improvement, showing almost complete mitigation of the bias as well as an improvement for the classification accuracy. Similar effects can be observed for the M2M-100 results for en↔zh, whereas the SMaLL-100 results show an improvement of the bias scores for the language pairs cs↔en and de↔fr. The normalized SMaLL-100 results exhibit the least amount of decreased bias.

	NLLB-dist-600M	M2M-100-418M	SMaLL-100
en↔zh	40.97 | **2.40**	20.12 | **0.08**	10.69 | 18.99
– Accuracy	56.10 | **63.89**	73.79 | **74.94**	76.84 | 71.51
en↔uk	28.12 | **16.04**	3.50 | 6.24	2.19 | 9.20
– Accuracy	65.22 | 62.45	75.39 | 73.87	75.53 | 71.02
cs↔en	23.31 | 30.72	0.33 | 3.81	6.10 | **2.46**
– Accuracy	67.12 | 57.74	74.38 | 70.01	72.24 | 70.04
de↔fr	18.69 | **7.49**	21.39 | **19.69**	13.61 | **11.51**
– Accuracy	67.40 | 59.90	69.51 | 63.99	72.93 | 69.18

Table 11: Bias and accuracy (%) for NMT subset of dataset (before | after) normalization.

	NLLB-dist-600M	M2M-100-418M	SMaLL-100
en↔zh	41.35| **8.47**	22.08| **5.54**	12.46 | 20.51
– Accuracy	50.25 | **53.09**	66.48 | 66.28	68.05 | 66.12
en↔uk	25.98 | **25.70**	0.17| **0.05**	1.56 | 12.62
– Accuracy	54.56 | 52.84	71.72 | 71.27	71.33 | 67.42
cs↔en	22.92 | 27.63	6.34 | **4.56**	1.95 | 2.37
– Accuracy	57.34 | 54.44	64.46 | 62.86	64.05 | **64.66**
de↔fr	19.57 | **9.90**	21.37| **19.62**	14.01 | **11.54**
– Accuracy	70.56 | 61.90	68.65 | 63.27	71.42 | 67.44

Table 12: Bias and accuracy for HT subset of dataset (before | after) normalization.

A further point to note in this setting is that although normalization seems to have a positive effect in terms of bias mitigation, it rarely improves the translation direction detection accuracy. Most of the results show a slightly decreased accuracy. The only exceptions are the results for en↔zh by NLLB-dist-600M and M2M-100, which show an improvement of 7.79% and 1.15%. The SMaLL-100 accuracy scores decrease for all language pairs.

The results for the HT dataset in Table 12 paint a similar picture, especially for NLLB-dist-600M. There are, however, multiple notable differences. Firstly, bias scores for M2M-100 are lowered for all four language pairs, even the ones that were almost unbiased during the initial experiment (en↔uk), indicating a positive effect in terms of bias mitigation for HT results from this model. The SMaLL-100 results show once again improvement for de↔fr in terms of bias. Although the bias score for cs↔en has not been lowered as above, a small rise in accuracy can be noted. However, the overall accuracy scores do not seem to be positively affected by the normalization with small losses in performance by all models for almost all the language pairs.

Finally, the subset of tested pre-NMT data is presented in Table 13. In this setting, a positive effect on all results has been shown in terms of translation direction detection accuracy and for SMaLL-100, the bias seems to have been reduced. However, the accuracy stays in all cases far below 50%.

In summary, the comparison of the initial and normalized results has shown that while the normalized results indicate a lowered bias for previously heavily biased

	NLLB-dist-600M	M2M-100-418M	SMaLL-100
cs↔en	19.95 | 25.0	0.94 | 1.18	4.69 | **2.42**
– Accuracy	30.73 | **36.52**	38.72 | **41.09**	38.98 | **40.4**

Table 13: Bias and accuracy on pre-NMT subset of dataset (before | after) normalization.

cases, the classification accuracy experiences losses for all translation strategies but pre-NMT. Not only has the normalization different effects on translations with different strategies but also different on different language pairs depending on the NMT model that was used to generate the probabilities. Since the results prove inconclusive in most cases, the following results are based on unnormalized translation probabilities.

5.3 Results at Different Text Lengths

This section presents the results across varying text lengths, employing two approaches to scrutinize the sentence-level data. Initially, by majority vote aggregated sentence-level outcomes are inspected to understand how translation direction detection fares at document level. Secondly, the influence of text length at sentence level is examined, providing a nuanced view of the performance across different text sizes.

5.3.1 From Sentences to Documents

In this subsection, the results for translation direction detection at document level are presented. The aggregation from sentence level to document level was performed using a majority vote as explained in 4.1.3. For the high-resource language pairs, the follow-up experiment was only performed on the results of the best-performing system, SMaLL-100. For the low-resource language pairs both SMaLL-100 as well as M2M-100 were considered, due to M2M-100's superior performance on the low-resource NMT dataset. The document-level results for all three data categories are presented in Table 14 for high-resource language pairs and Table 15 and 16 for low-resource language pairs.

	NMT	Human	Pre-NMT	Macro-Avg.
de↔fr	78.05	78.66	-	78.36
de↔en	79.10	68.45	**19.45**	55.67
cs↔uk	80.05	71.57	-	75.81
en↔zh	83.44	70.75	-	77.10
cs↔en	84.20	76.23	16.53	58.99
en↔ru	85.60	67.88	10.03	54.50
en↔uk	**86.04**	**85.64**	-	**85.84**
Macro-Avg.	82.35	74.17	15.34	54.37

Table 14: Accuracy scores for HT, Pre-NMT, and NMT datasets at document level using SMaLL-100 at document level.

Table 14 demonstrates a noticeable improvement at the document level compared to the results at the sentence level. While SMaLL-100 results at sentence level for NMT data show a range between 72% and 76% with an average of 74.57%, at document level a boost of almost 8% can be observed, with accuracy ranging from little below 80% to 86%. An equivalent effect is seen for the results for the HT dataset: The sentence-level average of 66% is raised to 74%. However, in contrast to the positive enhancement of the results for NMT and HT data, the accuracy scores for pre-NMT have experienced a decrease to an average accuracy of 15%.

The document-level results for low-resource language pairs with SMaLL-100 did not seem to benefit from the majority vote strategy. Table 15 illustrates the results, showing that there is a drop in performance for both the NMT and HT dataset.

	NMT	HT	Macro-Avg.
bn↔hi	46.33	44.35	45.34
en↔ha	50.00	**50.00**	**50.00**
xh↔zu	**51.62**	44.78	48.20
Macro-Avg.	49.32	46.36	47.85

Table 15: Low resource accuracy scores for HT, Pre-NMT, and NMT datasets using SMaLL-100 at document level.

	NMT	HT	Macro-Avg.
bn↔hi	**62.12**	41.14	**51.63**
en↔ha	50.00	**50.00**	50.00
xh↔zu	58.81	43.71	51.26
Macro-Avg.	57.31	44.95	51.13

Table 16: Low-resource accuracy scores for HT, Pre-NMT, and NMT datasets using M2M-100 at document level.

The results produced with M2M-100 in Table 15, however, paint a slightly better picture. While the results for HT score lower than above, the results for NMT reach an accuracy score of over 60% for the language pair bn↔hi and 58% for xh↔zu. The results for en↔ha score the same accuracy as in this setting as they do with SMaLL-100.

5.3.2 From Characters to Sentences

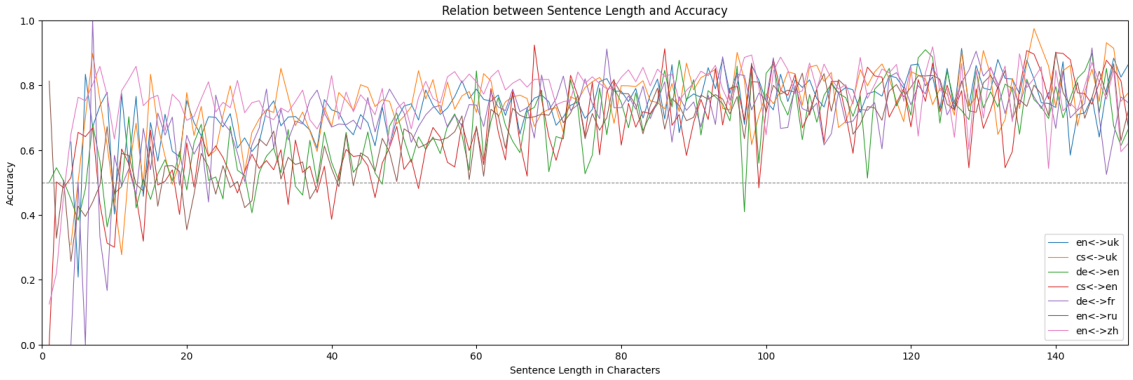

Figure 2: Mean accuracy per language pair over sentence length (in characters). The dashed line marks 50% accuracy.

Figure 2 illustrates how the sentence length (in characters) affects the mean accuracy per language pair for the SMaLL-100 results. For visibility purposes this figure has been cut at 150 characters – a larger range is shown in Appendix A. The figure shows a steep increase in accuracy within the first 20 characters. For longer sentences, the trend is still increasing with all language pairs passing the 50% accuracy mark at a sentence length of 60 characters. With approximately 80 characters a plateau is reached with an average accuracy of 70%. This plateau stays consistent for longer

sentences apart from occasional outliers.

5.4 Results for Different WMT Datasets

In this section of the report, the data was segregated based on the WMT year. This division allows to provide a more granular and specific examination of the results. By analyzing each dataset separately, corresponding to its respective year, trends, patterns, and anomalies that are specific to each time period can be recognized.

Table 17 depicts the accuracy scores for NMT data. The upper row shows the results when the different NLLB models are used. Each of these tables shows that the direction detection was more successful for the newer datasets, which correspond to the system outputs of newer NMT models. All NLLB models achieve scores around 50% for WMT16 translations, scoring as low as 36% for en↔ru with NLLB-1.3B and as high as 57.65%. The results for WMT22 system outputs show an increase in performance compared to the older data. For this subset accuracy scores range from 52.56% to 72.83% with the majority of language pairs reaching well above 60%. Another small increase in performance can be seen for the WMT23 dataset. There, a range between 56.97% and 79.97% can be observed. In general, the NLLB models show a trend of performing best for the latest system outputs.

NLLB-dist-600M

	WMT16	WMT22	WMT23
de↔en	50.87	**73.51**	**79.97**
en↔uk	-	64.87	65.61
en↔zh	-	53.75	57.74
cs↔en	**57.65**	68.20	-
de↔fr	-	67.40	-
en↔ru	46.90	66.67	71.79
uk↔cs	-	73.37	-

NLLB-dist-1.3B

	WMT16	WMT22	WMT23
de↔en	40.36	**72.27**	**77.79**
en↔uk	-	65.52	66.63
en↔zh	-	52.64	56.97
cs↔en	**51.63**	68.83	-
de↔fr	-	67.77	-
en↔ru	36.38	62.17	70.18
uk↔cs	-	72.06	-

NLLB-1.3B

	WMT16	WMT22	WMT23
de↔en	42.73	**72.83**	**78.59**
en↔uk	-	64.52	65.74
en↔zh	-	52.56	56.97
cs↔en	**52.86**	68.36	-
de↔fr	-	67.98	-
en↔ru	39.32	64.51	70.32
uk↔cs	-	72.20	-

Opus-MT

	WMT16	WMT22	WMT23
de↔en	60.91	**77.19**	**80.32**
en↔uk	-	70.18	65.36
en↔zh	-	66.28	65.10
cs↔en	**73.02**	74.36	-
de↔fr	-	60.77	-
en↔ru	56.48	66.59	73.16
uk↔cs	-	73.61	-

M2M-100

	WMT16	WMT22	WMT23
de↔en	63.98	74.03	**79.93**
en↔uk	-	**77.53**	72.95
en↔zh	-	74.28	73.45
cs↔en	**69.48**	74.94	-
de↔fr	-	69.51	-
en↔ru	64.79	70.15	74.55
uk↔cs	-	74.53	-

SMaLL-100

	WMT16	WMT22	WMT23
de↔en	64.04	74.06	**79.74**
en↔uk	-	76.95	73.90
en↔zh	-	**77.62**	76.29
cs↔en	**68.08**	72.72	-
de↔fr	-	72.93	-
en↔ru	64.39	72.34	75.08
uk↔cs	-	76.17	-

Table 17: Accuracy (%) for NMT sorted by language pair and WMT year for each model.

The second row, which displays the result for Opus, M2M-100, and SMaLL-100, shows a similar trend, but less evidently so. The lowest scores are still observed for the WMT16 dataset, ranging from 56.48% to 73.02% (both by using Opus). Again, an increase in performance on the newer dataset, WMT22, is observable, but the increase is less substantial than above. Opus ranges from 60.77% to 77.19%, while M2M-100 ranges from 69.51% to 77.53%, and SMaLL-100 for 72.34% to 77.62%. And although the results for the latest system outputs range from 65.10% to 80.32% (Opus), from 72.95% to 79.93% (M2M-100), and 73.90% to 79.74% (SMaLL-100), in contrast to the NLLB models, the results for WMT23 do not surpass the results for WMT22 in all cases.

In Table 18 the results for each NMT model on the HT data are presented. The first row again shows the results when using the probabilities generated by the NLLB models. The differences between the results for each WMT subset are less prominent than for NMT. While results for the earliest translations range from 49.30% to 53.90%, the WMT22 results range from 49.63% to 73.04%, and WMT23 from 49.59% to 71.86%. Hence, the results for HT by NLLB have a higher overlap performance overlap between each dataset than for NMT data. The largest range and highest scores are achieved for the WMT22 dataset. However, the WMT22 covers more language pairs than the other two datasets, providing more possibilities to perform better. For those language pairs that are covered with both WMT22 and WMT23, the NLLB scores are higher for the former for en↔uk in all cases and for en↔zh in the larger two models. The models achieve better results for de↔en and en↔ru in all cases for the WMT23 dataset.

The second row shows a similar pattern, although with generally higher scores. The scores for WMT16 data show a range from 56.44% to 59.23% (Opus), from 56.17% to 58.77% (M2M-100), and from 56.48% to 60.32% (SMaLL-100). For WMT22 translations, the results start at 53.94% and reach up to 66.58% (Opus-MT), 61.84% up to 75.42% (M2M-100), and 60.79 up to 75.57% (SMaLL-100). The results show a similar but slightly lower range for the latest translations ranging from 49.84% to 72.95% (Opus), from 59.08% to 73.77% (M2M-100), and from 59.81% to 73.32% (SMaLL-100). Similarly to the results in the upper row, the best scores are found for the WMT22 and WMT23. And when comparing those two years in terms of language pairs, en↔uk and en↔zh show higher scores for WMT22, while de↔en and en↔ru do so for WMT23.

NLLB-dist-600M

	WMT16	WMT22	WMT23
de↔en	**53.90**	53.51	**71.49**
en↔uk	-	57.28	50.51
en↔zh	-	50.12	50.48
cs↔en	51.90	60.12	-
de↔fr	-	**70.56**	-
en↔ru	50.57	55.11	57.46
uk↔cs	-	64.33	-

NLLB-dist-1.3B

	WMT16	WMT22	WMT23
de↔en	50.13	52.11	**70.95**
en↔uk	-	57.57	50.23
en↔zh	-	49.80	49.62
cs↔en	**51.29**	61.04	-
de↔fr	-	**73.04**	-
en↔ru	47.46	52.42	56.50
uk↔cs	-	64.40	-

NLLB-1.3B

	WMT16	WMT22	WMT23
de↔en	50.00	53.12	**71.86**
en↔uk	-	57.99	50.97
en↔zh	-	49.63	49.59
cs↔en	**51.26**	60.73	-
de↔fr	-	**72.99**	-
en↔ru	49.30	52.99	57.60
uk↔cs	-	65.05	-

Opus-MT

	WMT16	WMT22	WMT23
de↔en	**59.27**	53.94	**72.95**
en↔uk	-	60.83	49.84
en↔zh	-	58.21	55.47
cs↔en	57.97	64.30	-
de↔fr	-	62.53	-
en↔ru	56.44	59.99	60.74
uk↔cs	-	**66.58**	-

M2M-100

	WMT16	WMT22	WMT23
de↔en	**58.77**	61.84	**73.77**
en↔uk	-	**75.42**	65.93
en↔zh	-	67.67	64.22
cs↔en	57.91	67.83	-
de↔fr	-	68.65	-
en↔ru	56.17	62.40	59.08
uk↔cs	-	66.40	-

SMaLL-100

	WMT16	WMT22	WMT23
de↔en	**60.37**	60.79	**73.32**
en↔uk	-	**75.57**	64.69
en↔zh	-	69.30	65.66
en↔cs	58.94	66.68	-
de↔fr	-	71.42	-
en↔ru	56.48	64.45	59.81
cs↔uk	-	67.41	-

Table 18: Accuracy (%) for HT sorted by language pair and WMT year for each model.

5.5 Real-World Example

This chapter has provided insights into how unsupervised translation direction detection as described in this work has performed given different models and a varied test dataset. The results suggest that under most circumstances the translation direction can be detected. To illustrate this further, the approach is tested on real-world data from the plagiarism allegation case described in Chapter 1. Since this data set consists of parallel sentences in German and English, the approach has been tested once with the overall best-performing NMT model, SMaLL-100, and, additionally, with the best-performing model for this language pair for NMT, Opus. The results are presented in Table 19. They show that with translation probabilities by both models being higher for de→en than for en→de in well over 65% of the data, the results align with the current stage of the investigation in the plagiarism allegation case: The text seems to have been translated from German to English, indicating against the alleged plagiarism.

	Opus	SMaLL-100
de→en	66.27	**69.77**

Table 19: Results for data from the plagiarism allegation case.

5.6 Translation Probabilities

This section is intended to illustrate the distribution of the translation probabilities based on the three translation strategy categories: HT, NMT, and pre-NMT. The left side of Figure 5.6 shows the translation probability distribution when the translation model was confronted with the original direction, while the right side of the figure depicts the probability distribution of the inverse scenario.

Looking at the left figure one can observe that when confronted with HT for the original direction, the model generates on average the lowest probabilities, while pre-NMT probabilities are on average slightly higher, and NMT-based probabilities are higher than both other categories while covering also a broader spectrum of probabilities.

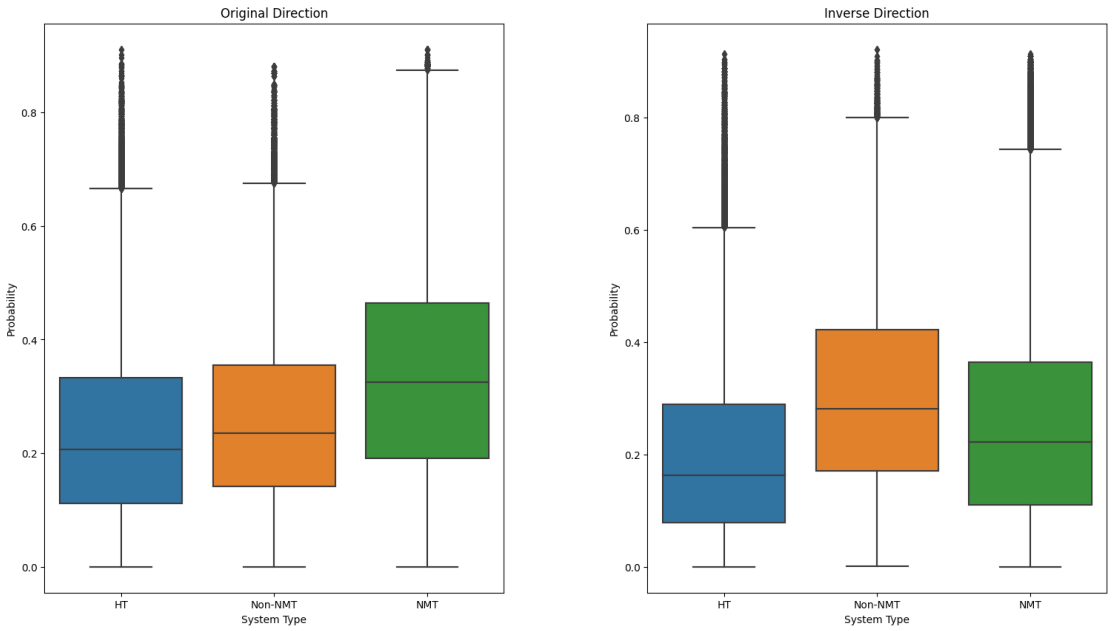

Figure 3: Comparison of translation probabilities in both translation directions generated by SMaLL-100.

The right figure, on the other hand, shows that this pattern does not stay consistent for the inverse direction. In this scenario, the model generated the highest probabilities for the pre-NMT dataset followed by NMT and, finally, by HT at equal intervals.

In summary, this section highlights the notable differences in translation probabilities among the three translation strategies - HT, NMT, and pre-NMT. It underscores the discrepancies that emerge from the previously reported results, emphasizing the varying performance for these strategies in the translation direction detection task.

6 Discussion

6.1 Result Interpretation

The primary objective of this thesis was to explore the capability of identifying the translation direction in parallel sentence pairs in an unsupervised manner. This was pursued by comparing the conditional translation probabilities generated by NMT models when confronted with the same sentence pair in both possible translation directions. The assumption was that the NMT model would produce a higher probability for the original direction.

In Section 1.2 the following research questions were formulated:

1. Can translation probabilities be used for translation direction detection?
2. Can potential biases be mitigated by normalizing the translation probabilities?
3. What other properties do translation probabilities display in terms of translation direction?

The results at sentence level and particularly at document level indicate a positive response to the first research question: Translation probabilities can be used to detect translation direction. For a high- to mid-resource setting on the language pairs tested in this work SMaLL-100 is an appropriate choice without normalization by zero-shot translation probabilities. The approach works best for sentence pairs where the translations were generated by NMT models. It can also be applied to manual translations, but the accuracy decreases. The approach seems to be least effective when faced with translations that were generated by pre-NMT systems. In a low-resource setting, the accuracies are lower, but they can still be regarded as an indication for a specific translation direction – for the low-resource language pairs that were tested here, M2M-100 seems to be the most suitable choice. This short summary shows that the approach leads to a range of different results depending on several factors, such as the NMT model, which generates the probabilities, the language pair, and the translation strategy. How these factors influence the results is analyzed in more detail in the Subsections 6.1.1 and 6.1.2.

Secondly, the next research question concerns biases and their mitigation. The results suggest that there is a detectable bias when trying to detect the translation direction. This bias varies similarly to the detection accuracy depending on the model and language pair. A mitigation attempt by normalizing the probabilities before making the binary decision has proven to be semi-successful. Subsection 6.1.3 provides a closer look into this topic.

Furthermore, the variety of results advocates for a closer analysis of the translation probabilities that were produced during these experiments. Subsection 6.1.4 formulates an attempt to answer the third research question by showing that the probabilities can be grouped into distinctive categories depending on the translation strategies. This observation can be used to explain the results obtained in this work to some extent by linking them to the linguistic observations of previous research. Moreover, it can be used as a foundation for future work.

Finally, the results are compared to previous work in Subsection 6.1.5 to contextualize the performance of the approach described in this work, before discussing the limitations in Section 6.2 and, finally, outlining the future work in Section 6.3.

6.1.1 Model Related Influences

The experiments cover several NMT models to produce translation probabilities. The models differ in terms of size, training data, and training strategy. On average, SmaLL-100 is the best-performing model for the high-resource dataset for all translation strategies, as well as for the HT subset of low-resource data. M2M-100 proves to closely follow the SMaLL-100 results and even outperformed it on the NMT dataset for low-resource language pairs. The bilingual Opus models scored almost comparably to M2M and SMaLL-100 – even outperforming them for one language pair in the NMT dataset, while the NLLB models take the last place.

This outcome suggests the following: A larger model size does not necessarily have a positive influence on the translation direction detection capabilities, nor do more recent models necessarily outperform older ones. Furthermore, bilingual models are not generally better or worse than multilingual ones, but they require more extensive preparation beforehand, as one must collect and align comparable models for each specific translation direction, a process that can be more time-consuming and restrictive.

A last model-related point is the effect of model distillation. Three distilled models were considered in the experiments: NLLB-dist-600M, NLLB-dist-1.3B, and

SMaLL-100. Two out of three models outperformed their undistilled counterparts. Additionally, the smaller NLLB-dist-600M outperformed its larger relative NLLB-dist-1.3B. This could be connected to the in Subsection 2.2.2 mentioned characteristic that translations by distilled NMT models display more interference. Hence, they might assign higher probabilities where interference is present on the target side. In order to ascertain this hypothesis, further investigation of the translation probabilities paired with a qualitative analysis of the translations would be required.

6.1.2 Robustness for Diverse Datasets

This subsection evaluates the robustness of the translation direction detection method across diverse datasets that are characterized by various translation strategies, languages, resource settings, text lengths, and domains.

Translation Strategy The test dataset is categorized by translation strategy into HT, NMT outputs, and pre-NMT outputs. This categorization was aimed at capturing the effects of the strategies on translation direction detection. The results consistently showed the highest accuracy for NMT outputs, suggesting that the NMT systems are more adept at identifying translations similar to their own output. HT data, while slightly more challenging, still produced accuracies above 50% in most cases. However, for pre-NMT data, the accuracy drops significantly below chance levels, indicating the model's inability to recognize translation patterns that are vastly different from its training data. This outcome aligns with the descriptions of translationese types and suggests that non-neural translations deviate considerably from the NMT model's "expectations" of translation style.

There is an overall tendency across the experiments for results for pre-NMT to yield the opposite of the other two translation strategies. Considering this consistency, one could argue in favor of the opposite hypothesis for pre-NMT data, namely, that the inverse translation direction yields higher translation probabilities by an NMT model. In this case, the results would reach accuracy up to approximately 80%.

Languages and Language Pairs: The detection's effectiveness varies across different language pairs and models. This variation in performance points to an important observation. It suggests that the disparities are mainly due to the model's underlying training data. The intrinsic linguistic properties of the languages seem to play a secondary role.

Text Length: Subsection 5.3.2 reveals a positive correlation between text length and translation direction detection accuracy. This observation is consistent with

prior research and seems intuitively logical — more text provides more context for accurate classification. As little as 60 characters are sufficient for the system to reliably determine translation direction across higher-resourced language pairs.

The thought that more text improves the detection accuracy is followed in the experiments, by providing document-level results by sentence-level label-based majority voting. The document-level results show that for settings, in which the detection is performing well at sentence-level, the accuracy can be improved even further.

Resource Setting: The scores for higher-resourced language pairs were higher than the ones for lower-resourced language pairs, indicating that resource levels form a caveat for this approach as they do in many other NLP subjects. However, the lower-resourced language pairs still achieved above-chance accuracy scores, especially for NMT outputs. An interesting observation was the accuracy drop and bias in predictions for language pairs with uneven resource availability, such as en↔zh. This is further discussed in Subsection 6.1.3.

Age of the Data: In Section 5.4, the accuracy scores for individual WMT datasets are presented. The results from the NMT datasets indicate an interesting trend: Newer systems tend to perform better on outputs generated by newer systems, while older systems show better results on outputs from older systems. Furthermore, the particularly high accuracy scores for the WMT22 HT dataset suggest that translations from the 2022 WMT are more easily identifiable in terms of translation direction. On the other hand, the lower scores observed for the WMT16 HT dataset could imply that the test dataset from that year might not be as meticulously curated in terms of translation direction as those from subsequent years. The addition of multiple reference translations to the WMT22 test set by the organizers underlines this point. Another point of interest is the consistently high accuracy scores for the de↔en pair across all models and translation strategies in the WMT23 dataset. This anomaly calls for further exploration, possibly through qualitative analyses or examining simple statistics like sentence length, to better understand the underlying factors.

Domain: The test dataset encompasses 11 different domains, including genres like news, manuals, and reviews, and mediums such as text and speech. Despite this variety, the top-performing systems yield similar results as previous work that was working with fewer domains. This indicates a certain degree of domain-independence.

Real-World Dataset: Finally, the real-world dataset experiment, employing both the overall best-performing model (SMaLL-100) and the best model for the specific language pair (Opus), shows SMaLL-100's superiority. The accuracies, although

lower than those for NMT outputs, are higher than for human translations. A lower classification accuracy can be expected when the experiment setting is less controlled. One aspect that is not controlled in this example is the translation strategy. So far, it is unknown whether the translations in this dataset were produced manually, automatically, or automatically with manual editing. The result could therefore be interpreted as an indication that the translations were indeed a product of a mixture of strategies. This would align with the accuracy positioned between the accuracies that are scored during the experiments for NMT and HT, respectively.

6.1.3 Bias and Normalization

One effect observable over all NMT models is a bias towards one of the translation directions in a language pair. The biases are found to be most pronounced in the NLLB models' results for language pairs like en↔zh, en↔cs, en↔uk, and en↔ha. One aspect that those language pairs have in common is the difference between data availability within the languages in the pair. English is a very high-resource language, with most of the world's research and language applications centering around it [Fan et al., 2021]. Languages like Czech and Ukranian provide less resources – the most extreme case in this set of examples being Hausa [Mohammadshahi et al., 2022]. Results for language pairs, where resource levels are more equal, show less bias. Taking these observations into account discrepancy between resource levels within the language pair might be the reason for the bias.

As a potential solution, reconstruction normalization as proposed by Vamvas and Sennrich [2022] is explored. The attempt proves to be semi-successful, because this strategy shows a positive effect in terms of bias mitigation only for models and language pairs, for which bias was most pronounced, e.g.: NLLB-dist-600M for en↔zh NMT. It is less successful where bias is small to begin with, e.g.: SMaLL-100 for en↔uk NMT. Furthermore, it has a negative effect on the classification accuracy of almost all NMT and HT data – for pre-NMT all accuracy scores improved. This outcome suggests that different models might require different normalization strategies and allows room for research for more elaborate and specific normalization and bias mitigation techniques according to the NMT model and translation strategy.

The question also arises why these biases are more pronounced in the newer and larger NLLB models, whose capability to translate from and into low-resource languages is emphasized in its corresponding paper [Akula et al., 2022]. The models employed several different techniques to compensate for the low-resource settings, such as sharing the model's capacity over multiple translation directions, elaborate

data mining techniques, and back-translation. All of these techniques might be reflected in their model's translation probabilities.

One characteristic, which sets the NLLB models apart from the other models is its large-scale use of back-translation. While it improves the translation performance on low-resource languages, introducing training data inverse to the actual translation direction of the model might affect the model's ability to detect translationese. However, this hypothesis would have to be tested in a more controlled setting.

6.1.4 Translation Probabilities

The view on the translation probability per translation strategy in Subsection 5.5 underlines the existing results and provides room for interpretation. The translation probabilities for HT are on average the lowest in both the original and the inverse direction. This implies that the NMT model "perceives" HT as less probable than both NMT and pre-NMT in both directions. Pre-NMT data in the original direction yield similar probabilities as HT, in the opposite direction, however, the probabilities are on average higher than the other two, almost displaying the same probabilities as NMT in the original translation direction, which is ultimately the reason for the low accuracy scores in this category. The probabilities for NMT are the highest in the original direction, which seems intuitive, given that the model producing the probabilities is an NMT model – it assigns the highest output to those translations that it would most likely generate itself. The inverse probabilities are slightly lower, thus, providing grounds for the successful translation direction detection. Overall, the probabilities for MT for both directions are higher than those for HT, which reflects the underlying difference between those strategies.

6.1.5 Comparison to Previous Work

Thematically, the closest previous work to what has been presented here is by Sominsky and Wintner [2019] (see: Section 3.1). Their work on sentence-level direction detection reaches results in a similar range for de↔fr to the ones reported in Section 5.1 using the SMaLL-100. Hence, the results achieved in this work can be regarded as competitive. However, the unsupervised approach here comes with two great benefits over the supervised approach by Sominsky and Wintner [2019]. For one, this approach requires no training data at all, being therefore much less resource intensive and practical to use. Secondly, the results from this approach are calculated over 11 domains, while the maximum reported in Sominsky and Wintner [2019] is

3 domains. Nonetheless, the single-domain results by Sominsky and Wintner [2019] are not exceeded.

This work's results can be compared to other unsupervised approaches for similar tasks. For instance, Rabinovich and Wintner [2015] achieved an average accuracy of 82% in mixed-domain settings. These settings involved 1 to 3 domains (again en↔fr) and used text chunks of 2000 character length. In contrast, the approach proposed in this work shows almost comparable performance. It reaches an accuracy of 78% at the document level in the same language pair. Notably, it requires less data to achieve this result. Furthermore, this work includes over 11 domains in its dataset. This suggests that the approach is more robust across different domains.

A third comparison with Vamvas and Sennrich [2022] should provide insight into how using translation probabilities for translation direction detection compares to other methods in which translation probabilities are used. Although the scores are not directly comparable, because Vamvas and Sennrich [2022] use the probabilities for a different task, they are compared nonetheless to emphasize the capabilities of translation probability-based approaches. The accuracy scores reached with translation-based measures for text similarity show a range of 65% to 77% (without normalization). This aligns with the range for single sentence pairs in the translation direction detection task here. One major difference to Vamvas and Sennrich's approach is, however, that the reconstruction normalization improves the results for their task, which is not the case here.

6.2 Limitations

Although the presented method has its advantages, which have been highlighted multiple times within this work, there are some limitations to be aware of. One of the most evident limitations is the approach's susceptibility to bias toward one translation direction. The results suggest that the bias varies to high degrees, depending on which language directions are to be detected and which model is used to generate the probabilities. Although the experiments in this work have given insight into which models are most prone to the bias for which language pairs and a normalization approach to mitigate the bias has been proposed, an optimal solution was not found.

Another limitation is weakness when it comes to low resource settings. These two limitations are presumably closely tied to the NMT model's training data (im)balance and the resulting translation performance, which leads to the next lim-

itation, namely, the method's reliance on existing large models. Training a whole translation model for such a task seems uneconomical; hence, one needs to rely on publicly available models.

Fortunately, the machine translation community provides a number of publicly available models. These models, however, can be very large and sometimes require special hardware to run, which drastically reduces their accessibility. If the models can be used without specialized hardware, they are still large and require a substantial amount of time to generate the probabilities for a large number of translations in both directions.

Ultimately, even when all the specified conditions are fulfilled, the approach outlined in this paper does not fully resolve the translation direction task. There is significant potential for improvement, especially for pre-NMT data. While the results for NMT and HT data are satisfactory, they are not without flaws. Additionally, efforts to mitigate bias have not been entirely effective. This underscores the fact that this research is intended to serve as a first foundation for future studies.

6.3 Future Work

This research has opened up several avenues for future exploration in the field of translation probability-based translation direction detection. Being one of the first studies of its kind, it paves the way for a more in-depth understanding and enhancement of the methodologies used, as well as exploration of new areas. The following subsections outline some potential directions for future research.

6.3.1 Expanding the Described Experiments

This study represents a starting point for translation probability-based translation direction detection. Future research can build upon this foundation, exploring new methods and refining existing ones to improve accuracy and reliability.

In this work, one possibility is described of how to generate translation probabilities. Alternative approaches to generating translation probabilities could be explored. Different methods might yield more accurate or insightful results.

This can also be applied to the probability normalization process, where a more nuanced method could be developed, perhaps also specified for certain models and language pairs. For example, the Opus models were not covered by the normalization

method in this work.

Additionally, the document-level result inference could be explored with a more subtle voting method, e.g., basing the vote on the average translation probabilities (soft voting) rather than the sentence labels.

Lastly, in this work, only a handful of possible influences on the system were chosen for investigation. This set of influences could be expanded and explored in more detail. For example, the influence of back-translation in training data might be an interesting aspect to investigate in a more controlled setting, where NMT models are specifically trained for the experiments at hand.

6.3.2 Research on Low-Resource Languages

The performance of translation models on the direction detection task seems closely linked to the resource availability of the languages involved. Future research should focus even more on improving translation model performances for low-resource languages, particularly those overshadowed by high-resource languages like English. This would not only enrich the research field of machine translation and improve the translation quality of MT systems but also broaden the accessibility of these systems.

6.3.3 Translation Detection

As Section 3 illustrated, a lot of previous research has focused on translation detection, rather than translation *direction* detection. Reformulating the task to the former has one main advantage: There is no need for parallel text. Monolingual data suffice to detect whether the text is an original or a translation. Future research could explore the application of the methodologies proposed in this work for translation detection as an even less resource-intensive alternative.

6.3.4 Translation Strategy Identification

Lastly, the translation probabilities that have been explored in this work have shown to be distinctive properties for each translation strategy (HT, NMT, pre-NMT). These probabilities could potentially be used to automatically identify the translation strategy. Future studies might explore the use of clustering methods, leveraging probabilities and differences in probabilities as key features. Alternatively, compar-

ing the translation probabilities to the probabilities of a reference corpus might give insight into which translation strategy has been applied.

7 Conclusion

The main goal of this thesis was to investigate the potential for determining the translation direction in parallel sentence pairs using an unsupervised approach. Although there are existing solutions for this problem, all previous work required a substantial amount of homogeneous parallel data to succeed, making it unfit for most real-world scenarios.

A different stream of research has looked into exploiting translation probabilities produced by neural machine translation models for text similarity tasks. This has proven successful for tasks, in which a high level of attention to detail for the texts was necessary, requiring minimal amounts of parallel data. In this work, these two research streams have been combined.

Here, the conditional translation probabilities generated by NMT models have been compared for the same sentence pair in each of the two possible translation directions. It was hypothesized that the NMT model would assign a higher probability to the sentence pair in its original translation direction, and thereby uncover which direction is the original one. Once the sentence-level results had been obtained, they were used to infer the results for whole documents. Furthermore, the results were checked for biases, which, in turn, led to a bias mitigation attempt using probability normalization.

This approach was tested using a selection of different NMT models to generate the translation probabilities based on a diverse dataset. The dataset included a variety of language pairs with different resource availabilities, human, pre-neural, and neural machine translations, a variety of domains, and an extra test set from a real plagiarism allegation case.

The findings confirm that the translation-probability-based approach is valid for unsupervised translation direction detection. Good results were shown at sentence level and even better results at document level. The best results in this work have shown this approach to be comparable to previous methods, but also offer a less resource-intensive option and demonstrate robustness, particularly when analyzing neural machine translations and human translations. However, there are multiple

factors to consider. Although there is one system that works best for nearly all conditions, it is advised to choose the NMT model and normalization strategy according to the application setting, which is defined by language pair, resource level of the included languages, and translation strategy.

The work aligns with previous research on translation probabilities, indicating that they are indeed sensitive to the stylistic differences between parallel texts. The findings from this work suggest that these translation probabilities can be connected to translation studies and provide room for linguistically motivated explanations.

While the unsupervised translation direction detection approach demonstrated in this thesis shows positive results, it also uncovers areas requiring further investigation and refinement. Considering the novelty of this approach, I hope, this work can be used as a foundation for future research.

References

R. Aharoni, M. Johnson, and O. Firat. Massively multilingual neural machine translation. *arXiv preprint arXiv:1903.00089*, 2019.

F. Akhbardeh, A. Arkhangorodsky, M. Biesialska, O. Bojar, R. Chatterjee, V. Chaudhary, M. R. Costa-jussa, C. España-Bonet, A. Fan, C. Federmann, M. Freitag, Y. Graham, R. Grundkiewicz, B. Haddow, L. Harter, K. Heafield, C. Homan, M. Huck, K. Amponsah-Kaakyire, J. Kasai, D. Khashabi, K. Knight, T. Kocmi, P. Koehn, N. Lourie, C. Monz, M. Morishita, M. Nagata, A. Nagesh, T. Nakazawa, M. Negri, S. Pal, A. A. Tapo, M. Turchi, V. Vydrin, and M. Zampieri. Findings of the 2021 conference on machine translation (WMT21). In *Proceedings of the Sixth Conference on Machine Translation*, pages 1–88, Online, Nov. 2021. Association for Computational Linguistics. URL `https://aclanthology.org/2021.wmt-1.1`.

B. Akula, L. Barrault, G. M. Gonzalez, P. Hansanti, and J. Hoffman. No Language Left Behind: Scaling Human-Centered Machine Translation - Meta Research. 2022.

N. Arivazhagan, A. Bapna, O. Firat, D. Lepikhin, M. Johnson, M. Krikun, M. X. Chen, Y. Cao, G. Foster, C. Cherry, W. Macherey, Z. Chen, and Y. Wu. Massively Multilingual Neural Machine Translation in the Wild: Findings and Challenges, 2019. URL `https://arxiv.org/abs/1907.05019`.

M. Artetxe, G. Labaka, E. Agirre, and K. Cho. Unsupervised neural machine translation. *arXiv preprint arXiv:1710.11041*, 2017.

M. Artetxe, S. Ruder, and D. Yogatama. On the cross-lingual transferability of monolingual representations. *arXiv preprint arXiv:1910.11856*, 2019.

M. Artetxe, G. Labaka, and E. Agirre. Translation artifacts in cross-lingual transfer learning. *EMNLP 2020 - 2020 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, pages 7674–7684, 2020. doi: 10.18653/v1/2020.emnlp-main.618.

E. A. Avner, N. Ordan, and S. Wintner. Identifying translationese at the word and sub-word level. *Digit. Scholarsh. Humanit.*, 31:30–54, 2016. URL https://api.semanticscholar.org/CorpusID:1389695.

D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate, 2016.

M. Baker. Corpus Linguistics and Translation Studies — Implications and Applications. *Text and Technology*, page 233, jun 1993. doi: 10.1075/Z.64.15BAK. URL https://benjamins.com/catalog/z.64.15bak.

M. Baker. Corpus linguistics and translation studies*: Implications and applications. In *Researching translation in the age of technology and global conflict*, pages 9–24. Routledge, 2019.

M. Baroni and S. Bernardini. A new approach to the study of translationese: Machine-learning the difference between original and translated text. *Literary and Linguistic Computing*, 21(3):259–274, 2006. ISSN 02681145. doi: 10.1093/llc/fqi039.

L. Barrault, O. Bojar, M. R. Costa-jussà, C. Federmann, M. Fishel, Y. Graham, B. Haddow, M. Huck, P. Koehn, S. Malmasi, C. Monz, M. Müller, S. Pal, M. Post, and M. Zampieri. Findings of the 2019 conference on machine translation (WMT19). In O. Bojar, R. Chatterjee, C. Federmann, M. Fishel, Y. Graham, B. Haddow, M. Huck, A. J. Yepes, P. Koehn, A. Martins, C. Monz, M. Negri, A. Névéol, M. Neves, M. Post, M. Turchi, and K. Verspoor, editors, *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy, Aug. 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-5301. URL https://aclanthology.org/W19-5301.

Y. Bizzoni, T. S. Juzek, C. España-Bonet, K. Dutta Chowdhury, J. van Genabith, and E. Teich. How Human is Machine Translationese? Comparing Human and Machine Translations of Text and Speech. pages 280–290, 2020. doi: 10.18653/v1/2020.iwslt-1.34.

S. Blum and E. A. Levenston. Universals of lexical simplification. *Language learning*, 28(2):399–415, 1978.

S. Blum-Kulka. Shifts of cohesion and coherence in translation shoshana blum-kulka, hebrew university of jerusalem. *Interlingual and intercultural communication: Discourse and cognition in translation and second language acquisition studies*, 272:17, 1986.

N. Bogoychev and R. Sennrich. Domain, Translationese and Noise in Synthetic Data for Neural Machine Translation. (November 2019), 2019. URL http://arxiv.org/abs/1911.03362.

O. Bojar, R. Chatterjee, C. Federmann, Y. Graham, B. Haddow, M. Huck, A. Jimeno Yepes, P. Koehn, V. Logacheva, C. Monz, M. Negri, A. Névéol, M. Neves, M. Popel, M. Post, R. Rubino, C. Scarton, L. Specia, M. Turchi, K. Verspoor, and M. Zampieri. Findings of the 2016 Conference on Machine Translation. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 131–198, Berlin, Germany, aug 2016. Association for Computational Linguistics. doi: 10.18653/v1/W16-2301. URL https://aclanthology.org/W16-2301.

J. Brownlee. How to develop voting ensembles with python, 2021. URL https://machinelearningmastery.com/voting-ensembles-with-python/.

F. Burlot and F. Yvon. Using monolingual data in neural machine translation: a systematic study. *arXiv preprint arXiv:1903.11437*, 2019.

C. Callison-Burch, M. Osborne, and P. Koehn. Re-evaluating the Role of BLEU in Machine Translation Research. In *Proceedings of the European Association of Computational Linguistics (EACL)*, pages 249–256, 2006. URL http://www.aclweb.org/anthology/E/E06/E06-1032.pdf.

A. Chesterman, A. Mauranen, P. Kujama, et al. Translation universals: Do they exist, 2004.

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

J. Czakon. 24 evaluation metrics for binary classification (and when to use them), 2023. URL https://neptune.ai/blog/evaluation-metrics-binary-classification.

T. Domhan and F. Hieber. Using target-side monolingual data for neural machine translation through multi-task learning. 2017.

U. Ebbinghaus. Der quacksalber. *Frankfurter Allgemeine Zeitung*, 2022. URL https://www.faz-biblionet.de/faz-portal/document?uid=FAZE__fazein_8488193&token=8862546b-1857-4e0a-8125-778711b00c03&p._scr=

faz-archiv&p.q=colchicine&p.source=&p.max=30&p.sort=&p.offset=0&p._ts=1699730134436&p.DT_from=01.11.1949&p.timeFilterType=0.

S. Edunov, M. Ott, M. Auli, and D. Grangier. Understanding back-translation at scale. *arXiv preprint arXiv:1808.09381*, 2018.

A. Fan, S. Bhosale, H. Schwenk, Z. Ma, A. El-Kishky, S. Goyal, M. Baines, O. Celebi, G. Wenzek, V. Chaudhary, N. Goyal, T. Birch, V. Liptchinsky, S. Edunov, E. Grave, M. Auliy, and A. Jouliny. Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research*, 22: 1–38, 2021. ISSN 15337928.

O. Firat, K. Cho, and Y. Bengio. Multi-way, multilingual neural machine translation with a shared attention mechanism. *arXiv preprint arXiv:1601.01073*, 2016.

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

X. Garcia, A. Siddhant, O. Firat, and A. P. Parikh. Harnessing multilinguality in unsupervised machine translation for rare languages. *arXiv preprint arXiv:2009.11201*, 2020.

J. Gehring, M. Auli, D. Grangier, D. Yarats, and Y. N. Dauphin. Convolutional sequence to sequence learning. In *International conference on machine learning*, pages 1243–1252. PMLR, 2017.

M. Gellerstam. Translationese in swedish novels translated from english. *Translation studies in Scandinavia*, 1:88–95, 1986.

N. Goyal, C. Gao, V. Chaudhary, P.-J. Chen, G. Wenzek, D. Ju, S. Krishnan, M. Ranzato, F. Guzman, and A. Fan. The flores-101 evaluation benchmark for low-resource and multilingual machine translation, 2021.

Y. Graham, B. Haddow, and P. Koehn. Statistical power and translationese in machine translation evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 72–81, Online, Nov. 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.6. URL https://aclanthology.org/2020.emnlp-main.6.

T. Harison. Machine translate: Wmt, 2023. URL https://machinetranslate.org/wmt.

G. Hinton, O. Vinyals, and J. Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

M. Hu, Y. Peng, F. Wei, Z. Huang, D. Li, N. Yang, and M. Zhou. Attention-guided answer distillation for machine reading comprehension. In E. Riloff, D. Chiang, J. Hockenmaier, and J. Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2077–2086, Brussels, Belgium, Oct.-Nov. 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1232. URL https://aclanthology.org/D18-1232.

I. Ilisei and D. Inkpen. Translationese traits in Romanian newspapers: A machine learning approach. *International Journal of Computational Linguistics and Applications*, 2(1-2):319–332, 2011.

I. Ilisei, D. Inkpen, G. Corpas Pastor, and R. Mitkov. Identification of Translationese: A Machine Learning Approach Iustina. In *Computational Linguistics and Intelligent Text Processing*, 2010. ISBN 978-3-642-12115-9. doi: 10.1007/978-3-642-12116-6. URL http://www.scopus.com/inward/record.url?eid=2-s2.0-78650435597&partnerID=tZOtx3y1.

T. Joachims. Text categorization with support vector machines: Learning with many relevant features. In *European conference on machine learning*, pages 137–142. Springer, 1998.

T. Joachims. Making large-scale svm learning. *Practical Advances in Kernel Methods-Support Vector Learning*, 1999.

M. Johnson, M. Schuster, Q. V. Le, M. Krikun, Y. Wu, Z. Chen, N. Thorat, F. Viégas, M. Wattenberg, G. Corrado, et al. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351, 2017.

I. T. Jolliffe. Principal component analysis. *Technometrics*, 45(3):276, 2003.

M. Junczys-Dowmunt. Dual Conditional Cross-Entropy Filtering of Noisy Parallel Corpora. *WMT 2018 - 3rd Conference on Machine Translation, Proceedings of the Conference*, 2:888–895, 2018. doi: 10.18653/v1/w18-6478.

M. Junczys-Dowmunt, R. Grundkiewicz, T. Dwojak, H. Hoang, K. Heafield, T. Neckermann, F. Seide, U. Germann, A. Fikri Aji, N. Bogoychev, A. F. T.

Martins, and A. Birch. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia, July 2018. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/P18-4020.

P. Juola. Authorship attribution. *Foundations and Trends® in Information Retrieval*, 1:233–334, 03 2008. doi: 10.1561/1500000005.

Y. Kim and A. M. Rush. Sequence-level knowledge distillation. In J. Su, K. Duh, and X. Carreras, editors, *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1317–1327, Austin, Texas, Nov. 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1139. URL https://aclanthology.org/D16-1139.

W.-J. Ko, A. El-Kishky, A. Renduchintala, V. Chaudhary, N. Goyal, F. Guzmán, P. Fung, P. Koehn, and M. Diab. Adapting high-resource nmt models to translate low-resource related languages without parallel data. *arXiv preprint arXiv:2105.15071*, 2021.

T. Kocmi, C. Federmann, R. Grundkiewicz, C. Monz, M. Novák, R. Bawden, M. Fishel, B. Haddow, M. Morishita, M. Popel, O. Bojar, T. Gowda, R. Knowles, M. Nagata, M. Popović, A. Dvorkovich, Y. Graham, P. Koehn, T. Nakazawa, and M. Shmatova. Findings of the 2022 Conference on Machine Translation (WMT22). *Conference on Machine Translation - Proceedings*, pages 1–45, 2022. ISSN 27680983.

M. Koppel and N. Ordan. Translationese and its dialects. *ACL-HLT 2011 - Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 1:1318–1326, 2011.

M. Koppel, J. Schler, and S. Argamon. Computational methods in authorship attribution. *JASIST*, 60:9–26, 01 2009. doi: 10.1002/asi.20961.

T. Kudo and J. 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, Brussels, Belgium, Nov. 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-2012. URL https://aclanthology.org/D18-2012.

D. Kurokawa, C. Goutte, and P. Isabelle. Automatic Detection of Translated Text and its Impact on Machine Translation. *MT-Summit-2009*, pages 81–88, 2009. URL http://www.mt-archive.info/MTS-2009-Kurokawa.pdf.

G. Lample, M. Ott, A. Conneau, L. Denoyer, and M. Ranzato. Phrase-based & neural unsupervised machine translation. *arXiv preprint arXiv:1804.07755*, 2018.

S. Läubli, R. Sennrich, and M. Volk. Has machine translation achieved human parity? a case for document-level evaluation. *arXiv preprint arXiv:1808.07048*, 2018.

S. Läubli, S. Castilho, G. Neubig, R. Sennrich, Q. Shen, and A. Toral. A set of recommendations for assessing human-machine parity in language translation. *Journal of Artificial Intelligence Research*, 67:653–672, 2020. ISSN 10769757. doi: 10.1613/JAIR.1.11371.

S. Laviosa. *Corpus-based translation studies: theory, findings, applications*, volume 17. Brill, 2021.

S. Lloyd. Least squares quantization in pcm. *IEEE Transactions on Information Theory*, 28(2):129–137, 1982. doi: 10.1109/TIT.1982.1056489.

J. Mallinson, R. Sennrich, and M. Lapata. Paraphrasing revisited with neural machine translation. *15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017 - Proceedings of Conference*, 2(2003):881–893, 2017. doi: 10.18653/v1/e17-1083.

H. A. Maurer, F. Kappe, and B. Zaka. Plagiarism-a survey. *J. Univers. Comput. Sci.*, 12(8):1050–1084, 2006.

A. Mohammadshahi, V. Nikoulina, A. Berard, C. Brun, J. Henderson, and L. Besacier. SMaLL-100: Introducing Shallow Multilingual Machine Translation Model for Low-Resource Languages. *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022*, pages 8348–8359, 2022. doi: 10.18653/v1/2022.emnlp-main.571.

R. C. Moore and W. Lewis. Intelligent selection of language model training data. In J. Hajič, S. Carberry, S. Clark, and J. Nivre, editors, *Proceedings of the ACL 2010 Conference Short Papers*, pages 220–224, Uppsala, Sweden, July 2010. Association for Computational Linguistics. URL `https://aclanthology.org/P10-2041`.

T. Q. Nguyen and D. Chiang. Transfer learning across low-resource, related languages for neural machine translation. *arXiv preprint arXiv:1708.09803*, 2017.

J. Ni, Z. Jin, M. Freitag, M. Sachan, and B. Schölkopf. Original or Translated? A Causal Analysis of the Impact of Translationese on Machine Translation Performance. *NAACL 2022 - 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference*, pages 5303–5320, 2022. doi: 10.18653/v1/2022.naacl-main.389.

S. Nisioi. Unsupervised classification of translated texts. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 9103:323–334, 2015. ISSN 16113349. doi: 10.1007/978-3-319-19581-0_29.

J. Olsson and J. Luchjenbroers. *Forensic linguistics*. Bloomsbury Publishing, 2014.

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

A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala. Pytorch: An imperative style, high-performance deep learning library, 2019.

M. Popescu. Studying translationese at the character level. *International Conference Recent Advances in Natural Language Processing, RANLP*, (September):634–639, 2011. ISSN 13138502.

D. Pylypenko, K. Amponsah-Kaakyire, K. D. Chowdhury, J. van Genabith, and C. España-Bonet. Comparing Feature-Engineering and Feature-Learning Approaches for Multilingual Translationese Classification. *EMNLP 2021 - 2021 Conference on Empirical Methods in Natural Language Processing, Proceedings*, pages 8596–8611, 2021. doi: 10.18653/v1/2021.emnlp-main.676.

E. Rabinovich and S. Wintner. Unsupervised Identification of Translationese. *Transactions of the Association for Computational Linguistics*, 3:419–432, 2015. doi: 10.1162/tacl_a_00148.

E. Rabinovich, S. Nisioi, N. Ordan, and S. Wintner. On the similarities between native, non-native and translated texts. *54th Annual Meeting of the Association for Computational Linguistics, ACL 2016 - Long Papers*, 4:1870–1881, 2016. doi: 10.18653/v1/p16-1176.

P. Riley, I. Caswell, M. Freitag, and D. Grangier. Translationese as a language in" multilingual" nmt. *arXiv preprint arXiv:1911.03823*, 2019.

R. Rubino, E. Lapshinova-Koltunski, and J. van Genabith. Information density and quality estimation features as translationese indicators for human translation classification. In K. Knight, A. Nenkova, and O. Rambow, editors, *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 960–970, San Diego, California, June 2016. Association for Computational Linguistics. doi: 10.18653/v1/N16-1110. URL `https://aclanthology.org/N16-1110`.

R. Sennrich. Advanced techniques of machine translation 01: Introduction + refresher deep learning and lms: A probabilistic model of translation. University Lecture Slides, 2022.

R. Sennrich, B. Haddow, and A. Birch. Improving neural machine translation models with monolingual data. In K. Erk and N. A. Smith, editors, *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany, Aug. 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1009. URL `https://aclanthology.org/P16-1009`.

I. Sominsky and S. Wintner. Automatic detection of translation direction. *International Conference Recent Advances in Natural Language Processing, RANLP*, 2019-Septe:1131–1140, 2019. ISSN 13138502. doi: 10.26615/978-954-452-056-4_130.

R. Sousa-Silva. Computational forensic linguistics: an overview of computational applications in forensic contexts. *Language and Law/Linguagem e Direito*, 5(2): 114–143, 2018.

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

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

J. Tiedemann. Parallel data, tools and interfaces in OPUS. In N. Calzolari, K. Choukri, T. Declerck, M. U. Doğan, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, and S. Piperidis, editors, *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey, May 2012. European Language Resources Association (ELRA). URL http://www.lrec-conf.org/proceedings/lrec2012/pdf/463_Paper.pdf.

J. Tiedemann and O. de Gibert. The OPUS-MT dashboard – a toolkit for a systematic evaluation of open machine translation models. In D. Bollegala, R. Huang, and A. Ritter, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 315–327, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-demo.30. URL https://aclanthology.org/2023.acl-demo.30.

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

A. Toral. Post-editese: an Exacerbated Translationese. 2019. URL http://arxiv.org/abs/1907.00900.

A. Toral and V. M. Sánchez-Cartagena. A multifaceted evaluation of neural versus phrase-based machine translation for 9 language directions. *15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017 - Proceedings of Conference*, 2(i):1063–1073, 2017. doi: 10.18653/v1/e17-1100.

A. Toral, S. Castilho, K. Hu, and A. Way. Attaining the unattainable? reassessing claims of human parity in neural machine translation. *arXiv preprint arXiv:1808.10432*, 2018.

G. Toury. *In search of a theory of translation*. Porter Institute for Poetics and Semiotics, Tel Aviv University, 1980.

G. Toury. Descriptive translation studies: And beyond. *Descriptive Translation Studies*, pages 1–366, 2012.

J. Vamvas and R. Sennrich. NMTSCORE: A Multilingual Analysis of Translation-based Text Similarity Measures. *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 198–213, 2022. doi: 10.18653/v1/2022.findings-emnlp.15.

H. Van Halteren. Source language markers in EUROPARL translations. *Coling 2008 - 22nd International Conference on Computational Linguistics, Proceedings of the Conference*, 1(August):937–944, 2008. doi: 10.3115/1599081.1599199.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

E. Voita, R. Sennrich, and I. Titov. Language Modeling, Lexical Translation, Reordering: The Training Process of NMT through the Lens of Classical SMT. *EMNLP 2021 - 2021 Conference on Empirical Methods in Natural Language Processing, Proceedings*, pages 8478–8491, 2021. doi: 10.18653/v1/2021.emnlp-main.667.

V. Volansky, N. Ordan, and S. Wintner. On the features of translationese. *Digital Scholarship in the Humanities*, 30(1):98–118, 2015. ISSN 2055768X. doi: 10.1093/llc/fqt031.

J. H. Ward. Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association*, 58(301):236–244, 1963. doi: 10.1080/01621459.1963.10500845. URL `https://www.tandfonline.com/doi/abs/10.1080/01621459.1963.10500845`.

T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, and A. M. Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, Oct. 2020. Association for Computational Linguistics. URL `https://www.aclweb.org/anthology/2020.emnlp-demos.6`.

J. Zenthöfer. Was wurde hier gefälscht? *Frankfurter Allgemeine Zeitung*, 2022a. URL `https://www.faz-biblionet.de/faz-portal/document?uid=FAZ_FD120221008500022145786l4&token=ff1d730a-820d-4bc7-8598-92cc827481c6&p._scr=faz-archiv&p.q=colchicine&p.source=&p.max=30&p.sort=&p.offset=0&p._ts=1699733084333&p.DT_from=01.11.1949&p.timeFilterType=0`.

J. Zenthöfer. Chronik einer plagiats-intrige. *Frankfurter Allgemeine Zeitung*, 2022b. URL `https://www.faz-biblionet.de/faz-portal/document?uid=FAZN_`

_20221018_8395173&token=7e6e3835-bb30-4993-86a8-3ad649989fd0&p._scr=faz-archiv&p.q=colchicine&p.source=&p.max=30&p.sort=&p.offset=0&p._ts=1699729207907&p.DT_from=01.11.1949&p.timeFilterType=0.

B. Zhang, P. Williams, I. Titov, and R. Sennrich. Improving massively multilingual neural machine translation and zero-shot translation. *arXiv preprint arXiv:2004.11867*, 2020.

M. Zhang and A. Toral. The effect of translationese in machine translation test sets. *arXiv preprint arXiv:1906.08069*, 2019.

C. Zhou, G. Neubig, and J. Gu. Understanding knowledge distillation in non-autoregressive machine translation. *arXiv preprint arXiv:1911.02727*, 2019.

B. Zoph, D. Yuret, J. May, and K. Knight. Transfer learning for low-resource neural machine translation. *arXiv preprint arXiv:1604.02201*, 2016.

A Figures

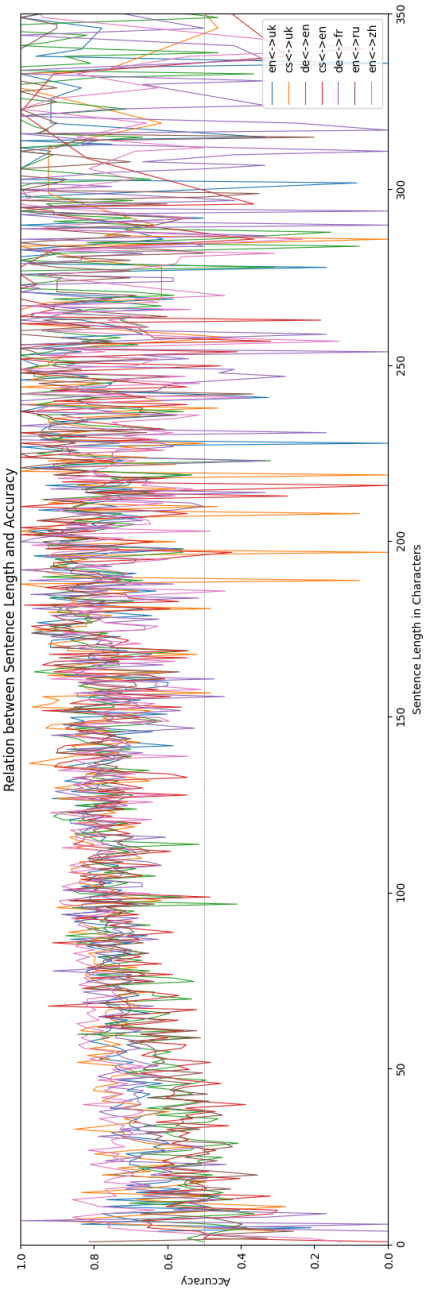

Figure 4: Larger view on the effect of sentence length on accuracy per language pair with more outliers towards the end while trend stays the same.

Curriculum Vitae

Personal Details

Full name	Michelle Wastl
Address	Viktoriastrasse 23
	8057 Zürich
E-mail address	michelle.wastl@uzh.ch

Education

since 2020	Master of Arts in Computational Linguistics and Language Technology and Methods – Data – Society at the University of Zurich
2016-2020	Bachelor of Arts in German Language and Literature and Comparative Linguistics at the University of Zurich

Relevant Professional Activities

since 2021	Programming Assistant at the Zurich School of Applied Sciences, School of Management and Law
2022	Teaching Assistant for the module *Intermediate Methods and Programming in Digital Linguistics* at the University of Zurich, Institute of Computational Linguistics
2019-2021	Student Research Assistant at the Swiss Federal Institute of Technology Zurich, Department of Management, Technology, and Economics

Declaration of Independent Authorship

Original work

I expressly declare that the written work I submitted to the University of Zurich in the ~~spring~~/autumn semester of 2023 with the title

Unsupervised Translation Direction Detection

is an original work written by myself, in my own words, and without unauthorized assistance. If it is a work by several authors, I confirm that the relevant parts of the work are correctly and clearly marked and can be clearly assigned to the respective author.

I also confirm that the work has not been submitted in whole or in part to receive credit for another module at the University of Zurich or another educational institution, nor will it be submitted in the future.

Use of sources

I expressly declare that I have identified all references to external sources (including tables, graphics etc.) contained in the above work as such. In particular, I confirm that, without exception and to the best of my knowledge, I have indicated the authorship both for verbatim statements (citations) and for statements by other authors reproduced in my own words (paraphrases).

Use of text generation models

I expressly declare that I have not only identified existing external sources, but also any automatically generated text that is contained in the above work. I have used the same citation style as if the text had been generated by a human to indicate the source of the automatically generated text. If the contribution of text generation models cannot be linked to specific text passages (see the associated guidelines), I have included a chapter describing the contributions of the text generation model. I acknowledge that no explicit citation is necessary where text generation models are merely used correctively (to improve grammar or idiomaticity of my own words).

Sanctions

I acknowledge that a thesis that is used to acquire credit and proves to be plagiarism with the meaning of the document *Erläuterung des Begriffs „Plagiat“* leads to a grade

deduction in minor cases, a grade 1 (one) in more severe cases, without the possibility of revision, and in very severe cases can have the corresponding legal and disciplinary consequences according to §§ 7ff of the “Disziplinarordnung der Universität Zürich” and § 36 of the “Rahmenordnung für das Studium in den Bachelor- und Master-Studiengängen der Philosophischen Fakultät der Universität Zürich”.

I confirm with my signature that this information is correct:

Name: Wastl First Name: Michelle

Matriculation number: 16-727-398

Date: 01.12.2023 Signature: