Evaluating and Improving Lexical Language Understanding in Neural Machine Translation

Denis Emelin

Doctor of Philosophy

Institute for Language, Cognition and Computation
School of Informatics
University of Edinburgh

2024

Abstract

Lexical understanding is an inalienable component of the translation process. In order to correctly map the meaning of a linguistic unit to the appropriate target language expression, the meaning of its constituent words has first to be identified and disambiguated, followed by the application of compositional operations. This thesis examines the competency of contemporary neural machine translation (NMT) models on two core aspects of lexical understanding – word sense disambiguation (WSD) and coreference resolution (CoR), both of which are well-established and much-studied natural language processing (NLP) tasks. Certain linguistic properties that are underspecified in a source language (e.g. the grammatical gender of a noun in English) may need to be stated explicitly in the chosen target language (e.g. German). Doing so correctly requires the accurate resolution of the associated ambiguities.

While recent modeling advances appear to suggest that both WSD and CoR are largely solved challenges in machine translation, the work conducted within the scope of this thesis demonstrates that this is not yet the case. In particular, we show that NMT systems are prone to relying on surface-level heuristics and data biases to guide their lexical disambiguation decisions, rather than engaging in deep language understanding by correctly recognizing and leveraging contextual disambiguation triggers. As part of our investigation, we introduce a novel methodology for predicting WSD errors a translation model is likely to make and utilize this knowledge to craft adversarial attacks with the aim to elicit disambiguation errors in model translations. Additionally, we create a set of challenging CoR benchmarks that uncover the inability of translation systems to identify referents of pronouns in contexts that presuppose commonsense reasoning, caused by their pathological over-reliance on data biases.

At the same time, we develop initial solutions for the identified model deficiencies. As such, we show that fine-tuning on de-biased data and modifying the learning objective of a model can significantly improve disambiguation performance by counteracting the harmful impact of data biases. We furthermore propose a novel extension to the popular transformer architecture that is found to strengthen its WSD capabilities and robustness to adversarial WSD attacks by facilitating the accessibility of lexical features across all layers of the model and increasing the extent to which contextual information is encapsulated with its latent representations. Despite the so effected improvements to WSD and CoR, both tasks remain far from solved, posing a veritable challenge for the current generation of NMT models, as well as for large language models that have risen to prominence within NLP in recent years.

Lay Summary

Translating text from one language into another necessarily requires a solid understanding of its constituent words. Since ambiguity is a pervasive feature of natural language, the meaning of a word can vary depending on its context. The disambiguation choice directly informs the choice of the corresponding target language expression during the translation step, e.g. in case of ambiguous nouns that can express several senses realized by different target language words, or pronouns that can point to different referents that must be explicitly marked in the generated translation. This thesis examines the extent to which contemporary neural machine translation (NMT) systems are able to correctly disambiguate polysemous words, a task commonly referred to as word sense disambiguation (WSD), and identify the intended referents of ambiguous pronouns, a task known as coreference resolution (CoR).

More specifically, the work presented herein proposes novel evaluation methods and benchmarks that can be used to assess the WSD and CoR capabilities of NMT systems. Evaluating models in this manner reveals that, rather than engaging in human-like reasoning about the intended meaning of words, they are guided in their disambiguation decisions by undesirable biases and shallow heuristics. This, in turn, leads to high error rates on the proposed evaluation suits and heightened vulnerabilities to malicious inputs designed to elicit incorrect translations.

In addition to uncovering the shortcomings of translation systems with regard to lexical language understanding, the thesis considers different ways of alleviating them. This includes exposing models to training data that is demonstrably free of undesirable biases, explicitly encouraging correct disambiguation choices during model training, and modifying model design in a manner that encourages better utilization of the disambiguation-relevant information contained within the translated text. While the proposed strategies are shown to be effective, WSD and CoR continue to present a significant challenge to computational models of various sizes, designs, and complexities, requiring continued efforts from machine translation researchers.

Acknowledgements

The road leading up to the completion of this manuscript had been a long and winding one, strewn with personal challenges to be overcome. A global pandemic, an unjust war, and several deeply affecting personal events made me question the choice of pursuing the PhD candidacy, doubt my confidence in my abilities as a scholar, and caused my mental health to plummet to new lows. However, these challenges also encouraged me to address deeply rooted personal issues, find support in the words and actions of friends, and re-evaluate my life's goals and aspirations. Throughout all of this, and despite my best efforts to cause their trust in me to falter, I had the complete support of my PhD supervisors, Dr. Rico Sennrich and Dr. Ivan Titov, for which I will forever be grateful. Thank you, very much, for your sage advice, generosity, clarity of vision, and enduring tolerance. I am extremely fortunate to have had you as my advisors.

Many of the ideas that found their way onto the pages of this thesis arose from conversations with my peers at the Informatics Forum, many of whom have become dear friends. Sabine Weber and Tom Pelsmaeker, in particular, offered valuable insights that informed the work on lexical ambiguity presented herein, as well as invaluable emotional support. Equally, I am fortunate to be able to name many wonderful people who have enriched my PhD experience through their presence and companionship, including Ida, Seraphina, Stefanie, Ramon, Yevgen, Kate, Sander, Daniel et alia. Similarly, my family has always been a source of support and comfort for me, even when I haven't been deserving of it. To all of you, as well, I extend my deepest gratitude.

My life in Edinburgh has been full of high and lows, just like a day spent wandering around the city's streets may allow one to experience brilliant sunshine and torrential downpour, sometimes at the same time. Many of the highs have been due to the brilliant, unique, and unconventional thespians that make up the Beltane Fire Society, which I have been part of for some time. Without you, my creativity would not have found such strange and beautiful outlets. In my mind, you are an inalienable part of the cultural tapestry that pervades this city.

To anyone else who should have been included here: I'm sorry for not mentioning you and thank you all the same!

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*Denis Emelin*)

To mom.

Contents

Chapter 1

Introduction

Anfangen ist leicht, Beharren eine Kunst.

German proverb

In recent years, the field of machine translation saw a rapid transition from phrase-based statistical models to neural architectures utilizing dense representations of linguistic units. Such neural machine translation (NMT) models extract relevant features from the supplied training data according to the distributional hypothesis that posits that the meaning of a linguistic unit, e.g. a word or a phrase, can be derived from the contexts in which it occurs. Most often defined as a supervised task that relies on parallel, multi-lingual datasets and incorporates a likelihood maximization[1] objective, NMT has lead to marked qualitative improvements of translations produced by automated systems, especially for language pairs for which abundant training resources are available. But even in such ideal settings, NMT systems still struggle to adequately handle certain sub-tasks that are integral to translation, such as word sense disambiguation (WSD) and coreference resolution (CoR). Within the scope of this thesis, several projects were completed with the goal of exploring the limitations of popular NMT architectures through the lens of *lexical understanding*, i.e. errors and other deficiencies observed when reasoning about the meaning of words as part of the translation process, as well as their potential causes.

Our focus on lexical understanding draws a connecting line between the individual research projects completed as part of the PhD candidacy, leading up to the writing of this thesis. Specifically, it foregrounds lexical ambiguity as a shared cause of challenges associated with both WSD and CoR (see Chapters 3 and 4, respectively), and

[1]Or, more precisely, negative likelihood minimization.

highlights that improvements to NMT model design can improve the efficacy with which this ambiguity is resolved (see Chapter 5). This, in turn, results in a more self-contained and impactful contribution to a specific area of the highly diverse and ever-evolving NLP research landscape.

The main disadvantage to this framing is the simplifying assumptions made by the evocation of the term *lexical*. Although linguistic phenomena like WSD and CoR are of primary relevance to the translation of specific lexical units, i.e. words that are ambiguous in the source context but must be disambiguated as part of the translation step, the language understanding capabilities required to competently navigate these tasks extend beyond lexical boundaries by incorporating the wider context in which said lexical units occur. Therefore, our use of *lexical* is a simplification aimed at the outcome of the studied translation competencies, i.e. WSD and CoR, as their visible and easily verifiable consequence. This assumes the understanding, on the part of the reader, of the role played by the information found in the surroundings of the ambiguous lexical units in their translation. Additionally, it must be noted that since contemporary NMT systems predominantly operate on sub-word units (Sennrich et al., 2016) rather than word tokens, it is not inherently obvious in how far the linguistic notion of *lexical units* aligns itself with mechanistic realities of NMT implementations.

For the purposes of this thesis, we chose to subscribe to the notion of words as lexical units to better connect our work with the larger body of past research in both linguistics and NLP. Having thus established the intended meaning of *lexical understanding* as used within the context of this thesis, we hope that our chosen framing will ultimately be beneficial rather than distracting to the reader.

Past machine translation research introduced a variety of methods for delineating the limitations of translation systems. Among those, common strategies involve manual analysis of model outputs (Shi et al., 2016; Liu et al., 2018; Yehudai et al., 2023), examination of models' internal representations such as layer activations or attention distributions (Belinkov et al., 2017b,a; Marvin and Koehn, 2018; Tang et al., 2019), use of benchmark datasets (Guillou and Hardmeier, 2016; Müller et al., 2018; Raganato et al., 2019; He et al., 2020), and the engineering of adversarial attacks intended to target assumed model weaknesses (Cheng et al., 2019; Stojanovski et al., 2020; Zhang et al., 2020). Empirically verified insights into the deficiencies of best-performing models are one of the central prerequisites for devising better, more capable methods, making them essential to the directed progress towards improved, robust, more human-like translation quality. This thesis aims to contribute towards this goal, while also proposing extensions to popular NMT architectures that are intended to address identified model shortcomings.

Despite gradual improvements in translation accuracy and over-promising claims of human parity (Hassan et al., 2018), the landscape of commonly encountered machine translation errors remains a rich one. A non-exhaustive list includes premature termination of the generated translations (Wu et al., 2016; Murray and Chiang, 2018), failure to adequately translate terminology (Dinu et al., 2019; Skianis et al., 2020) and figurative language (Toral and Way, 2018; Dankers et al., 2022), incorrect nominal declination and verbal inflection (Moryossef et al., 2019; Voita et al., 2019a), hallucinations of target words not present in the input sentence (Wang and Sennrich, 2020; Raunak et al., 2021), and the failure to correctly disambiguate ambiguous or polysemous terms (Rios et al., 2017; Campolungo et al., 2022b). As addressing all possible error types goes well beyond the scope of a thesis, our investigation instead focuses on errors related to WSD and CoR, building upon and extending past work (Raganato et al., 2019; Sakaguchi et al., 2020).

The decision to situate WSD and CoR at the core of this thesis may seem surprising, since both represent long-standing research areas within NLP and may appear to offer little space for novel contributions as evidenced by saturated challenge sets (Raganato et al., 2019) or the apparent focus on long-tail phenomena in recent publications (Sakaguchi et al., 2020). We argue that both research directions are far from "solved" and attribute any indications to the contrary to insufficiently challenging benchmarks and the often downplayed significance of model robustness to edge-cases.

As a translation step, WSD is concerned with the disambiguation of polysemous source words given the available context, with the goal of producing an accurate target language translation that correctly captures the intended meaning of the input text. CoR, on the other hand, describes the process of determining the meaning of a source word (e.g. a pronoun) specifically by resolving its intended reference, i.e. by identifying – within the source context – mentions of the same entity, object, or event that the word in question points to. Given the ubiquity of ambiguity in natural language, competent and robust ambiguity resolution capabilities of translation systems are essential for ensuring the correctness of automatic translations. That such capabilities remain far from guaranteed is evidenced by the following examples obtained from Google Translate, a highly popular machine translation service that supports a large number of translation directions:

1. EN (input): *The bat flew across the field after waking up.*
   DE (output): *Der Schläger flog nach dem Aufwachen über das Feld.*[2]

2. EN (input): *The hurricane couldn't blow over the windmill because it was weak.*
   DE (output): *Der Hurrikan konnte die Windmühle nicht überwältigen, da sie schwach war.*

The first example represents an obvious WSD failure, where the English noun ***bat*** is incorrectly translated into German as ***Schläger*** (i.e. in the "sports equipment" sense) rather than as ***Fledermaus*** (i.e. in the "flying mammal" sense). The second example, on the other hand, shows an instance of erroneous CoR, whereby ***windmill*** is wrongly chosen to be the antecedent of the ambiguous pronoun ***it*** in the NMT translation, as indicated by the agreement in grammatical gender between ***Windmühle*** (the German translation of "windmill") and ***sie*** (one possible German translation of "it"). The resulting translation is clearly incorrect as it violates commonsense assumptions despite being grammatical. Although anecdotal, such evidence is symptomatic of the insufficient robustness of NMT models to disambiguation errors which, in our estimation, necessitates continued research into topics such as WSD and CoR.

Translation errors can arise from a variety of causes, some of which have been studied in depth. Among these, several stem from properties of the data that the model is exposed to during training, e.g. gender bias (Stanovsky et al., 2019; Renduchintala et al., 2021), positional bias (Stojanovski et al., 2020), low frequency phenomena (Sennrich et al., 2016; Zhang et al., 2022), or domain mismatch in its various manifestations (Müller et al., 2020; Saunders and Byrne, 2020). The other category includes errors attributable to architectural and algorithmic choices made when designing the model. Among this category are truncated translations due to the modeling of the end-of-sequence tag (Newman et al., 2020) or large beam sizes during decoding (Murray and Chiang, 2018), hallucinations due to exposure bias and domain shift (Wang and Sennrich, 2020), and promotion of frequent patterns over infrequent – but more desirable – ones as a result of the loss minimization process (Vanmassenhove et al., 2021).

The goal of this thesis is to illuminate some of the understudied pathologies of contemporary NMT systems, discuss their potential causes, and propose workable and empirically verified solutions where feasible. Additionally, initial investigations are conducted into whether translation challenges that were found to elicit undesirable behaviour in traditional NMT models are equally problematic for the recently popu-

[2]Translations collected on 6.2.2024; EN = English, DE = German.

larized large language models (LLMs). The latter have been shown to perform comparably to state-of-the-art (SOTA) NMT systems on a variety of translation directions (Hendy et al., 2023) and, as such, appear to present a valid alternative to conventional translation models. In the following section, the content of the thesis is discussed in greater detail, summarizing its main points and stating its intended contributions to the field of natural language processing (NLP).

1.1 Structure of Thesis

The central contribution of this thesis is the **principled discussion of several lexical phenomena shown to pose a substantial challenge to contemporary NMT systems**, with a particular focus on the highly prolific transformer architecture (Vaswani et al., 2017). Two of the projects discussed herein investigate the limitations of translation models by uncovering their over-reliance on spurious correlations found in the training data (Emelin et al., 2020; Emelin and Sennrich, 2021), while the third proposes a simple architectural modification intended to improve lexical reasoning (Emelin et al., 2019).

Chapter 2: Background: The study of lexical understanding in the context of machine translation and beyond received limited but consistent attention in recent years from the academic community. The literature review provided in this chapter aims to summarize findings and developments that are of relevance to studies discussed in subsequent chapters and is intended to equip the reader with the prerequisite background knowledge. As such, the review includes summaries of works on WSD in the context of machine translation, commonsense CoR, transformer models that have risen to uncontested prominence in NLP, as well as LLMs responsible for SOTA results on a variety of language understanding and generation tasks.

Chapter 3: Over-fitting on Dataset Artifacts Facilitates Word Sense Disambiguation Errors We propose that one prominent category of translation errors observed in the output of NMT models – the incorrect disambiguation of polysemous nouns – is the result of model behaviour being largely guided by correlations between individual word senses and context words that they frequently co-occur with during training. As part of our investigation into the validity of this assumption, we introduce a new statistical measure of the associative strength between any particular sense of an ambiguous word and the sentence context surrounding it, which we denote as a model's *disambiguation bias* (DB). We examine several strategies for calculating DB based on different estimates of lexical correlation in the training data, showing their respective efficacy for the prediction of WSD errors. Subsequently, we utilize our insights to craft targeted, model-agnostic adversarial attacks that successfully exploit a translation model's DB. Within the broader context of this thesis, this work sheds light onto how the assumptions of distributional semantics that underpin the learning of lexical representations in NMT can give rise to errors due to properties of the training data.

The chapter is complemented by a more recently conducted investigation into whether LLMs are similarly susceptible to DBs and exhibit notable vulnerabilities towards adversarial WSD attacks. The reported findings suggest that LLMs offer a promising path towards translations that are more accurate and robust to WSD errors, due to their superior lexical understanding capabilities.

The uncovered flaws in the lexical disambiguation step are a direct consequence of the assumptions guiding the learning process. Complementing this finding, the work presented in Chapter 4 lays out further causes for undesirable translation model behavior.

Chapter 4: Deficits in Text Understanding Lead to Coreference Resolution Errors The second type of translation errors examined in this thesis arises from a model's failure to correctly establish the coreference between a pronoun and its intended antecedent, resulting in the incorrect translation of the pronoun into the target language. A particularly challenging CoR setting are Winograd schemas (Winograd, 1972), since they require the comprehender to apply commonsense reasoning (Sakaguchi et al., 2020) in order to correctly identify the referent of an ambiguous pronoun among two possible alternatives. While this task poses little difficulty for humans, it has been found to confound computational models (Levesque et al., 2012) . We test NMT models in this setting on the assumption that their aptitude for solving difficult instances of a task (such as CoR) is a useful indicator for their overall robustness and worst case performance on said task. To enable this evaluation, we construct a challenge set of English schemas paired with their translations into three typologically distinct languages. Importantly, each target language enforces agreement between the grammatical gender of a pronoun and its referent. Models are evaluated based on whether their translations of schema samples exhibit agreement between the translation of the pronoun and its true referent. Across all examined translation directions, we find NMT models to perform close to chance. Expanding upon the conclusions drawn in Chapter 3, we trace likely reasons for the observed errors to a combination of faulty model behavior and inherent properties of the training data. More specifically, we find that NMT models exhibit gender bias, frequently favouring masculine referent candidates over their alternatives, and fail to identify words that provide the necessary context for the disambiguation of the pronoun referent. By fine-tuning models on a small amount of unbiased data and modifying their training objective, both shortcomings can be mitigated.

Building upon this evaluation of NMT models, a series of more recent experiments examines the ability of a popular SOTA LLM to solve Winograd schemas, both across different languages as well as when assessing the adequacy of contrasting translations. Here, too, LLMs show promise as a new generation of translation engines by outperforming traditional NMT models provided suitable prompting. Even so, however, they remain far below human parity. Thus, translation that presupposes commonsense reasoning remains a difficult challenge.

Taken together with the findings presented in Chapter 3, insights from this study suggest that contemporary NMT architectures fall short of effectively utilizing lexical information throughout the translation process. In Chapter 5, we consider a straightforward modification to the default transformer architecture (Vaswani et al., 2017) in order to make embedding features more accessible and improve the leveraging of lexical context, thereby aiding lexical reasoning.

Chapter 5: Better Access to Embedding Features in Transformers Improves Translation Quality Within the workflow of transformer models, features in the embedding layer encode local information pertaining to a (sub-)word's meaning in isolation from any textual context. Through repeated application of the self-attention mechanism, these initial features are enriched with relevant contextual information that may have a disambiguating effect for polysemous and otherwise ambiguous words. Intuitively, improving the extent to which an NMT model can access and utilize both local and contextual information – represented by corresponding latent feature tensors – could alleviate some of the observed word-level translation errors and improve overall translation quality. In this work, we propose a simple extension to the standard transformer encoder-decoder model in the form of gated residual connections drawn between the embedding layer and any of the subsequent layers. These *lexical shortcut connections* can be deployed in the encoder of an NMT model, in its decoder, or both. We evaluate the resultant architecture on a variety of translation directions, finding that it outperforms the standard transformer on automatic metrics of translation quality, especially for smaller models that are more limited in their representational capacity due to the low dimensionality of their hidden states. Similarly, we observe notable improvements in WSD accuracy. We furthermore verify that lexical shortcuts enable the transformer to incorporate more contextual information into its hidden states as evidenced by diagnostic classifiers (Belinkov et al., 2017b).

To establish the usefulness of our modifications beyond the experiments included in the original publication, we additionally benchmark the lexical shortcuts transformer (LST) on the WSD and CoR benchmarks introduced in Chapters 3 and 4, comparing its performance with that of the default transformer model. We discuss the results, finding that they generally confirm the efficacy of shortcut connections for improved lexical reasoning.

Chapter 6: Conclusion In the concluding chapter, we distill the main findings of the thesis and discuss their implications for the current state of machine translation research as well as their limitations. Lastly, we consider potential directions for future research that could address some of the flaws of current NMT models identified here.

1.2 Contributions

The primary contributions made in this thesis to the field of machine translation and, more generally, NLP are as follows:

- The presented work systematically **uncovers and quantifies some of the under-explored weaknesses and failure modes** of contemporary NMT models, with a focus on WSD and CoR
  - *Disambiguation bias* is introduced as a novel concept and found to be highly relevant to the prediction and elicitation of WSD errors in automatic translation
  - NMT models are shown to be incapable of applying commonsense reasoning to CoR as part of the translation step, instead relying on undesirable data artifacts
- **Novel resources in form of multiple challenge sets** are constructed, verified, and used to evaluate contemporary NMT models

  - Challenge sets can be used to assess WSD and CoR capabilities of models across different domains and languages, offering a comprehensive coverage
  - All challenge sets are constructed semi-automatically and can, therefore, be expanded and modified to meet the requirements of subsequent studies

- A simple but effective **extension of the transformer architecture** is proposed with the aim to improve lexical understanding capabilities of NMT models

  - The extension demonstrably improves translation quality and WSD accuracy compared to the unmodified transformer

- A thorough **evaluation of LLMs' ability to perform cross-lingual WSD and CoR** is conducted that mirrors the presented NMT studies

  - Findings suggest that LLMs generally outperform NMT models while nonetheless failing to demonstrate human-like competency

Chapter 2

Background: Lexical Understanding in NMT

> There is no greater impediment to the advancement of knowledge than the ambiguity of words.
>
> Thomas Reid, *Thomas Reid's Inquiry and Essays*

This section outlines the necessary scientific background for the subsequent chapters, presenting relevant work that motivated the hypotheses explored in this thesis and paved the way for the conducted studies.

2.1 WSD and Biases in Neural Machine Translation

Polysemous terms represent a long-standing challenge for NMT. As such, various WSD strategies have been employed in the past to improve the disambiguation accuracy of translation systems. As previously touched upon, the objective of WSD within machine translation is to disambiguate polysemous source words by leveraging the available source context, to ensure that the produced translations adequately capture the intended meaning of the input text in the target language. More generally, WSD is central to the task of language understanding and, as such, integral to a wide range of NLP downstream tasks beyond machine translation, see e.g. (Bovi et al., 2015; Shimura et al., 2019).

The study of WSD in NMT, which represents one of the foci of this thesis, can be roughly subdivided into two categories – the evaluation of models' WSD capabilities as well as their improvement. Within the former category, past investigations have sought to understand the disambiguation process by analysing the internal representations of NMT models (Marvin and Koehn, 2018; Tang et al., 2019), or to quantify their WSD efficacy through the use of challenge sets. Such evaluation datasets are constructed either primarily through manual efforts as in (Lefever and Hoste, 2013; Campolungo et al., 2022a; Futeral et al., 2022) or by relying on automated procedures as in (Vickrey et al., 2005; Rios et al., 2017; Raganato et al., 2019), and usually task the evaluated model with scoring a pair of contrastive translations that differ only in the word sense assigned to a specific polysemous source word as indicated by its target language translation. In contrast, works belonging to the latter category have either relied on supervised learning with sense-annotated data (Campolungo et al., 2022b), or leveraged external lexical knowledge resources such as WordNet (Pu et al., 2018) as well as multi-modal data sources (Futeral et al., 2022) and large language models (LLMs)(Iyer et al., 2023) in order to facilitate the correctness of WSD during translation. To our knowledge, no study prior to (Emelin et al., 2020) – discussed in Chapter 3 – has examined the interaction between training data artifacts and WSD performance.

Dataset artifacts, on the other hand, have been shown to enable models to make correct predictions based on incorrect or insufficient information (McCoy et al., 2019; Gururangan et al., 2018) by over-relying on spurious correlations present in the training data. Within NMT, models were found to exhibit gender-bias that reinforces harmful stereotypes (Vanmassenhove et al., 2018; Stanovsky et al., 2019; Sarti et al., 2023). As a response, strategies have been proposed for de-biasing the training data (Li and Vasconcelos, 2019; Le Bras et al., 2020), as well as for making models more robust to data biases through adversarial training (Belinkov et al., 2019).

Adversarial attacks have previously been extended as an effective model analysis tool from vision to language tasks (Samanta and Mehta, 2017; Alzantot et al., 2018; Glockner et al., 2018; Zhang et al., 2019), including NMT (Cheng et al., 2019, 2020), where the focus so far has been on strategies that require direct access to the loss gradient or output distribution of the victim model. Recent surveys suggested that state-of-the-art attacks often yield ungrammatical and meaning-destroying samples, thus diminishing their usefulness for the evaluation of model robustness (Michel et al., 2019; Morris et al., 2020). More specifically, targeted attacks on WSD abilities of translation models have remained under-explored.

2.2 CoR and Commonsense Reasoning in NLP

Closely related to WSD due to addressing ambiguity in natural language, the study of coreference has a longstanding tradition in machine translation. In contrast to WSD, CoR aims to identify the meaning of ambiguous words – such as pronouns – by detecting mentions of the same entity, object, or event that the ambiguous word refers to within the available text context, rather than by selecting its intended meaning from an inventory of possible word senses. Likewise, past research into CoR focused on either establishing the CoR capabilities and limitations of translation models, or on improving CoR performance either by incorporating auxiliary linguistic information into the translation process, or by improving the extent to which the context surrounding ambiguous words influences their translation.

As with WSD, CoR evaluation commonly utilizes contrastive test sets, due to the insufficient sensitivity of popular translation quality metrics such as BLEU (Papineni et al., 2002), since the impact of CoR errors is commonly confined to individual words within any given sentence (Müller et al., 2018). Specifically, such benchmarks assess CoR accuracy according to the likelihood assigned by the evaluated model to a pair of contrasting translations – one where coreference has been resolved correctly resulting in the intended translation of a particular ambiguous source word, and one where the word in question is translated incorrectly due to failed CoR. Several such benchmarks had been proposed in the past, including (Bawden et al., 2018; Guillou and Hardmeier, 2016; Müller et al., 2018; Stojanovski et al., 2020). Here, too, dataset construction methods differ, ranging from manual (Bawden et al., 2018) to largely automated (Stojanovski et al., 2020). Among those, (Stojanovski et al., 2020) bears most relevance to the work presented in Chapter 4 of this thesis. While their benchmark includes some examples of coreference relations that require commonsense knowledge to be resolved correctly, the corresponding test samples are constructed from a fixed set of templates and remain limited to the EN-DE translation direction. Such limitations do not apply to the *Wino-X* benchmark introduced in in Chapter 4.

Among strategies proposed to improve CoR in NMT, the currently dominant methodology involves incorporating document-level context of the source sentence – and, in some cases, the target sentence – into the translation process by utilizing mechanisms such as additional encoder and attention layers (Jean et al., 2017; Bawden et al., 2018; Werlen et al., 2018; Stojanovski and Fraser, 2019; Herold and Ney, 2023) and two-pass translation (Voita et al., 2019b), or by simply concatenating the source context with the sentence to be translated (Tiedemann and Scherrer, 2017). The related field of document-level machine translation remains an active research area and its advancement can be expected to lead to more effective and robust CoR in translation models in the future.

While the success of CoR as a translation step manifests itself in whether individual, ambiguous source words are disambiguated correctly or not within the target language output, the disambiguation choice is informed by the context in which ambiguous words occur, both within and outside the surrounding text (e.g. if the decision is informed by *commonsense* or *world* knowledge). As such, lexical understanding alone is not sufficient for CoR to be carried out – semantic and pragmatic reasoning on sentence-, discourse-, and extra-textual level undoubtedly play an essential role, as well. The framing of CoR as a competency predominantly linked to lexical understanding as assumed in this thesis is therefore a convenient shortcut that foregrounds its outcome – i.e. the correct or incorrect choice of a specific target *word* – as a visible, empirically verifiable reflection of the combined cognitive processes informing the disambiguation step. Moreover, since the correctness and appropriateness of lexical choices is one of the most immediately apparent indicators of the overall correctness of translations, and the contributions of this thesis are centered around the limitations of NMT systems with respect to the adequacy of their outputs, viewing CoR through the lens of lexical understanding is a conscious, natural choice. However, we stress that this is but one of the many aspects intrinsic to this complex, multifaceted task.

Winograd schemas, in turn, have been widely adopted in recent years for the study of pronominal coreference and commonsense reasoning (Kocijan et al., 2020). Several datasets had been proposed, differing in whether schemas are authored by experts (Levesque et al., 2012; Wang et al., 2019) or composed by crowd-workers (Isaak and Michael, 2019; Sakaguchi et al., 2020). Crucially, the majority of such resources is in English, with the notable exception of (Amsili and Seminck, 2017; Bernard and Han, 2020; Melo et al., 2019) (each contain a few hundred examples).

Finally, while cross-lingual transfer in multi-lingual language models (MLLMs) has received much attention in the past (Conneau et al., 2020, 2018; Hu et al., 2020; Liang et al., 2020), research on commonsense reasoning in multiple languages remains limited, with (He et al., 2020) being the only relevant machine translation study known to us. Outside of the translation context, (Lin et al., 2021) examine whether MLLMs can perform multilingual commonsense reasoning on tasks unrelated to Winograd schemas.

2.3 Transformer Models and Lexical Representations

The transformer architecture has been widely adopted by the NLP community, becoming the de-facto workhorse for many applications, especially NMT. First introduced in (Vaswani et al., 2017), it employs an attention mechanism to model the source and target contexts during translation, implicitly learning to align them. Parallel training and scalability make it a popular choice for translation tasks with access to large quantities of data where its use has yielded marked improvements in translation quality.

Within recent literature, several strategies for altering the flow of information within the transformer have been proposed, including adaptive model depth (Dehghani et al., 2018), layer-wise transparent attention (Bapna et al., 2018), and dense inter-layer connections (Dou et al., 2018). The investigation presented in Chapter 5 of this thesis bears strongest resemblance to the latter work, by introducing additional connectivity to the model. However, rather than establishing new connections between layers indiscriminately, it explicitly seeks to facilitate the accessibility of lexical features across network layers.

Likewise, studies investigating the role of lexical features in NMT have informed parts of Chapter 5. Among them, (Nguyen and Chiang, 2018) note that improving accessibility of source words in the decoder benefits translation quality for low-resource settings. In a similar vein, (Wu et al., 2018) attend both encoder hidden states and source embeddings as part of decoder-to-encoder attention, while (Kuang et al., 2018) provide the decoder-to-encoder attention mechanism with improved access to source word representations. Research concerning itself with the analysis of the internal dynamics and learned representations within deep neural networks (Karpathy et al., 2015; Qian et al., 2016; Shi et al., 2016; Bisazza and Tump, 2018) was equally relevant to the chapter. Here, (Belinkov et al., 2017a) and (Belinkov et al., 2017b) serve as primary points of reference by offering a thorough and principled investigation of the extent to which neural translation models are capable of learning linguistic properties from raw text. However, it must be noted that the effectiveness of diagnostic classifiers, e.g. as used in (Belinkov et al., 2017b), has been called into question by more recent studies (Dhar and Bisazza, 2020), published after (Emelin et al., 2019).

The view of the transformer as a model learning to refine input representations through the repeated application of attention is consistent with the iterative estimation paradigm introduced in (Greff et al., 2016). According to this interpretation, given a stack of connected layers sharing the same dimensionality and interlinked through highway or residual connections, the initial layer generates a rough version of the stack's final output, which is iteratively refined by successive layers, e.g. by enriching localized features with information drawn from the surrounding context. This analysis is supported and further extended by the findings presented in Chapter 5 which indicate that individual layers can learn novel information provided there is sufficient representational capacity to do so.

2.4 Large Language Models

Recent years saw the emergence of LLMs as one of the focal points of NLP research, owing to their effectiveness on a variety of diverse language comprehension and generation problems (Zhao et al., 2023). The core defining characteristics of LLMs are their large parameter count, usually in the billions, as well as the enormous quantities of data used in their unsupervised (pre-)training, which can include up to trillions of tokens for text-based models (Hoffmann et al., 2022). As a result of their unprecedented scale, LLMs exhibit powerful emergent capabilities (Wei et al., 2022a) that can be invoked via methods such as prompting (Brown et al., 2020) to solve complex tasks. Some of the representative models falling within this category include the GPT family (Radford et al., 2018, 2019; Brown et al., 2020; OpenAI, 2023), PaLM (Chowdhery et al., 2022; Anil et al., 2023), Llama (Touvron et al., 2023), and Pythia (Biderman et al., 2023). For most LLMs, the majority of training data originates from sources authored in English (e.g. websites, books, or academic publications), with other languages making up a far smaller fraction of training samples by comparison (Zhao et al., 2023). Nonetheless, LLMs have been found to exhibit strong multilingual capabilities even if not explicitly optimized for this purpose, performing well on tasks such as multilingual natural language generation and multilingual question answering in few-shot and fine-tuning settings (Chowdhery et al., 2022).

One of the most studied and leveraged emergent abilities of LLMs is *in-context learning*, i.e. the models' capacity to generate appropriate responses for previously unseen tasks after being provided with a task instruction in natural language and several demonstrations of the expected model behavior (Brown et al., 2020). Since the formulation of the instruction as well as the properties of the selected demonstrations have been found to notably impact model performance (Liu et al., 2023), much research has been dedicated to developing strategies for optimal prompt construction, commonly referred to as *prompt engineering* (White et al., 2023). One notable prompting strategy that has enabled LLMs to better handle complex reasoning tasks is Chain-of-Thought (CoT) prompting (Wei et al., 2022b) where the prompt incorporates intermediate reasoning steps that directly motivate and inform the expected model output. Since both WSD and CoR can require in-depth linguistic and commonsense reasoning within the translation setting, CoR presents a potentially necessary tool for obtaining accurate translations from LLMs that conventional NMT systems may be unable to yield.

Despite their impressive performance on a variety of NLP benchmarks, LLMs adoption in machine translation has remained gradual. While several LLM have been developed specifically with multilingual applications in mind, including such models as mT5 (Xue et al., 2021) and BLOOM (Scao et al., 2022), they have not yet found widespread adaptation in translation applications. Among publications exploring the use of LLMs for NMT (e.g. (Guerreiro et al., 2023)), (Hendy et al., 2023) and (Zhang et al., 2023) are of particular relevance to the experiments conducted within the scope of this thesis. Specifically, (Hendy et al., 2023) demonstrate that the GPT-3 and Chat-GPT LLMs perform on-par with the winning WMT22[1] (Kocmi et al., 2022) systems with respect to the quality of the generated translations for high-resource directions, but fare less well on low-resource languages underrepresented in their training data. On the other hand, (Zhang et al., 2023) investigate the efficacy of different prompting strategies for the translation task, finding that relatively simple English prompts that incorporate (up to ten) high-quality demonstrations work best empirically. LLM studies conducted in this thesis are directly motivated by these findings in both model selection and prompt design.

[1]Seventh Conference on Machine Translation: https://www.statmt.org/wmt22.

Chapter 3

Dataset Artifacts Inform Word Sense Disambiguation Errors

Light boats sail swift, though greater hulks draw deep.

William Shakespeare, *King Henry VI*

Abstract: Word sense disambiguation is a well-known source of translation errors in NMT. We posit that some of the incorrect disambiguation choices are due to models' **over-reliance on dataset artifacts found in training data**, specifically superficial word co-occurrences – rather than a deeper understanding of the source text – which becomes particularly evident in sentence contexts that contain contrasting lexical disambiguation cues. We introduce a **method for the prediction of disambiguation errors** based on statistical data properties, demonstrating its effectiveness across several domains and model types. Moreover, we develop a simple **adversarial attack strategy** that minimally perturbs sentences in order to elicit disambiguation errors to further probe the robustness of translation models. Our findings indicate that disambiguation robustness varies substantially between domains and that different models trained on the same data are vulnerable to different attacks.[1]

[1]This section is based on work previously published at EMNLP 2020 (Emelin et al., 2020). Experimental codebase is available at http://github.com/demelin/detecting_wsd_biases_for_nmt.

3.1 Introduction

Consider the sentence *John met his wife in the hot spring of 1988*. In this context, the polysemous term *spring* unambiguously refers to the season of a specific year. Its appropriate translation into German would therefore be *Frühling* (the season), rather than one of its alternative senses, such as *Quelle* (the source of a stream). To contemporary machine translation systems, however, this sentence presents a non-trivial challenge, with Google Translate (GT) producing the following translation: *John traf seine Frau in der heißen Quelle von 1988*[2].

Prior studies have indicated that neural machine translation (NMT) models rely heavily on source sentence information when resolving lexical ambiguity (Tang et al., 2019). This suggests that the combined source contexts in which a specific sense of an ambiguous term occurs in the training data greatly inform the models' disambiguation decisions. Thus, a stronger correlation between the English collocation *hot spring* and the German translation *Quelle*, as opposed to *Frühling*, in the training corpus may explain this disambiguation error. Indeed, *John met his wife in the spring of 1988* is translated correctly by GT.

We propose that our motivating example is representative of an undesirable behaviour NMT systems have yet to overcome when performing word sense disambiguation (WSD). Specifically, we hypothesize that translation models are unable to reliably identify and utilise informative disambiguation triggers in source sentences containing multiple potential candidates, resulting in incorrect diambiguation of polysemous source terms. We attribute this primarily to the models' reliance on lexical correlations observed in the training data and inability to generalize beyond them. As a result, disambiguation errors are likely to arise when an ambiguous word co-occurs with words that are strongly correlated in the training corpus with a sense that differs from the reference.

To test our hypothesis, we evaluate whether dataset artifacts are predictive of disambiguation decisions made in NMT. First, given an ambiguous term, we define a strategy for quantifying how much its context biases NMT models towards its different target senses, based on statistical patterns in the training data. We validate our approach by examining correlations between this bias measure and WSD errors made by baseline models. Furthermore, we investigate whether such biases can be exploited for the generation of minimally-perturbed adversarial samples that trigger disambigua-

[2] Last verified to be the case on 7.8.2023.

tion errors. Our method does not require access to gradient information nor the score distribution of the decoder, generates samples that do not significantly diverge from the training domain, and comes with a clearly-defined notion of attack success and failure.

The main contributions of this study are:

1. We present **evidence for the misleading reliance of NMT systems on lexical correlations** when incorrectly translating polysemous source words.

2. We propose a **method for quantifying WSD biases** that can predict disambiguation errors.

3. We leverage data artifacts for the creation of **adversarial samples** that facilitate WSD errors.

3.2 Can WSD errors be predicted?

To evaluate whether WSD errors can be effectively predicted, we first propose a method for measuring the bias of sentence contexts towards different senses of polysemous words, based on lexical co-occurrence statistics of the training distribution. We restrict our investigation to English→German, although the presented findings can be assumed to be language-agnostic. To bolster the robustness of our results, we conduct experiments in two domains - movie subtitles characterized by casual language use, and the more formal news domain. For the former, we use the OpenSubtitles2018 (OS18) (Lison et al., 2019) corpus[3], whereas the latter is represented by data made available for the news translation task of the Fourth Conference on Machine Translation (WMT19)[4]

The WMT19 data is obtained by concatenating the Europarl v9, Common Crawl, and News Commentary v14 parallel corpora. Basic data cleaning is performed for both domains, which includes removal of pairs containing sentences classified by langid[5] as neither German or English and pairs with a source-to-target sentence length ratio exceeding 2. We create development and testing splits for the OS18 domain by removing 10K sentence pairs from the full, shuffled corpus in each case. For each domain, we additionally hold out 20% of pairs to be used for the extraction of test pairs containing homographs, as described in section 3.2.2. Final statistics for the OS18 domain are reported in table 3.1, and in 3.2 for the WMT19 domain.

[3] http://opus.nlpl.eu
[4] http://statmt.org/wmt19
[5] http://github.com/saffsd/langid.py

Statistic	train	dev	test	held-out
# sentences	14,993,062	10,000	10,000	3,751,765
# words (EN)	106,873,835	71,719	71,332	26,763,351
# words/sentence (EN)	7.13	7.17	7.13	7.13
# words (DE)	100,248,893	67,185	66,799	25,094,166
# words/sentence (DE)	6.69	6,71	6.68	6.69

Table 3.1: Corpus statistics for the OS18 domain.

Statistic	train	dev (test18)	test14	test19	held-out
# sentences	4,861,743	2,998	3,003	1,997	1,215,435
# words (EN)	100,271,426	58,628	59,325	42034	25,057,036
# words/sentence (EN)	20.62	19.56	19.76	21.05	20.62
# words (DE)	93,900,343	54,933	54,865	42,087	23,467,086
# words/sentence (DE)	19.31	18.32	18.27	21.08	19.31

Table 3.2: Corpus statistics for the WMT19 domain.

Each dataset is subsequently tokenized and truecased using Moses (Koehn et al., 2007) scripts[6]. For model training and evaluation, we additionally learn and apply BPE codes (Sennrich et al., 2016) to the data using the subword-NMT implementation[7], with 32k merge operations and the vocabulary threshold set to 50.

3.2.1 Quantifying disambiguation biases

An evaluation of cross-lingual WSD errors presupposes the availability of certain resources, including a list of ambiguous words, a lexicon containing their possible translations, and a set of parallel sentences serving as a disambiguation benchmark.

Resource collection

Since lexical ambiguity is a pervasive feature of natural language, we limit our study to **homographs** - polysemous words that share their written form but have multiple,

[6]http://github.com/moses-smt/mosesdecoder
[7]http://github.com/rsennrich/subword-nmt

unrelated meanings. We further restrict the set of English homographs to nouns that are translated as distinct German nouns, so as to confidently identify disambiguation errors, while minimizing the models' ability to disambiguate based on syntactic cues. English homographs are collected from web resources[8], excluding those that do not satisfy the above criteria. The full list of homographs used in our experiments is as follows: *anchor*, *arm*, *band*, *bank*, *balance*, *bar*, *barrel*, *bark*, *bass*, *bat*, *battery*, *beam*, *board*, *bolt*, *boot*, *bow*, *brace*, *break*, *bug*, *butt*, *cabinet*, *capital*, *case*, *cast*, *chair*, *change*, *charge*, *chest*, *chip*, *clip*, *club*, *cock*, *counter*, *crane*, *cycle*, *date*, *deck*, *drill*, *drop*, *fall*, *fan*, *file*, *film*, *flat*, *fly*, *gum*, *hoe*, *hood*, *jam*, *jumper*, *lap*, *lead*, *letter*, *lock*, *mail*, *match*, *mine*, *mint*, *mold*, *mole*, *mortar*, *move*, *nail*, *note*, *offense*, *organ*, *pack*, *palm*, *pick*, *pitch*, *pitcher*, *plaster*, *plate*, *plot*, *pot*, *present*, *punch*, *quarter*, *race*, *racket*, *record*, *ruler*, *seal*, *sewer*, *scale*, *snare*, *spirit*, *spot*, *spring*, *staff*, *stock*, *subject*, *tank*, *tear*, *term*, *tie*, *toast*, *trunk*, *tube*, *vacuum*, *watch*.

Homograph	**Sense 1**	**Sense 2**	**Sense 3**
bat	*Chiroptera*, *Fledertier*, *Handflügler*, *Fledermaus*, *Flattertier*	*Schlagstock*, *Schlagholz*, *Baseballschläger*, *Baseballkeule*, *Schläger*	-
case	*Karton*, *Kiste*, *Päckchen*, *Packung*, *Schachtel*, *Kasten*, *Behälter*, *Box*	*Fall*, *Zustand*, *Sache*, *Gegebenheit*, *Lage*, *Kontext*, *Umstand*, *Status*, *Sachverhalt*, *Stand*, *Situation*	*Prozess*, *Gerichtsverfahren*, *Fall*, *Gerichtsverhandlung*, *Sache*, *Prozeß*, *Rechtsstreit*, *Ermittlung*, *Antrag*, *Rechtsfall*, *Gerichtsfall*, *Klage*, *Verhör*, *Rechtssache*
letter	*Sendschreiben*, *Papierbrief*, *Musterbrief*, *Anschreiben*, *Post*, *Schreiben*, *Brief*	*Buchstabe*, *Großbuchstabe*, *Charakter*, *Letter*, *Kleinbuchstabe*, *Zeichen*	-
spring	*Ringfeder*, *Spiralfeder*, *Sprungfeder*, *Feder*, *Tellerfeder*, *Federung*, *Gummifeder*	*Frühling*, *Lenz*, *Frühjahr*	*Quelle*, *Brunnen*, *Quell*, *Wasserquelle*
vacuum	*Vakuum*, *Nichts*, *Unterdruck*, *Leerraum*, *Leere*, *Luftleere*	*Industriestaubsauger*, *Staubsauger*, *Handstaubsauger*, *Teppichkehrer*, *Bodenstaubsauger*, *Allessauger*, *Sauger*, *Kesselsauger*	-

Table 3.3: Non-exhaustive examples of homograph-specific sense clusters.

[8]http://7esl.com/homographs, http://en.wikipedia.org/wiki/List_of_English_homographs

We next compile a parallel lexicon of homograph translations, prioritizing a high coverage of all possible senses. Similar to (Raganato et al., 2019), we obtain sense-specific translations from cross-lingual BabelNet (Navigli and Ponzetto, 2010) synsets. Since BabelNet entries vary in their granularity, we iteratively merge related synsets as long as they have at least three German translations in common or share at least one definition.[9] This leaves us with multiple sense clusters of semantically related German translations per homograph. To further improve the quality of the lexicon, we manually clean and extend each homograph entry to address the noise inherent in BabelNet and its incomplete coverage. Table 3.3 lists some of the identified sense clusters for several homographs. All homographs used in our experiments have at least two sense clusters associated with them.

In order to identify sentence contexts specific to each homograph sense, parallel sentences containing known homographs are extracted from the training corpora in both domains. We lemmatize homographs, their senses, and all sentence pairs using *spaCy* (Honnibal and Montani, 2017) to improve the extraction recall. Homographs are further required to be aligned with their target senses according to alignments learned with *fast_align* (Dyer et al., 2013). Each extracted pair is assigned to one homograph sense cluster based on its reference homograph translation. Pairs containing homograph senses assigned to multiple clusters are ignored, as disambiguation errors are impossible to detect in such cases.

Bias measures

It can be reasonably assumed that context words co-occurring with homographs in a corpus of natural text are more strongly associated with some of their senses than others. Words that are strongly correlated with a specific sense may therefore bias models towards the corresponding translation at test time. We refer to any source word that co-occurs with a homograph as an *attractor* associated with the sense cluster of the homograph's translation. Similarly, we denote the degree of an attractor's association with a sense cluster as its *disambiguation bias* (DB) towards that cluster. Table 3.4 lists the most frequent attractors identified for the different senses of the homograph *spring* in the OS18 training set.

Intuitively, if an NMT model disproportionately relies on simple surface-level correlations when resolving lexical ambiguity, it is more likely to make WSD errors when

[9] A manual inspection found the clusters to be meaningful.

season	*water source*	*device*
summer	hot	like
winter	water	back
come	find	thing

Table 3.4: Examples of attractors for *spring*.

translating sentences that contain strong attractors towards a wrong sense. To test this, we collect attractors from the extracted parallel sentences, quantifying their DB using two metrics: Raw co-occurrence frequency (FREQ) and positive point-wise mutual information (PPMI) between attractors and homograph senses. FREQ is defined in Eqn. 3.1, while Eqn. 3.2 describes PPMI, with $w \in V$ denoting an attractor term in the source vocabulary[10], and $sc \in SC$ denoting a sense cluster in the set of sense clusters assigned to a homograph. For PPMI, $P(w_i, sc_j)$, $P(w_i)$, and $P(sc_j)$ are estimated via relative frequencies of (co-)occurrences in training pairs.

$$FREQ(w_i, sc_j) = Count(w_i, sc_j) \tag{3.1}$$

$$PPMI(w_i, sc_j) = max(\frac{P(w_i, sc_j)}{P(w_i)P(sc_j)}, 0) \tag{3.2}$$

The disambiguation bias associated with the entire context of a homograph is obtained by averaging sense-specific bias values $DB(w_i, sc_j)$ of all attractors in the source sentence $S = \{w_1, w_2, ..., w_{|S|}\}$, as formalized in Eqn. 3.3, where $DB(w_i, sc_j)$ can be either $FREQ(w_i, sc_j)$ or $PPMI(w_i, sc_j)$. Context words that are not known attractors of sc_j, i.e. that have not been observed in the training corpus as accompanying an occurrence of the homograph corresponding to sc_j, are assigned a disambiguation bias value of 0.

$$DB(S, sc_j) = \frac{1}{|S|} \sum_{i=1}^{|S|} DB(w_i, sc_j) \tag{3.3}$$

As a result, sentences containing a greater number of strong attractors are assigned a higher bias score.

3.2.2 Probing NMT models

To evaluate the extent to which sentence-level disambiguation bias is predictive of WSD errors made by NMT systems, we train baseline translation models for both

[10]We consider any word that co-occurs with a homograph in the training corpus as an attractor of the homograph's specific sense cluster, except for the homograph itself which is not regarded as an attractor for any of its known sense clusters.

domains. The baselines include Transformer (Vaswani et al., 2017), LSTM (Luong et al., 2015), and convolutional Seq-to-Seq (ConvS2S) (Gehring et al., 2017) models of comparable size. Table 3.5 provides implementation and training details for each architecture. Same settings are used for training identical model types in different domains. We use standard fairseq[11] (Ott et al., 2019) implementations for all model types and train them on NVIDIA 1080ti or NVIDIA 2080ti GPUs. Model translations are obtained by averaging the final 5 model checkpoints and decoding using beam search with beam size 5.

Parameter	**Transformer**	**LSTM**	**ConvS2S**
batch size (subwords)	24,576	4,096	4,096
# total updates	100,000	600,000	600,000
# warm-up updates	4,000	-	-
# updates between checkpoints	1,000	4,000	4,000
# epochs between validations	1	1	1
optimizer	Adam	Adam	Adam
Adam betas	0.9, 0.98	0.9, 0.98	0.9, 0.98
learning rate	scheduled (*inverse_sqrt*)	0.0002 (+ decay)	0.0003 (+ decay)
# total parameters (OS18)	60,641,280	59,819,008	64,548,328
# total parameters (WMT19)	61,714,432	60,892,160	66,696,728
embedding size	512	512	512
Tied embeddings	Yes	Yes	Yes
hidden size	2,048	512	512
# encoder layers	6	5 (bidirectional)	8
# decoder layers	6	5	8
kernel size	-	-	3
dropout	0.1	0.2	0.2
label smoothing	0.1	0.1	0.1

Table 3.5: Training settings and model hyperparameters.

SacreBLEU (Post, 2018) scores given in Table 3.6 indicate that all models are

[11] http://github.com/pytorch/fairseq

reasonably competent, being comparable to results previously reported for these test sets. e.g. by (Gehring et al., 2017; Vaswani et al., 2017).

		WMT	
Architecture	**OS18 test**	**test 2014**	**test 2019**
Transformer	29.7	27.5	38.2
LSTM	27.7	24.1	34.3
ConvS2S	27.7	23.5	32.5

Table 3.6: EN-DE translation performance (BLEU).

Test sets for WSD error prediction are constructed by extracting parallel sentences from held-out, development, and test data (see tables 3.1 and 3.2 for details). The process is identical to that described in section 3.2.1, with the added exclusion of source sentences shorter than 10 tokens, as they may not provide enough context. For each source sentence, disambiguation bias values are computed according to equation 3.3, with sc_j corresponding to either the correct sense cluster ($DB_{\checkmark}$) or the incorrect sense cluster with the strongest bias ($DB_{\times}$). Additionally, we consider the difference DB_{DIFF} between $DB_{\times}$ and $DB_{\checkmark}$ which can be interpreted as the overall statistical bias in a source sentence towards an incorrect homograph translation. All bias scores are computed either using FREQ or PPMI.

We examine correlations between the proposed bias measures and WSD errors produced by the in-domain baseline models. Translations are considered to contain WSD errors if the target homograph sense does not belong to the same sense cluster as its reference translation. We check this by looking up target words aligned with source homographs according to fast_align. To estimate correlation strength we employ the ranked biserial correlation (RBC) metric[12] (Cureton, 1956) and measure statistical significance using the Mann-Whitney U (MWU) test (Mann and Whitney, 1947).

In order to compute the RBC values, test sentences are divided into two groups - one containing correctly translated source sentences and another comprised of source sentences with incorrect homograph translations. Next, all possible pairs are constructed between the two groups, pairing together each source sentence from one group with all source sentences from the other. Finally, the proportion of pairs f where the

[12] We additionally used the non-parametric generalization of the Common Language Effect Size (Ruscio, 2008) for correlation size estimation, but couldn't detect any advantages over RBC in our experimental setting.

DB score of the incorrectly translated sentence is greater than that of the correctly translated sentence is computed, as well as the proportion of pairs u where the opposite relation holds. The RBC value is then obtained according to Eqn. 3.4.

$$RBC = f - u \tag{3.4}$$

Statistical significance, on the other hand, is estimated by ranking all sentences in the test set according to their DB score in ascending order while resolving ties, and computing the U-value according to Eqn. 3.5 - 3.7, where R_1 denotes to the sum of ranks of sentences with incorrectly translated homographs and n_1 their total count, while R_2 denotes the sum of ranks of correctly translated sentences and n_2 their respective total count.

$$U = min(U_1, U_2) \tag{3.5}$$

$$U_1 = R_1 - \frac{n_1(n_1+1)}{2} \tag{3.6}$$

$$U_2 = R_2 - \frac{n_2(n_2+1)}{2} \tag{3.7}$$

To obtain the p-values, U-values are subjected to tie correction and normal approximation.[13]

Model	**FREQ**✓	**PPMI**✓	**FREQ**✗	**PPMI**✗	**FREQ$_{\text{DIFF}}$**	**PPMI$_{\text{DIFF}}$**	**Length**
OS18 Transformer	-0.532	-0.578	0.327	0.474	**0.708**	0.674	0.018
OS18 LSTM	-0.468	-0.504	0.386	0.486	**0.690**	0.630	0.008
OS18 ConvS2S	-0.477	-0.514	0.391	0.492	**0.723**	0.658	0.021
WMT19 Transformer	-0.610	-0.668	0.415	0.579	**0.687**	0.677	-0.004
WMT19 LSTM	-0.661	-0.698	0.376	0.574	**0.725**	0.708	-0.009
WMT19 ConvS2S	-0.648	-0.678	0.408	0.599	**0.731**	0.710	0.000

Table 3.7: Rank biserial correlation between disambiguation bias measures and lexical disambiguation errors.

Table 3.7 summarizes the results[14], including correlation estimates between WSD errors and source sentence length, as a proxy for disambiguation context size. Statistically significant correlations are discovered for all bias estimates based on attractors

[13] We use Python implementations of RBC and MWU provided by the pingouin library (Vallat, 2018).

[14] Positive values denote a positive correlation between bias measures and the presence of disambiguation errors in model translations, whereas negative values denote negative correlations. The magnitude of the values, meanwhile, indicates the correlations' effect size.

($p <$ 1e-5, two-sided). Moreover, the observed correlations exhibit a strong effect size (McGrath and Meyer, 2006). For all models and domains the strongest correlations are observed for DB_{DIFF} derived from simple co-occurrence counts.

A brief diversion is required to further elaborate upon our interpretation of model-specific effect size thresholds. Whether the effect size of correlations between dichotomous and quantitative variables can be considered strong depends on the size ratio between the two groups denoted by the dichotomous variable, i.e. its base rate. As the standard formulation of RBC is sensitive to the base rate, the estimated effect size decreases as the base rate becomes more extreme (see (McGrath and Meyer, 2006) for details). Applied to our experimental setting, this means that the observed correlation values are sensitive to the number of sentences containing disambiguation errors relative to the amount of those that do not. This is an undesirable property, as we are only interested in the predictive power of our quantitative variables, regardless of how often disambiguation errors are observed. Thus, we adjust the thresholds for the interpretation of correlation strength to account for WSD errors being less frequent than WSD successes overall, in analogy to (McGrath and Meyer, 2006). Doing so enables the direct comparison of correlation strength between domains and model types, as each combination of the two factors exhibits a different disambiguation success base rate.

A common practice for interpreting effect size strength that does not account for base rate inequalities is the adoption of Cohen's benchmark (Cohen, 2013), which posits that the effect size d is large if $d >= 0.8$, medium if $d >= 0.5$, and small if $d >= 0.2$. To adjust these threshold values for the observed base rates, they are converted according to Eqn. 3.8, where $p1$ and $p2$ represent the proportions of groups described by the dichotomous variable, with $p_2 = 1 - p_1$:

$$threshold = \frac{d}{\sqrt{d^2 + \frac{1}{p_1, p_2}}} \tag{3.8}$$

The adjusted effect size interpretation thresholds for WSD error correlation values as given in Table 3.7 are provided in Table 3.8.

Challenge set evaluation

To establish the predictive power of the uncovered correlations, a challenge set of 3,000 test pairs with the highest $\text{FREQ}_{\text{DIFF}}$ score is subsampled from the full WSD test pair pool in both domains. In addition, we create secondary sets of equal size (i.e. 3,000 sentence pairs) by randomly selecting pairs from each pool. As Figure 3.1

Model	small	medium	large
OS18 Transformer	0.0542	0.1344	0.2121
OS18 LSTM	0.0666	0.1647	0.2581
OS18 ConvS2S	0.0710	0.1753	0.2740
WMT19 Transformer	0.0381	0.0949	0.1508
WMT19 LSTM	0.0458	0.1138	0.1803
WMT19 ConvS2S	0.0502	0.1247	0.1971

Table 3.8: Base-rate adjusted thresholds for the interpretation of WSD error prediction correlations.

illustrates, our translation models exhibit a significantly higher WSD error rate - by a factor of up to **6.1** - on the challenge sets as compared to the randomly chosen pairs. While WSD performance is up to 96% on randomly chosen sentences, performance drops to 77–82% for the best-performing model (Transformer). This suggests that lexical association artifacts, from which the proposed disambiguation bias measure is derived, can be an effective predictor of lexical ambiguity resolution errors across model architectures and domains.

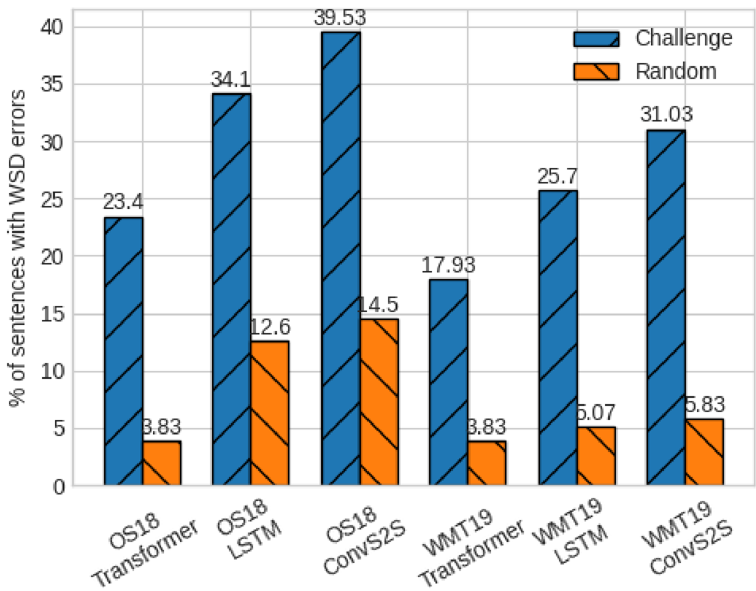

Figure 3.1: WSD errors in subsampled challenge sets.

The observed efficacy of attractor co-occurrence counts for WSD error prediction may be partially due to sense frequency effects, since more frequent senses occur in more sentence pairs, yielding more frequent attractors. NMT models are known to underperform on low-frequency senses of ambiguous terms (Rios et al., 2017), prompting us to investigate if disambiguation biases capture the same information. For this purpose, another challenge set of 3000 pairs is constructed by prioritizing pairs assigned to the rarest among each homograph's sense sets. We find that the new challenge set has a 72.63% overlap with the disambiguation bias challenge set in the OS18 domain and 64.4% overlap in the WMT19 domain. Thus, disambiguation biases appear to indeed capture some sense frequency effects, which themselves represent a dataset artifact, but also introduce novel information.

Our experimental findings indicate that translation models can be mislead by surface-level correlations found in the training data when resolving lexical ambiguity and are prone to disambiguation errors in cases where learned statistical patterns are violated. Next, we use these insights for the construction of adversarial samples that cause disambiguation errors by minimally perturbing source sentences.

3.3 Adversarial WSD attacks on NMT

Adversarial attacks probe model robustness by attempting to elicit incorrect predictions with perturbed inputs (Zhang et al., 2020). By crafting adversarial samples that explicitly target WSD capabilities of NMT models, we seek to provide further evidence for their susceptibility to dataset artifacts.

3.3.1 Generating adversarial WSD samples

Our proposed attack strategy is based on the assumption that introducing an attractor into a sentence can flip its inherent disambiguation bias towards the attractor's sense cluster. Thus, translations of the so perturbed sentence will be more likely to contain WSD errors. The corresponding sample generation strategy consists of four stages:

1. Select *seed* sentences containing homographs to be adversarially perturbed.

2. Identify attractors that are likely to yield fluent and natural samples.

3. Apply perturbations by introducing attractors into seed sentences.

4. Predict effective adversarial samples based on attractor properties.

The targeted attack is deemed successful if a victim model accurately translates the homograph in the seed sentence, but fails to correctly disambiguate it in the adversarially perturbed sample, instead translating it as one of the senses belonging to the attractor's sense cluster. This is a significantly more challenging attack success criterion than the general reduction in test BLEU typically employed for evaluating adversarial attacks on NMT systems (Cheng et al., 2019). Samples are generated using homographs and attractors collected in section 3.2.1, while all test sentence pairs extracted in section 3.2.2 form the domain-specific seed sentence pools. Attack success is evaluated on the same baseline translation models as used throughout section 3.2.

Seed sentence selection

In order to generate informative and interesting adversarial samples, we focus on seed sentences that are likely to be unambiguous. We thus apply three filtering heuristics to seed sentence pairs:

- Sentences have to be at least 10 tokens long.
- We mask out the correct homograph sense in the reference translation and use a pre-trained German BERT model (Devlin et al., 2019)[15] to predict it. Pairs are rejected if the most probable sense does not belong to the correct sense cluster which suggests that the sentence context may be insufficient for correctly disambiguating the homograph. As a result, WSD errors observed in model-generated translations of the constructed adversarial samples are more likely to be due to the applied adversarial perturbations.
- 10% of pairs with the highest disambiguation bias towards incorrect sense clusters are removed from the seed pool.

Setting the rejection threshold above 10% can further reduce WSD errors in seed sentences. At the same time, it would likely render minimal perturbations ineffective, due to the sentences' strong bias towards the correct homograph sense. Thus, we aim for a working compromise.

IH	During this **first** **spring**, he planted another tree that looked the same.
RH	A **hot** ~~new~~ **spring** will conquer the dark nights of winter.
InH	Come the **spring**, I will be invading the **whole** country called Frankia.
RnH	After a long, **eternal** ~~fallow~~ winter, **spring** has come again to Fredericks Manor.

Table 3.9: Perturbation examples; seed sense: *season*, adversarial sense: *water source*. Insertion/replacement in red.

Perturbation types

Naively introducing new words into sentences is expected to yield disfluent, unnatural samples. To counteract this, we constrain candidate attractors to adjectives, since they can usually be placed in front of English nouns without violating grammatical constraints. We consider four perturbation types:

- **Insertion** of the attractor adjective in front of the **homograph** (IH)
- **Replacement** of a seed adjective modifying the **homograph** (RH)
- **Insertion** of the attractor adjective in front of a **non-homograph noun** (InH)
- **Replacement** of a seed adjective modifying a **non-homograph noun** (RnH)

Replacement strategies require seed sentences to contain adjectives, but can potentially have a greater impact on the sentence's disambiguation bias by replacing attractors belonging to the correct sense cluster. Examples for each generation strategy are given in Table 3.9, with homographs highlighted in blue and added attractors in red.

Attractor selection

Since adjectives are subject to selectional preferences of homograph senses, not every attractor will yield a semantically coherent adversarial sample. For instance, inserting the attractor *flying* in front of the homograph *bat* in a sentence about baseball will likely produce a nonsensical expression, whereas an attractor like *huge* would be more acceptable. We attempt to control for this type of disfluency by only considering attractors that had been previously observed to modify the homograph in its seed sentence

[15]We use the implementation provided by the Transformers library (Wolf et al., 2020). We do not fine-tune BERT, as our use case corresponds to its original masked language modeling objective.

sense. For non-homograph perturbations, attractors must have been observed modifying the non-homograph noun. This is ensured by obtaining a dependency parse for each sentence in the English half of the training data and maintaining a list of modifier adjectives for each known target homograph sense set and source noun.[16]

Lastly, to facilitate the fluency and naturalness of adversarial samples, the generation process incorporates a series of constraints:

- Comparative and superlative adjective forms are excluded from the attractor pool.
- Attractors may not modify compound nouns due to less transparent selectional preferences.
- Attractors are not allowed next to other adjectives modifying the noun, to avoid violating the canonical English adjective order.

As all heuristics rely on POS taggers or dependency parsers,[17] they are not free of noise, occasionally yielding disfluent or unnatural samples.

We restrict the number of insertions or replacements to one, so as to maintain a high degree of semantic similarity between adversarial samples and seed sentences. A single seed sentence usually yields several samples, even after applying the aforementioned constraints. Importantly, we generate samples using all retained attractors at this stage, without selecting for expected attack success.

Post-generation filtering

To further ensure the naturalness of generated samples, sentence-level perplexity is computed for each seed sentence and adversarial sample using a pre-trained English GPT2 (Radford et al., 2019) language model.[18] Samples are rejected if their perplexity exceeds that of their corresponding seed sentence by more than 20%. In total, we obtain a pool of ∼500k samples for the OS18 domain and ∼3.9M samples for the WMT19 domain. Each sample is translated by all in-domain models.

Model	$\mathbf{FREQ}_{\text{✗}}$	$\mathbf{PPMI}_{\text{✗}}$	$\mathbf{FREQ}_{\mathbf{DIFF}}$	$\mathbf{PPMI}_{\mathbf{DIFF}}$
OS18 Transformer	0.307	0.367	**0.438**	0.306
OS18 LSTM	0.258	0.261	**0.375**	0.227
OS18 ConvS2S	0.228	0.174	**0.325**	0.165
WMT19 Transformer	0.241	0.241	**0.264**	0.224
WMT19 LSTM	0.278	0.256	**0.316**	0.231
WMT19 ConvS2S	0.304	0.270	**0.328**	0.216

Table 3.10: Rank biserial correlation between attractors' disambiguation bias and attack success.

3.3.2 Identifying effective attractors

The success of the proposed attack strategy relies on the selection of attractors that are highly likely to flip the homograph translation from the correct *seed* sense towards an *adversarial* sense belonging to the attractors' own sense set. To identify such attractors, we examine correlations between attractors' disambiguation biases and the effectiveness of adversarial samples containing them. The attractors' bias values are based either on co-occurrence frequencies (Eqn. 3.1) or PPMI scores (Eqn. 3.2) with the homographs' sense clusters. In particular, we examine the predictive power of an attractor's bias towards the adversarial sense cluster ($DB_{\text{✗}}$) as well as the difference between its adversarial and seed bias values (DB_{DIFF}). As before, RBC and MWU measures are used to estimate correlation strength, with Table 3.10 summarizing the results. The corresponding base-rate adjusted effect size interpretation thresholds are summarized in Table 3.11.

Similarly to the findings reported in section 3.2.2, all uncovered correlations surpass the *large* effect size thresholds, and are statistically significant with $p <$ 1e-5. Importantly, $FREQ_{DIFF}$ exhibits the strongest correlation in all cases.

We are furthermore interested in establishing which of the proposed perturbation methods yields most effective attacks. For this purpose, we examine the percentage of attack successes per perturbation strategy in Figure 3.2, finding perturbations prox-

[16] This assumes correctness of homograph reference translations, which is unfortunately not always guaranteed.
[17] We use spaCy in all cases.
[18] As implemented in the Transformers library.

Model	small	medium	large
OS18 Transformer	0.0339	0.0846	0.1345
OS18 LSTM	0.0338	0.0842	0.1340
OS18 ConvS2S	0.0328	0.0817	0.1301
WMT19 Transformer	0.0166	0.0414	0.0661
WMT19 LSTM	0.0178	0.0446	0.0712
WMT19 ConvS2S	0.0219	0.0548	0.0874

Table 3.11: Base-rate adjusted thresholds for the interpretation of attack success correlations.

imate to the homograph to be most effective.

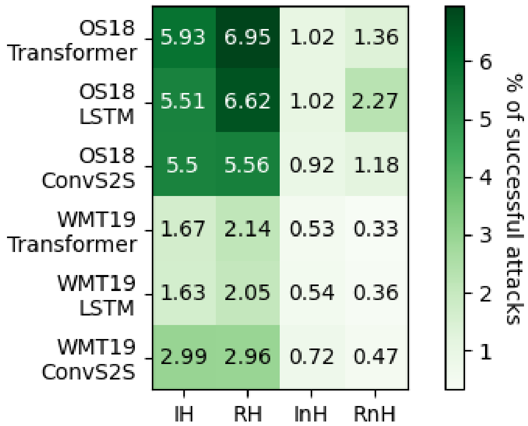

Figure 3.2: Successful attacks per perturbation.

Challenge set evaluation

Having thus identified a strategy for selecting attractors that are likely to yield successful attacks, we construct a challenge set of 10,000 adversarial samples with the highest attractor $\text{FREQ}_{\text{DIFF}}$ scores that had been obtained via the IH or RH perturbations. To enforce sample diversity, we limit the number of samples to at most 1,000 per homograph. Additionally, we create equally-sized, secondary challenge sets by drawing samples at random from each domain's sample pool. Figure 3.3 illustrates the attack success rate for both categories, while Table 3.12 shows some of the successful attacks on the OS18 transformer. Tables 3.13 - 3.18 list further examples of successful

Source input / Original output / Perturbed output	Seed sense	Adv. sense
S: We played the songs again until we felt they sounded right, worked out all the (**nasty**) **bugs**. **O**: Wir spielten die Lieder wieder, bis sie sich richtig anhörten und alle **Fehler**✓ ausarbeiteten. **P**: Wir spielten die Lieder wieder, bis sie sich richtig anhörten und alle **bösen** **Käfer**✗ ausarbeiteten.	*error*	*insect*
S: The driver gets out, opens the (**large**) **boot**, takes some flowers out to deliver. **O**: Der Fahrer steigt aus, öffnet den **Kofferraum**✓, nimmt ein paar Blumen zum Ausliefern mit. **P**: Der Fahrer steigt aus, öffnet den **großen** **Stiefel**✗, nimmt ein paar Blumen zum Ausliefern mit.	*trunk*	*shoe*
S: The doctor somehow got that wig mixed up with the newspapers and (**different**) **letters**. **O**: Der Arzt verwechselte die Perücke mit den Zeitungen und **Briefen**✓. **P**: Der Arzt verwechselte die Perücke mit den Zeitungen und **anderen** **Buchstaben**✗.	*message*	*character*
S: And he will not cease until every **last** **race** of the Four Lands is destroyed. **O**: Und er wird nicht aufgeben, bis jede **Rasse**✓ der Vier Länder ausgelöscht ist. **P**: Und er wird nicht aufhören, bis jedes **letzte** **Rennen**✗ der Vier Länder zerstört ist.	*ethnic group*	*contest*

Table 3.12: Examples of successful attacks on the OS18 transformer. Homographs are blue, attractors are red.

adversarial attacks across the examined model architectures and dataset domains and are included at the end of this chapter for readability.

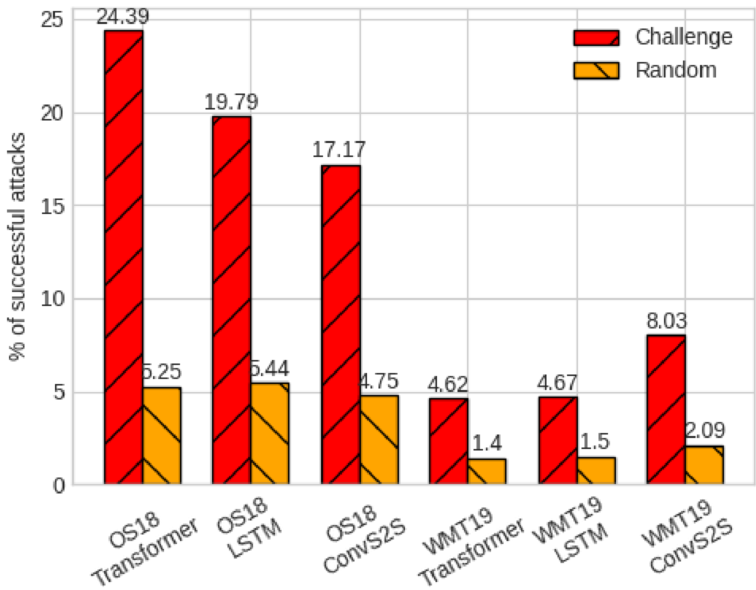

Figure 3.3: Successful challenge sets attacks.

The success rates are modest, ranging from 4.62% to 24.39%, but nonetheless showcase the capacity of targeted, minimal perturbations for flipping correct homograph translations towards a specific sense set. Since our attacks do not require access to model gradients or predictive score distributions, fall within the same domain as the models' training data, and have a strict notion of success, direct comparisons with previous work are difficult. Crucially, compared with a random sample selection strategy, subsampling informed by attractors' disambiguation bias is up to **4.25** times more successful at identifying effective adversarial samples.

While the relative improvement in attack success rate over the random baseline is comparable in both domains, the OS18 models are more susceptible to attacks in absolute terms. This may be due to their lower quality, or the properties of the training data, which can suffer from noisiness (Lison et al., 2019). Interestingly, the relative robustness of individual model architectures to WSD attacks also differs between domains, despite similar quality in terms of BLEU (see Table 3.6). A more thorough investigation of architecture-specific WSD vulnerabilities is left for future work.

3.3.3 Sample quality analysis

To examine whether our adversarial samples would appear trivial and innocuous to human translators, automatic and human evaluation of samples included in the challenge set is conducted. Following (Morris et al., 2020), we use a grammar checker[19] to evaluate the number of cases in which adversarial perturbations introduce grammatical errors. In the OS18 domain, only 1.04% of samples are less grammatical than their respective seed sentences, whereas this is the case for 2.04% of WMT19 samples, indicating a minimal degradation.

We additionally present two bilingual judges with 1,000 samples picked at random from adversarial challenge sets in both domains and 1,000 regular sentences from challenge sets constructed in section 3.2.2. For each adversarial source sentence, annotators were asked to choose whether the homograph's translation belongs to the correct or adversarial seed cluster. For each regular sentence, the choice was between the correct and randomly selected clusters. Across both domains, annotator error rate was 11.23% in the adversarial setting and 11.45% for regular sentences. As such, the generated samples display a similar degree of ambiguity to natural sentences that are likely to elicit WSD errors in NMT models. Annotator agreement was substantial (Cohen's kappa = 0.7).

The same judges were also asked to rate the naturalness of each sentence on a Likert scale from 1 to 5. Perturbed sentences were assigned a mean score of 3.94, whereas regular sentences scored higher at 4.18. However, annotator agreement was low (weighted Kappa = 0.17). The observed drop in naturalness is likely due to the selection of attractors that are not fully consistent with the selectional preferences of

[19]http://languagetool.org

homograph senses during sample generation. We attribute this to WSD errors in reference translations. For instance, we find that the attractor *vampire* is occasionally applied to seed sentences containing the homograph *bat* in its *sporting equipment* sense, which can only occur if the attractor has been observed to modify this sense cluster in the training data (see 3.3.1). Annotator instructions for both tasks are included at the end of this chapter.

3.3.4 Transferability of adversarial samples

An interesting question to consider is whether translation models trained on the same data are vulnerable to the same adversarial samples. We evaluate this by computing the Jaccard similarity index between successful attacks on each baseline model from the entire pool of adversarial samples described in section 3.3.2. We find the similarity to be low, raging between 10.1% and 18.2% for OS18 and between 5.7% and 9.1% for WMT19 samples, which suggests that different model architectures appear to be sensitive to different corpus artifacts, possibly due to differences in their inductive biases.

Considering the observed discrepancy in vulnerabilities between architectures, a natural follow-up question is whether two different instances of the same architecture are susceptible to the same set of attacks. We investigate this by training a second transformer model for each domain, keeping all settings constant with the initial models, but choosing a different seed for the random initialization. While the similarity between sets of successful adversarial samples is greater for two models of the same type, with 25.2% in the OS18 and 12.4% in WMT19 domain, is it still remarkably low.

3.3.5 Comparison with other adversarial attack strategies

As indicated in Chapter 2, this work joins previous efforts dedicated to identifying effective adversarial attacks on NMT models, and NLP models more broadly. This section briefly discusses the advantages and disadvantages of the proposed methodology compared to other existing adversarial attack strategies and considers the ways in which the former complements the latter. Since an extensive survey of adversarial attack methods in NLP goes well beyond the scope of this thesis, we refer the interested reader instead to relevant publications such as (Roth et al., 2021).

Following the taxonomy of adversarial attacks presented in (Shayegani et al., 2023), the attack methodology introduced in this chapter falls within the word-level category, but remains markedly distinct from related methods in that it does not utilize gradient information (Samanta and Mehta, 2017), importance scores (Jin et al., 2020), or random perturbations (Kuleshov et al., 2018), instead relying entirely on the properties of a model's training data to guide the construction of adversarial samples. As a black box method, its main advantage lies in not necessitating access to the internals of the victim model for the identification of effective attractor tokens and their placement within the modified source sentence. This, in turn, makes it broadly applicable to a wide range of NMT models trained on the same data (e.g. the widely used public datasets supplied by the yearly Workshops for Machine Translation (WMT) (Bojar et al., 2018; Kocmi et al., 2022)), thus reducing the need for the crafting of model-specific attack samples. Furthermore, the proposed attack methodology could be equally leveraged for the detection of vulnerabilities in translation system capabilities beyond WSD, such as the modeling of subject-verb agreement where attractors may introduce translation errors by flipping the inflection of the verb (Zacharopoulos et al., 2023), or on-target translations in multilingual NMT systems where attractors could effect target language drift leading to off-target translations (Sennrich et al., 2023).

The proposed attack method does, however, require access to the training data of the victim model, which – e.g. in case of proprietary translation systems – may not be easily available. This, together with the (one-time) manual annotation efforts required to refine BabelNet synsets, represents its primary disadvantage. Another potential weakness of the method is the reliance of the attractor detection protocol on the correctness of reference translations, particularly for settings where a model's training data had been collected from noisy sources, as is the case for the popular OpenSubtitles dataset (Lison et al., 2019). Overcoming these limitations represents a promising research direction for future work concerned with adversarial attacks on NMT models.

In addition to representing a novel class of adversarial attacks, attractor-based strategies can be complimentary to methods relying on gradient information or importance scores (e.g. as provided by attention weights). Specifically, strong attractors identified within the victim model's training data can be preferentially inserted in positions determined via the analysis of model gradients and other importance markers, potentially resulting in more effective adversarial attacks than can be constructed by utilizing the corresponding attack methods in isolation. Moreover, combining attacks that were obtained via a diverse set of strategies, e.g. white-box attacks exploiting model gradients and black-box attacks based on training data properties, can be expected to produce a more challenging robustness benchmark for NMT (and NLP) models compared to benchmarks relying on a single category of attacks, thus facilitating the development of more reliable, consistent, and accurate translation models.

3.4 Conclusion

We conducted an initial investigation into leveraging data artifacts for the prediction of WSD errors in machine translation and proposed a simple adversarial attack strategy based on the presented insights. Our results show that WSD is not yet a solved problem in NMT, and while the general performance of popular model architectures is high, we can identify or create sentences where models are more likely to fail due to data biases.

The effectiveness of our methods owes to neural models struggling to accurately distinguish between meaningful lexical correlations and superficial ones. As such, the presented approach is expected to be transferable to other language pairs and translation directions, assuming that the employed translation models share this underlying weakness. Given the model-agnostic nature of our findings, this is likely to be the case.

As a continuation to this work, we intend to evaluate whether multilingual translation models are more resilient to lexical disambiguation biases and, as a consequence, are less susceptible to adversarial attacks that exploit source-side homography. Extending model-agnostic attack strategies to incorporate other types of dataset biases and to target natural language processing tasks other than machine translation is likewise a promising avenue for future research. Lastly, the targeted development of models that are resistant to dataset artifacts is a promising direction that is likely to aid generalization across linguistically diverse domains.

Annotator instructions

The judges were presented with the following instructions for the described annotation tasks:

Your first task is to judge whether the meaning of the homograph as used in the given sentence is best described by the terms in the SENSE 1 cell or by those in the SENSE 2 cell. Please use the drop-down menu in the WHICH SENSE IS CORRECT? column to make your choice. If you think that neither sens captures the homograph's meaning, please select NONE from the options in the drop-down menu. If you think that the homograph as used in the given sentence can be equally interpreted both as SENSE 1 or SENSE 2, please select BOTH.

We're also asking you to give us your subjective judgment whether the sentence you've been evaluating makes sense to you, i.e. whether it's grammatical, whether it can be easily understood, and whether it sounds acceptable to you as a whole. Typos and spelling mistakes, on the other hand, can be ignored. Specifically, we would like you to assign each sentence a naturalness score, ranging from 1 to 5, according to the following scale:

- *1 = Completely unnatural (i.e. sentence is clearly ungrammatical, highly implausible, or meaningless / incoherent)*
- *2 = Somewhat unnatural (i.e. sentence is not outright incoherent, but sounds very strange)*
- *3 = Unsure (i.e. sentence is difficult to judge either way)*
- *4 = Mostly natural (i.e. sentence sounds good for the most part)*
- *5 = Completely natural (i.e. a well-formed English sentence)*

For instance a sentence like "John ate ten pancakes for breakfast." may get a ranking between 4 and 5, as it satisfies all of the above criteria. A sentence like "John ate green pancakes for breakfast." is grammatical but somewhat unusual and may therefore get a score between 3 and 4. "John ate late pancakes for breakfast.", on the other hand, does not sound very natural since pancakes cannot be "late" and may therefore be rated as 1 or 2. For this judgment we ask you to pay special attention to words in the neighborhood of the homograph. To submit your judgment please select the appropriate score from the drop-down menu in the DOES THE SENTENCE MAKE SENSE? column.

Examples of successful adversarial samples

Tables 3.13 - 3.18 list examples of successful adversarial attacks across the examined model architectures and dataset domains. As before, homographs are highlighted in blue, whereas the introduced attractors are emphasized in red.

Source input / Original output / Perturbed output	**Seed sense**	**Adv. sense**
S: The Penguin was beating him with an (**old**) **bat**, but it was Gordon that pulled the trigger. **O**: Der Pinguin hat ihn mit einem **Schläger**$_✓$ geschlagen, aber Gordon hat abgedrückt. **P**: Der Pinguin hat ihn mit einer **alten** **Fledermaus**$_✗$ geschlagen , aber Gordon hat abgedrückt.	*club*	*animal*
S: I'm not going to relax until that thing its back in its (**simple**) **case**. **O**: Ich werde mich nicht entspannen, bis dieses Ding nicht seinen Rücken in seinem **Koffer**$_✓$ hat. **P**: Ich werde mich nicht entspannen, bis das Ding nicht seinen Rücken in seinem **einfachen** **Fall**$_✗$ hat.	*container*	*instance*
S: "They rest in their mother's (**hot**) **lap**, enjoying the ultimate bliss" **O**: "Sie ruhen im **Schoß**$_✓$ ihrer Mutter, genießen das ultimative Glück" **P**: "Sie ruhen in der **heißen** **Runde**$_✗$ ihrer Mutter, genießen das ultimative Glück"	*body part*	*circuit*
S: That's mighty neighbourly, but I got to play the (**big**) **organ** for the parson tonight. **O**: Das ist mächtig nachbarschaftlich, aber ich muss heute Abend **Orgel**$_✓$ für den Pfarrer spielen. **P**: Das ist mächtig nachbarschaftlich, aber ich muss heute Abend das **Organ**$_✗$ für den Pfarrer spielen.	*instrument*	*body part*
S: I'm just gonna write a (**high**) **note**, and then we'll go. **O**: Ich schreibe nur einen **Zettel**$_✓$ und dann gehen wir. **P**: Ich schreibe einen **hohen** **Ton**$_✗$ und dann gehen wir.	*writing*	*tone*

Table 3.13: Additional examples of successful attacks on the OS18 transformer. Homographs are blue, attractors are red.

Source input / Original output / Perturbed output	Seed sense	Adv. sense
S: I only sell (**good**) **arms** to people who fight clean wars! sure! **O**: Ich verkaufe nur **Waffen**✓ an Leute, die saubere Kriege bekämpfen. **P**: Ich verkaufe nur **gute Arme**✗ an Leute, die saubere Kriege bekämpfen.	*weapon*	*body part*
S: We've heard they're trying to raise (**new**) **capital** to rebuild their armies. **O**: Wir haben gehört, sie wollen **Kapital**✓ sammeln, um ihre Armeen aufzubauen. **P**: Wir haben gehört, dass sie eine **neue Hauptstadt**✗ aufziehen wollen, um ihre Armeen aufzubauen.	*money*	*city*
S: Did you charge the Donellys for five (**closed**) **cases** of vodka? **O**: Haben Sie die Donellys für fünf **Kisten**✓ Wodka berechnet? **P**: Haben Sie die Donellys für fünf **geschlossene Fälle**✗ Wodka berechnet?	*container*	*court case*
S: All units, repeat. that is a battered yellow van, no (**separate**) **plates**. **O**: An alle Einheiten, das ist ein gegrillter gelben Van, keine **Nummernschilder**✓. **P**: An alle Einheiten, das ist ein gegrillter gelben Van, keine **getrennten Teller**✗.	*number plate*	*dish*
S: Um, (**old**) **seals** tell the truth, but a sea lion's always lyin'? **O**: **Robben**✓ sagen die Wahrheit, aber ein Seelöwen lügt immer ? **P**: **Alte Siegel**✗ sagen die Wahrheit, aber ein Seelöwen lügt immer?	*animal*	*emblem*

Table 3.14: Examples of successful attacks on the OS18 LSTM. Homographs are blue, attractors are red.

Source input / Original output / Perturbed output	Seed sense	Adv. sense
S: - Oh, well, keep the (**small**) **change** and have a drink on me. **O**: Behalten Sie den **Rest**✓ und trinken Sie auf mich. **P**: Oh, nun, behalte die **kleine Veränderung**✗ und trink einen auf mich.	*coins*	*development*
S: Do you know how that (**specific**) **date** went, by any chance? **O**: Wissen Sie, wie das **Date**✓ gelaufen ist? **P**: Wissen Sie, wie das **Datum**✗ gelaufen ist?	*meeting*	*calendar date*
S: Goal! (public address) An amazing last-minute third goal that takes Greenock into the (**strong**) **lead**. **O**: Ein erstaunliches drittes drittes Ziel, das Greenock in die **Führung**✓ führt. **P**: Ein erstaunliches drittes Ziel, das Greenock in die **starke Spur**✗ führt.	*first place*	*clue*
S: I mean, you seem like someone who plots out every (**fucking**) **move**. **O**: Ich meine, Sie scheinen jemand zu sein, der jeden **Schritt**✓ aussticht. **P**: Ich meine, Sie scheinen jemand zu sein, der jede **verdammte Bewegung**✗ ausschüttet.	*action*	*movement*
S: You know, if we get hungry, we eat some chips, have some (**crazy**) **punch** ... **O**: Weißt du, wenn wir hungrig werden, essen wir ein paar Chips, haben etwas **Punsch**✓ ... **P**: Weißt du, wenn wir hungrig werden, essen wir ein paar Chips, haben einen **verrückten Schlag**✗ ...	*drink*	*hit*

Table 3.15: Examples of successful attacks on the OS18 ConvS2S. Homographs are blue, attractors are red.

Source input / Original output / Perturbed output	**Seed sense**	**Adv. sense**
S: Copenhagen - Copenhagen, Denmark's (**financial**) **capital**, wants to be the world's first CO2-neutral city by 2025. **O**: Kopenhagen - Kopenhagen, die **Hauptstadt**✓ Dänemarks, will bis 2025 die erste CO2-neutrale Stadt der Welt sein. **P**: Kopenhagen - Kopenhagen, das **Finanzkapital**✗ Dänemarks, will bis 2025 die erste CO2-neutrale Stadt der Welt sein.	*city*	*money*
S: This is done by pricking the earlobe with a small lancet and taking a (**real**) **drop** of blood. **O**: Dies geschieht, indem der Ohrwurm mit einem kleinen Lancet geprickt wird und ein **Tropfen**✓ Blut eingenommen wird. **P**: Dies geschieht, indem der Ohrwurm mit einem kleinen Lancet geprickt wird und ein **richtiger** Blut**abfall**✗ entsteht.	*drop of liquid*	*decrease*
S: One (**small** ~~positive~~) **note** was from the Republic of Ireland, which saw its PMI grow to 57.3, its highest level since the end of 1999. **O**: Eine positive **Anmerkung**✓ war die aus der Republik Irland, wo das PMI auf 57,3 anstieg, das höchste Niveau seit Ende 1999. **P**: Ein **kleiner Schein**✗ stammt aus der Republik Irland, wo das PMI auf 57,3 anstieg, das höchste Niveau seit Ende 1999.	*remark*	*paper money*
S: His epoch-making (**full**) **record** "Free Jazz" was released by Atlantic Records at the dawn of that decade. **O**: Seine epochale **Platte**✓ "Free Jazz" wurde zu Beginn des Jahrzehnts von Atlantic Records veröffentlicht. **P**: Seine epochale **Aufzeichnung**✗ "Free Jazz" wurde zu Beginn des Jahrzehnts von Atlantic Records veröffentlicht.	*musical medium*	*document*
S: After winter delivered an early dose of (**natural**) **spring** last week, temperatures dropped again on Monday to a high of just 15.8C in the city. **O**: Nachdem der Winter vergangene Woche eine frühe **Frühjahr**sdosis✓ geliefert hatte, fielen die Temperaturen am Montag wieder auf einen Höchstwert von nur 15,8C in der Stadt. **P**: Nachdem der Winter letzte Woche eine frühe Dosis **Naturquelle**✗ lieferte, fielen die Temperaturen am Montag wieder auf einen Höchstwert von nur 15,8C in der Stadt.	*season*	*water source*

Table 3.16: Examples of successful attacks on the WMT19 transformer. Homographs are blue, attractors are red.

Source input / Original output / Perturbed output	**Seed sense**	**Adv. sense**
S: A Thousand Splendid Suns is a story of two women's lives in Afghanistan, where women are equal, as a table or the (**last**) **chair**. **O**: Ein Thousand Splendid Seine ist eine Geschichte von zwei Frauen in Afghanistan, wo Frauen gleich sind, als Tisch oder **Stuhl**✓. **P**: Ein Thousand Splendid Seine ist eine Geschichte von zwei Frauen in Afghanistan, wo Frauen gleich sind, als Tisch oder als **letzter** **Vorsitzender**✗.	*furniture*	*chairperson*
S: See a (**small** ~~rapid~~) **drop** in your CO level once you stop smoking. **O**: Sehen Sie sich einen schnellen **Rückgang**✓ Ihrer CO-Ebene an, sobald Sie das Rauchen einstellen. **P**: Sehen Sie einen **kleinen** **Tropfen**✗ auf Ihrem CO-Niveau, sobald Sie aufhören, Rauchen zu beenden.	*decrease*	*drop of liquid*
S: And moreover - each of our guests will get a (**different** ~~small~~) **present**! **O**: Und darüber hinaus wird jeder unserer Gäste ein kleines **Geschenk**✓ bekommen! **P**: Und darüber hinaus wird jeder unserer Gäste eine **andere** **Gegenwart**✗ bekommen!	*gift*	*current time*
S: A (**new**) **record** of every transaction made is kept, allowing for a complete audit if necessary. **O**: Ein **Datensatz**✓ jeder Transaktion wird gehalten, so dass erforderlichenfalls eine vollständige Prüfung möglich ist. **P**: Ein **neuer** **Rekord**✗ jeder Transaktion wird gehalten, so dass erforderlichenfalls eine vollständige Prüfung möglich ist.	*document*	*achievement*
S: Britain's new trade deals with non-EU countries would also probably involve (**political** ~~worse~~) **terms**. **O**: Die neuen Handelsvereinbarungen Großbritanniens mit Nicht-EU-Ländern würden wahrscheinlich auch schlechtere **Bedingungen**✓ beinhalten. **P**: Großbritanniens neue Handelsabkommen mit Nicht-EU-Ländern würden wahrscheinlich auch **politische** **Begriffe**✗ beinhalten.	*demand*	*expression*

Table 3.17: Examples of successful attacks on the WMT19 LSTM. Homographs are blue, attractors are red.

Source input / Original output / Perturbed output	**Seed sense**	**Adv. sense**
S: Not to mention (non) uniform loading and soring fingers, contaminated with (**common**) **lead**. **O**: Ganz zu schweigen von (nicht) einheitlichen Lade- und Sortierfingern, die mit **Blei**✓ kontaminiert sind. **P**: Ganz zu schweigen von (nicht) einheitlichen Lade- und Sortierfingern, die mit einer **gemeinsamen Führung**✗ kontaminiert sind.	*metal*	*first place*
S: If the symbol "&gt" is displayed, keep entering (**greek**) **letters** until predictive options are displayed. **O**: Wenn das Symbol "&gt" angezeigt wird, erhalten Sie die Eingabe von **Buchstaben**✓, bis prognostizierte Optionen angezeigt werden. **P**: Wenn das Symbol "&gt" angezeigt wird, erhalten Sie immer wieder Gruß**briefe**✗, bis prognostizierte Optionen angezeigt werden.	*character*	*message*
S: This film is not about dialogue or a (**little** ~~stringent~~) **plot**, but all about atmosphere - a feverish dream that has become a film. **O**: In diesem Film geht es nicht um einen Dialog oder um eine strenge **Handlung**✓, sondern um die Atmosphäre - ein feverser Traum, der zu einem Film geworden ist. **P**: In diesem Film geht es nicht um Dialog oder ein **wenig Grundstück**✗, sondern alles über die Atmosphäre - ein feverser Traum, der zu einem Film geworden ist.	*story*	*tract of land*
S: Manufacture of products from silicone and rubber, Production of springs, Manufacturing of springs, Winding of (**small**) **springs**. **O**: Herstellung von Produkten aus Silikon- und Gummi, Herstellung von Quellen, Herstellung von Quellen, **Federn**✓. **P**: Herstellung von Produkten aus Silikon- und Gummi, Herstellung von Quellen, Herstellung von Quellen, Winding von **kleinen Quellen**✗.	*device*	*water source*
S: In 1980, financial assets - (**large**) **stocks**, bonds, and bank deposits - totaled around 100% of GDP in the advanced economies. O; Im Jahr 1980 belief sich das Finanzvermögen - **Aktien**✓, Anleihen und Bankeinlagen - in den hochentwickelten Volkswirtschaften rund 100% des BIP. **P**: Im Jahr 1980 belief sich das Finanzvermögen - **große Bestände**✗, Anleihen und Bankeinlagen - in den hochentwickelten Volkswirtschaften rund 100% des BIP.	*investment*	*inventory*

Table 3.18: Examples of successful attacks on the WMT19 ConvS2S. Homographs are blue, attractors are red.

3.5 Post-Publication Comments

Despite being a relatively recent development, LLMs that have been pretrained on data from multiple languages (Liu et al., 2020; Xue et al., 2021; Scao et al., 2022; OpenAI, 2023) have shown some potential as versatile and controllable translation engines (Zhang et al., 2023; Hendy et al., 2023). Translation capabilities appear to emerge and improve in LLMs as a function of model scale, data size, and data variety, and are not contingent on any specialized optimization function other than the standard, monolingual language modeling objective (OpenAI, 2023).

Within the context of this chapter, the question of whether this new variety of translation models is less susceptible to WSD errors arises naturally. However, for most SOTA LLMs, accessing the training data for the estimation of disambiguation biases is not possible due to its proprietary nature and difficult-to-replicate data pre-processing steps. Moreover, given that LLM training relies heavily on monolingual text across different languages, identifying effective attractors for individual senses of polysemous terms becomes particularly challenging, as it is no longer possible to rely on the properties of parallel data alone, as done in Section 3.2.1. Consequently, the methodologies for the prediction of WSD errors and for the crafting of adversarial WSD attacks presented in this chapter with application to NMT models can not be applied to LLMs without significant modifications.

Nevertheless, given that OS18 and WMT19 data are both easily accessible online and are of reasonably high quality, both are highly likely to be included in the training distribution of a recent LLM trained on large quantities of crawled web data. As a consequence, the LLM's WSD preferences could reflect some of the biases associated with these collections of parallel text. To examine whether this is indeed the case, challenge sets referenced in Sections 3.2.2 (i.e. 3k naturally occurring sentences containing homographs) and 3.3.1 (i.e. 10k of synthetic sentences constructed as adversarial attacks on NMT models) were translated using the commercial ChatGPT API provided by OpenAI (OpenAI, 2022), with *gpt-3.5-turbo* as the backend model[20]. The API allows for convenient access to a much-studied and reported on, proprietary SOTA LLM and was therefore deemed as the preferred method for collecting model responses. Moreover, the GPT model family has been recently found to achieve highly competitive translation quality, which makes it an attractive subject for studying translation abilities in LLMs (Hendy et al., 2023).

To obtain translations of English sentences making up the aforementioned challenge sets, ChatGPT was provided with a task-specific prompt reproduced in Table 3.19. The prompt is limited to a concise task description and states the main requirements to be fulfilled by the generated response. It is subsequently referred to as the *basic prompt*. Furthermore, the model was given 10 examples (i.e. "shots") for the OS18 and 4 examples for the WMT19 data, since the latter contains on average much longer and less noisy sentences. These demonstrations were selected at random from each respective challenge set and subsequently excluded from the evaluation.

[20]All ChatGPT queries reported in this manuscript were issued in June - August 2023.

U: Your task is to generate a valid German translation for the provided English sentence. The correct translation must fully and accurately capture the meaning of the English sentence, be fluent, and grammatical. Your response should not contain anything else. English sentence to translate: ["My dad would jump overboard with an anchor before he ever talks to me again."]	
S: [Mein Vater würde mit einem Anker über Bord springen, bevor er wieder anfängt mit mir zu reden.]	

Table 3.19: Basic translation prompt used with ChatGPT. U = User, S = System. Segments given in square brackets are placeholders for in-context learning examples ("shots") provided to the model.

Dataset	# API failures (%)	# Invalid translations (%)	# Valid translations (%)
OS18 natural samples	83 (2.77%)	181 (6.03%)	2,736 (91.2%)
WMT19 natural samples	66 (2.2%)	78 (2.6%)	2,856 (95.2%)
OS18 adversarial samples	683 (6.83%)	776 (7.76%)	8,541 (85.41%)
WMT19 adversarial samples	539 (5.39%)	577 (5.77%)	8,884 (88.84%)

Table 3.20: Fractions of successful and unsuccessful WSD queries addressed to ChatGPT via the web API. *Invalid translations* refers to cases where the model did not produce a translation, instead returning an explanation as to why a translation can not be provided (e.g. due to toxicty).

We note that it was not possible to obtain valid model replies for all of our queries. This can be attributed to several reasons, including the API timing out or returning an error code, the input sentence being considered impossible to translate by the model (e.g. because of limited context, fragmented language, slang expressions, or typos and misspellings), or the input containing words or phrases deemed inappropriate by the model (e.g. mentions of reproductive anatomy, slurs and other offensive language). We jointly refer to the latter two categories in the following as *invalid replies*. Table 3.20 reports the percentage of samples based on whether ChatGPT returned a valid response to the API query, an invalid one, or whether the API call had failed altogether. For all of the evaluated challenge sets, 85%+ of samples were successfully translated.[21] As such, while the translation step eliminated some of the challenge samples from the

[21] Here, *valid translations* count represents the intersection of valid translations obtained using basic prompting and Chain-of-Thought prompting discussed in subsequent paragraphs of this section, to ensure a fair comparison between the two prompting methods. Some of the samples were successfully translated with one prompt format but yielded an invalid translation with the other, which explains the elevated API request failure rate.

evaluation, we deem the remaining set to be sufficiently large to offer useful insights into the LLM's WSD behaviour as well as its potential flaws. At the same time, we note that the omission of samples makes any comparison with results reported in the earlier sections of this chapter incomplete and approximate.

OS18 natural	**OS18 adversarial**	**WMT19 natural**	**WMT19 adversarial**
10.14%	15.4%	10.68%	3.21%

Table 3.21: ChatGPT failure rate for the examined WSD challenge sets.

As summarized in Table 3.21, ChatGPT remains somewhat susceptible to WSD errors despite its advanced language understanding capabilities. Still, it obtains a substantially better WSD accuracy on natural, non-adversarial samples compared to the best-performing NMT transformer model – 10.14% vs. 23.4% in the OS18 domain and 10.68% vs. 17.93% in the WMT19 domain. Unfortunately, these results do not indicate whether the improvement is primarily due to a fundamental superiority of the LLM's lexical disambiguation capabilities, the inclusion of the test sentences in its training data[22], or a difference in the models' disambiguation bias arising from different training distributions. On adversarial examples, ChatGPT similarly improves over the evaluated NMT transformer models – 15.4% vs. 24.39% in the OS18 domain and 3.21% vs. 4.62% in the WMT19 domain. These results also imply a certain degree of overlap in the disambiguation biases between ChatGPT and the examined translation models which, in turn, makes it more likely that the observed improvements on *non-adversarial* samples are due to the LLM being the superior lexical disambiguator[23].

It is, however, possible that the basic prompt used so far to elicit translations from the LLM (see Table 3.19) could be modified to further improve WSD accuracy, in light of the demonstrable importance of prompt formats in a variety of NLP tasks (White et al., 2023). To investigate whether this is the case, CoT prompting (Wei et al., 2022b) was leveraged to encourage the model to reason about homographs present in the input sentence and determine which of their senses is the most appropriate given the sentence context. The corresponding CoT prompt is reproduced in Table 3.22. In contrast to the initial prompt, it asks the LLM to identify relevant homographs, decide on their

[22]However, as the model's performance remains far from perfect, the model does not appear to have memorized reference translations within the OS18 and WMT19 corpora.

[23]As none of the challenge sets used to evaluate the WSD capabilities of ChatGPT in this section are accessible online, their potential inclusion in the LLM's training distribution is not a valid explanation for the observed performance difference compared to NMT models.

intended meaning, and incorporate these considerations when generating the translation hypothesis. We also manually construct four prompting examples representative of the intended reasoning format that are used together with both natural and adversarial samples across both test domains, selecting the underlying samples at random. The system response included in 3.22 represents one of the prompting examples. We ensure a fair evaluation of the impact the CoT prompt may have on the LLM's WSD accuracy relative to the previously examined basic prompt, by only taking into account samples that were successfully translated with both prompting strategies.

U: Your task is to return a valid German translation for the provided English sentence. The correct translation must fully and accurately capture the meaning of the English sentence, be fluent, and grammatical. Before providing your translation, you should reason about the ambiguous nouns (specifically homographs) found in the sentence and decide on their meaning within the sentence context, which should then inform your translation. Here are some examples of what this should look like. English sentence: "I've been dragging you through time like an anchor and now it's time to cut you loose."
S: A homonym present in this sentence is the English noun "anchor". In this sentence, "anchor" refers to a heavy object that is difficult to drag, and should therefore be translated as the German noun "Anker" which can be defined as a heavy object used to moor a ship to the bottom of the sea. Thus it follows that the German translation of the full English sentence is: "Ich habe dich wie einen Anker durch die Zeit gezogen und jetzt ist es an der Zeit, dich loszulassen."

Table 3.22: CoT prompt used with ChatGPT with a focus on accurate WSD. U = User, S = System.

Table 3.23 reports WSD error rates across the four challenge sets, showing mixed results. While a substantial improvement in robustness to adversarial attacks can be observed in the OS18 domain, vulnerability to attacks increases marginally in the WMT19 domain. For non-adversarial data, CoT prompting yields negative results in both evaluated domains, with failure rate increasing by up to 1.26%. Consequently, it cannot be asserted with certainty that CoT prompting represents a superior response elicitation method for WSD-sensitive translation, unlike for other NLP tasks. It is, however possible, that a differently structured CoT prompt would obtain markedly different results, which is left for future work to investigate. Concrete examples of CoT prompting rectifying disambiguation errors are provided in Tables 3.24 - 3.27. Interestingly, in some cases, the CoT prompt is effective even if the LLM fails to identify valid homographs and explicitly reason about their meaning. One of such examples is included at the bottom of Table 3.26.

OS18 natural	OS18 adversarial	WMT19 natural	WMT19 adversarial
10.69% (+0.55%)	**13.88% (-1.52%)**	11.94% (+1.26%)	3.4% (+0.19%)

Table 3.23: ChatGPT failure rate on the examined WSD challenge sets, using CoT prompting. Changes in failure rate relative to the basic prompt are given in parentheses.

Based on theses additional findings we can conclude that leveraging LLMs as translation engines is a highly promising path towards improved robustness of automatic translations to lexical disambiguation errors. While sophisticated prompting methods that facilitate lexical reasoning can yield improvements in WSD accuracy, understanding the full breath of their utility requires further research. Since the observed improvements can likely be at least in part attributed the LLM's scale and the vast amount of training data it had been exposed to, in contrast to conventional NMT models that are limited in scale and rely on – by comparison – small amounts of highly curated data, integrating LLMs into translation pipelines may become essential for obtaining high-quality translations in cases where WSD is a concern.

Follow-Up Work

The publication of (Emelin et al., 2020) has in part motivated a number of noteworthy follow up works. Among these, (Campolungo et al., 2022b) propose a strategy for integrating word sense annotations into the NMT training pipeline without requiring them at inference. In doing so, they demonstrate improvements to the lexical disambiguation capacity and robustness against adversarial WSD attacks of the evaluated NMT systems according to the evaluation strategy defined in (Emelin et al., 2020). On the other hand, (Vamvas and Sennrich, 2021) describe a reference-free method for detecting lexical disambiguation pathologies in translation models that parallels the construction of adversarial WSD attacks in (Emelin et al., 2020). Using this method, they find that sequence-level knowledge distillation (Kim and Rush, 2016) amplifies undesirable biases of the teacher model in its student. In a related vein, (Zeng and Xiong, 2021) conduct an investigation into model biases that drive WSD errors, extending the findings of the analysis conducted in (Emelin et al., 2020) among others. They reach the conclusion that NMT models are most susceptible to attacks on the source sentence positions aligned with the front positions of the respective target sentences and leverage these insights to craft effective adversarial WSD attacks.

Source / **R**eference translation / **Basic Prompt** reply / **CoT** reply
S: You can shut down the matches, but only for a little while. **R**: Sie können die Wettkämpfe verbieten, aber nur für kurze Zeit. **BP**: Du kannst die Streichhölzer ausmachen, aber nur für eine kurze Zeit. **CoT**: In this sentence, the word "matches" can be understood as either referring to a contest or competition, or to the objects used to start a fire. To determine the intended meaning, we can look at the context of the sentence. Since the sentence is about shutting down the matches, it suggests that it is referring to a contest or competition. Therefore, the word "matches" should be translated as "Wettkämpfe" in German. The German translation of the full English sentence is: "Du kannst die Wettkämpfe beenden, aber nur für eine kurze Zeit."
S: If not, why, in a moment look to see the blind and bloody soldier with foul hand defile the locks of your shrill-shrieking daughters. **R**: Wo nicht, erwartet Augenblicks zu sehen den blinden und blutigen Soldaten mit rauer Hand, besudeln die Locken eurer gellend schreienden Töchter. **BP**: Falls nicht, dann schau in diesem Moment, wie der blinde und verblutende Soldat mit dreckigen Händen das Schloss deiner laut kreischenden Töchter beschmutzt. **CoT**: A homonym present in this sentence is the English noun "locks". In this sentence, "locks" refers to the hair of someone's head, and should therefore be translated as the German noun "Locken" which can be defined as a curl of hair. Thus it follows that the German translation of the full English sentence is: "Wenn nicht, dann schau in diesem Moment zu, wie der blinde und blutige Soldat mit dreckiger Hand die Locken deiner schrill-schreienden Töchter entweiht."
S: Any clown on the street could tell you that battery charge won't stick. **R**: Jeder Trottel auf der Straße könnte Ihnen sagen, dass diese Anklage wegen einer Tätlichkeit nicht halten wird. **BP**: Jeder Trottel auf der Straße kann dir sagen, dass die Akkuladung nicht hält. **CoT**: A homograph present in this sentence is the English noun "charge". In this sentence, "charge" refers to a formal accusation brought against someone, and should therefore be translated as the German noun "Anklage" which can be defined as a formal accusation or charge made in a court of law. Thus it follows that the German translation of the full English sentence is: "Jeder Clown auf der Straße könnte dir sagen, dass die Anklage nicht halten wird."

Table 3.24: Examples of non-adversarial OS challenge set translations elicited from Chat-GPT, where the CoT prompting strategy improves over the basic prompt. Homographs are underlined, correct WSD choices are coloured **blue**, incorrect ones are colored **red**.

Source / Reference translation / **Basic Prompt** reply / **CoT** reply
S: They sat there with *small* arms folded with dirty looks on their face and I swear to God ... **R**: Sie saßen da mit verschränkten Armen und haben mich geil angestarrt. **BP**: Sie saßen da mit verschränkten Waffen und finsterem Blick im Gesicht und ich schwöre bei Gott ... **CoT**: A homograph present in this sentence is the English noun "arms". In this sentence, "arms" refers to a person's limbs, specifically their arms, and should therefore be translated as the German noun "Arme" which can be defined as the part of the human body between the shoulder and the hand. Thus it follows that the German translation of the full English sentence is: "Sie saßen dort mit verschränkten Armen und schmutzigen Blicken im Gesicht, und ich schwöre bei Gott ..."
S: I put a *new* vacuum inside a temporal field until a world developed. **R**: Ich habe ein Vakuum in ein temporales Feld gesteckt bis sich eine Welt entwickelt hat. **BP**: Ich habe einen *neuen* Staubsauger in ein zeitliches Feld gestellt bis sich eine Welt entwickelt hat. **CoT**: In this sentence, the noun "vacuum" is ambiguous. It can refer to a space with no matter or particles (a vacuum), or to a household cleaning device (a vacuum cleaner). Based on the context of the sentence, where the vacuum is put inside a temporal field and a world develops, it is more likely that the intended meaning is a "vacuum" as an empty space. Therefore, the German translation would be: "Ich habe ein *neues* Vakuum in einem zeitlichen Feld platziert, bis sich eine Welt entwickelte."
S: The *good* spring bubbling up you saw me drink from? **R**: Die Quelle, von der du mich hast trinken sehen? **BP**: Der *gute* Frühling, aus dem ich getrunken habe? **CoT**: The provided English sentence contains some ambiguities, making it difficult to determine the precise meaning without more context. However, assuming that "spring" refers to a natural source of water and "bubbling up" is used figuratively, it can be translated as: "Den *guten* Quell, aus dem du mich hast trinken sehen?"

Table 3.25: Examples of adversarial OS challenge set translations elicited from ChatGPT, where the CoT prompting strategy improves over the basic prompt. Homographs are underlined, attractors are set in *cursive*, correct WSD choices are coloured **blue**, incorrect ones are colored **red**. Note, the reference translation corresponds to the original source sentence without the attractor term.

Source / Reference translation / **Basic Prompt** reply / **CoT** reply
S: There are also quite a few races and individuals who have attained immortality after leading a mortal existence. **R**: Andererseits gibt es eine ganze Reihe Völker und Wesen, die nach einem Leben als Sterbliche die Unsterblichkeit erlangt haben. **BP**: Es gibt auch viele Rennen und Einzelpersonen, die Unsterblichkeit erlangt haben, nachdem sie ein sterbliches Leben geführt haben. **CoT**: A homonym present in this sentence is the English noun "races". In this sentence, "races" refers to different groups or species of beings, and should therefore be translated as the German noun "Rassen". However, it is important to note that the concept of immortality and leading a mortal existence can be tricky to express in German. To accurately capture the meaning, it is recommended to rephrase the sentence. Thus, the German translation of the full English sentence can be: "Es gibt auch einige Rassen und Individuen, die nach einem sterblichen Dasein Unsterblichkeit erlangt haben."
S: Currently, the record for image resolution among camera phones is 10 megapixels. **R**: Aktuell liegt der Auflösungsrekord unter Kamera-Handys bei 10 Megapixeln. **BP**: Die derzeitige Aufzeichnung für Bildauflösungen bei Kamerahandys beträgt 10 Megapixel. **CoT**: A homonym present in this sentence is the noun "record." In this sentence, "record" refers to the highest achieved level or value, and should therefore be translated as the German noun "Rekord." Thus the correct translation of the full English sentence is: "Derzeit liegt der Rekord für die Bildauflösung bei Handykameras bei 10 Megapixeln."
S: Its emblem, a stylized crane, is recognized almost everywhere as the symbol of the Polish flag carrier. **R**: Ihr Logo, ein stilisierter Kranich, ist weltweit als Symbol der polnischen Fluggesellschaft bekannt. **BP**: Das Emblem, ein stilisierter Kran, wird nahezu überall als das Symbol des polnischen Staatsfliegers anerkannt. **CoT**: The English noun "crane" can be interpreted as a bird or as a machine used for lifting heavy objects. In this context, "crane" refers to a bird. Thus, the German translation for "Its emblem, a stylized crane, is recognized almost everywhere as the symbol of the Polish flag carrier." is: "Sein Emblem, ein stilisierter Kranich, wird nahezu überall als das Symbol der polnischen Fluggesellschaft erkannt."
S: The mountain spring in May / June brings lush flower-filled meadows and sprinkles the higher slopes with gentians and all kinds of anemones. **R**: Der Bergfrühling im Mai / Juni bringt üppige Blumenwiesen im Tal und in der Höhe mit Enzianen und Anemonen aller Art übersäte Hänge. **BP**: Im Mai und Juni bringt die Gebirgsquelle üppige Blumenwiesen mit sich und besprüht die höheren Hänge mit Enzianen und allerlei Anemonen. **CoT**: **There are no ambiguous nouns in this sentence**. The correct German translation of the sentence is: Der Gebirgsfrühling im Mai / Juni bringt üppige, blumenreiche Wiesen und besprengt die höheren Hänge mit Enzianen und allerlei Anemonen.

Table 3.26: Examples of natural WMT challenge set translations elicited from ChatGPT, where the CoT prompting strategy improves over the basic prompt. Homographs are underlined, correct WSD choices are coloured **blue**, incorrect ones are colored **red**.

Source / Reference translation / **Basic Prompt** reply / **CoT** reply
S: In every *national* race we intend to thrill people and inspire them by our strength of purpose and our team spirit. **R**: Rennen für Rennen wollen wir so die Menschen mitreißen und mit unserem Willen und unserem Teamgeist anstecken. **BP**: In jeder *nationalen* Rasse beabsichtigen wir, die Menschen zu begeistern und durch unsere Zielstrebigkeit und unseren Teamgeist zu inspirieren. **CoT**: A homonym present in this sentence is the English noun "race". In this sentence, "race" refers to a competition among individuals or teams, and should therefore be translated as the German noun "Rennen" which can be defined as a competition involving speed. Thus it follows that the German translation of the full English sentence is: "Bei jedem *nationalen* Rennen beabsichtigen wir, die Menschen zu begeistern und sie mit unserer Zielstrebigkeit und unserem Teamgeist zu inspirieren."
S: - Acknowledges with interest the *first* note from the Presidency on the Science and Technology cooperation between the European Union and Mediterranean third countries; **R**: Nimmt mit Interesse den Vermerk des Vorsitzes zur wissenschaftlichen und technischen Zusammenarbeit zwischen der Europäischen Union und Drittländern im Mittelmeerraum zur Kenntnis; **BP**: - Anerkennt mit Interesse die *erste* Note der Präsidentschaft zur Zusammenarbeit in Wissenschaft und Technologie zwischen der Europäischen Union und den Mittelmeer-Drittländern; **CoT**: A homonym present in this sentence is the English noun "Note". In this sentence, "Note" refers to a written communication or message sent from one party to another, and should therefore be translated as the German noun "Notiz" which can be defined as a brief written record of something. Thus it follows that the German translation of the full English sentence is: "- nimmt mit Interesse die *erste* Notiz des Vorsitzes über die wissenschaftliche und technologische Zusammenarbeit zwischen der Europäischen Union und den Mittelmeerländern Dritter zur Kenntnis;
S: I wish to place on *new* record for Mr Bourlanges and for the House that the majority of my group was disposed to vote for the strategy in paragraph 8. **R**: Ich möchte für Herrn Bourlanges und das Haus zu Protokoll geben, daß die Mehrheit meiner Fraktion für die in Absatz 8 erläuterte Strategie stimmen wollte. **BP**: Ich möchte einen *neuen* Rekord für Herrn Bourlanges und für das Haus aufstellen, dass die Mehrheit meiner Gruppe bereit war, für die Strategie in Absatz 8 zu stimmen. **CoT**: A homonym present in this sentence is the English noun "record". In this sentence, "record" refers to an official documented statement, and should therefore be translated as the German noun "Protokoll" which can be defined as a written account of a meeting or session. Thus it follows that the German translation of the full English sentence is: "Ich möchte ein *neues* Protokoll für Herrn Bourlanges und für das Parlament anführen, dass die Mehrheit meiner Gruppe bereit war, für die Strategie im Absatz 8 zu stimmen."

Table 3.27: Examples of adversarial WMT challenge set translations elicited from ChatGPT, where the CoT prompting strategy improves over the basic prompt. Homographs are underlined, attractors are set in *cursive*, correct WSD choices are coloured **blue**, incorrect ones are colored **red**. Note, the reference translation corresponds to the original source sentence without the attractor term.

Chapter 4

Lacking Text Understanding Leads to Coreference Resolution Errors

Common sense is not so common.

Voltaire, *Dictionnaire Philosophique*

Abstract: Winograd schemas are a well-established tool for evaluating coreference resolution (CoR) and commonsense reasoning (CSR) capabilities of computational models. So far, schemas remained largely confined to English, limiting their utility in multilingual settings. This work presents ***Wino-X*, a parallel dataset of German, French, and Russian schemas, aligned with their English counterparts**. We use this resource to investigate whether neural machine translation (NMT) models can perform CoR that requires commonsense knowledge and whether multilingual language models (MLLMs) are capable of CSR across multiple languages. Our findings show *Wino-X* to be **exceptionally challenging for NMT systems** that are prone to undesirable biases and unable to detect disambiguating information. We quantify biases using established statistical methods and define ways to address both of these issues. We furthermore present evidence of active **cross-lingual knowledge transfer in MLLMs**, whereby fine-tuning models on English schemas yields CSR improvements in other languages.[1]

[1]This section is based on work previously published at EMNLP 2021 (Emelin and Sennrich, 2021). The *Wino-X* dataset and experimental codebase are available at https://github.com/demelin/Wino-X.

4.1 Introduction

Originally introduced in (Winograd, 1972), Winograd schemas (*schemas* from here on) have become an established tool for probing the ability of computational models to reason about natural language. Either viewed through the lens of coreference (CoR) as in (Levesque et al., 2012) or, more recently, framed as a gap-filling task (Sakaguchi et al., 2020), schemas are assumed to require commonsense knowledge to be resolved correctly.

Consider the following schema: *The trophy doesn't fit into the brown suitcase because **it** is too [large / small]*. Here, the pronoun ***it*** has two possible antecedents *(trophy* / *suitcase*), with the choice of the antecedent determined by the trigger word (*large* / *small*). To connect the pronoun to its true antecedent, a model must 'know' that objects that are too large cannot fit into containers and that containers that are too small cannot house objects.

When translating an *instance* of a schema (i.e. the schema with a fixed trigger word) into languages such as German, where pronouns and their antecedents must agree in their grammatical gender, translation models must implicitly perform the CoR step to produce accurate translations. A competent translation model is, therefore, expected to identify the correct antecedent as reflected by the target pronoun choice. The capacity to access and utilize everyday knowledge in a task-oriented manner, e.g. for CoR as in the aforementioned example, is commonly referred to as commonsense reasoning (Sakaguchi et al., 2020), and has been shown to emerge in language processing models trained on the next word prediction objective, in particular LMs of sufficiently large scale (Bhagavatula et al., 2019; Sakaguchi et al., 2020; Emelin et al., 2021). It is not yet clear whether NMT models – commonly trained on modest amounts of parallel, highly curated data – can exhibit this emergent capability but, given its importance for the translation process, an investigation into the commonsense reasoning capabilities of translation models is necessary to guide model development towards accurate and robust translation performance.

In the first part of this work, we construct cross-lingual instances by aligning English instances with their translations into morphologically rich languages, so as to probe the robustness of CoR in current NMT models, as illustrated in Figure 4.1 (top half). In doing so, we show that models follow simplistic heuristics when attempting to resolve coreference, while failing to detect disambiguating information.

A second category of models that is expected to correctly identify coreference in

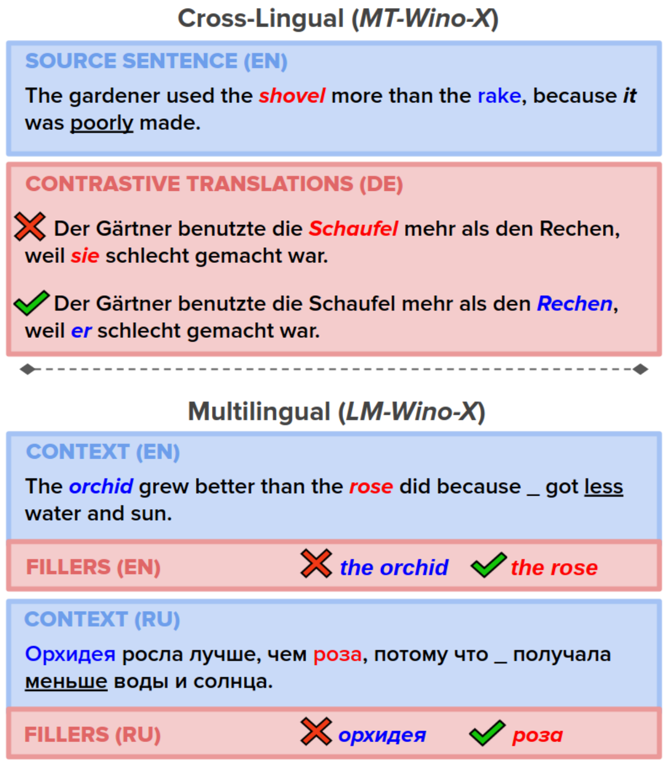

Figure 4.1: *Wino-X* examples. *Cross-lingual* samples are used to evaluate translation models, whereas *multilingual* instances are compatible with MLLMs. Coreferent words are highlighted with the same color, while disambiguating **trigger words** are underlined.

multiple languages are multilingual language models. Where translation models learn to map their input to semantically equivalent sequences in the target language, MLLMs are trained on a mask-filling objective and learn to encode sentences drawn from different languages into a shared semantic space. Accordingly, schema instances correctly solved by MLLMs in one language should be equally solvable in other languages, by leveraging the same, language-agnostic representations. Similarly, improvements to model performance in one language should transfer to other languages via the shared latent space. In the second part of our work, we empirically put these assumptions to the test with multilingual schema instances, as shown in Figure 4.1 (bottom half), finding evidence of active commonsense knowledge transfer across languages.

Our primary contributions are as follows:

1. We introduce ***Wino-X*: A dataset containing Winograd schemas in German, French, and Russian**, aligned with their English analogues.

2. We benchmark the **CoR performance of NMT models** for each language pair, finding it to be **close to chance**.

3. We identify two **causes underlying the poor performance** of the evaluated NMT models and define **ways to mitigate them**.

4. We show that ***Wino-X* presents a challenge for MLLMs**, and observe active **transfer of commonsense knowledge across languages**.

4.2 *Wino-X*: A Contrastive Dataset of Multilingual Winograd Schemas

In order to maximize the coverage and quality of *Wino-X*, we derive multilingual schemas from *WinoGrande* (Sakaguchi et al., 2020), a large-scale, crowd-sourced corpus of English Winograd schemas. Notably, *WinoGrande* uses a *gap* token in place of an ambiguous pronoun in each schema, which can be filled by one of two preceding nouns. Based on the chosen noun, the resulting sentence either satisfies or violates commonsense constraints. Schemas are divided into two domains - social and physical. Those belonging to the former category predominantly feature names of individuals (e.g. *Mary* or *Tom*) as fillers, whereas physical samples feature objects or entities (e.g. *vase* or *cat*). Constructing cross-lingual schemas suitable for evaluating translation models requires replacing the *gap* with the ambiguous pronoun *it*, which is not possible for the social domain. Consequently, we focus our attention on the physical subset of *WinoGrande* that contains 19,260 unique samples (9,630 schemas), with each sample representing a single instance of a monolingual, English schema.

4.2.1 Sample Formats

Wino-X includes samples in two formats - one for the evaluation of translation models and another for the evaluation of MLLMs. In both cases the dataset assumes a contrastive evaluation setup (Rios et al., 2017; Gardner et al., 2020), whereby evaluated models are used to rank two minimally different alternatives. Models are scored according to how frequently they rank the correct alternative above the incorrect one.

For the evaluation of NMT models, we replace the gap token with the ambiguous *it* in each sample, and pair the result with two contrastive translations. The translated *it* agrees in gender with a different antecedent in each case. For our investigation, we focus on German, French, and Russian as morphologically rich, high-resource target languages. In the following, we refer to these cross-lingual samples as *MT-Wino-X*.

Evaluation of MLLMs, on the other hand, adopts the *WinoGrande* format. We translate samples without additional modifications, obtaining a set of samples for each target language that we align with their English equivalents. We refer to such multilingual samples as the *LM-Wino-X* set. Example of *MT-Wino-X* entries are provided in Table 4.1, while Table 4.2 contains *LM-Wino-X* entries.

Dataset	Sample
EN-DE	**Source Sentence**: I dusted the **dresser** in the bedroom with a **rag** until **it** was free of dust. **Correct Translation**: Ich staubte die **Kommode** im Schlafzimmer mit einem Lappen ab, bis **sie** staubfrei war. **Incorrect Translation**: Ich staubte die Kommode im Schlafzimmer mit einem **Lappen** ab, bis **er** staubfrei war.
EN-FR	**Source Sentence**: Stacey used the company **credit card** to buy a **plane ticket**, but **it** was declined. **Correct Translation**: Stacey a utilisé **la carte de crédit** de l' entreprise pour acheter un billet d' avion, mais **elle** a été refusée. **Incorrect Translation**: Stacey a utilisé la carte de crédit de l' entreprise pour acheter un **billet d' avion**, mais **il** a été refusé.
EN-RU	**Source Sentence**: Dana could not hang the **artwork** on her **wall** because **it** was too thin. **Correct Translation**: Дана не могла повесить произведение искусства на **стену**, потому что **она** была слишком тонкой. **Incorrect Translation**: Дана не могла повесить **произведение искусства** на стену, потому что **оно** было слишком тонким.

Table 4.1: *MT-Wino-X* examples. Highlighting signifies coreference.

Dataset	Sample
EN-DE	**EN Context**: Adam chose to sleep on a **sofa** instead of a **bed** because _ was much more comfortable. **Correct Filler**: **the sofa** **Incorrect Filler**: **the bed** **DE Context**: Adam entschied sich dafür, auf einem **Sofa** statt auf einem **Bett** zu schlafen, weil _ viel bequemer war. **Correct Filler**: **das Sofa** **Incorrect Filler**: **das Bett**
EN-FR	**EN Context**: The bartender poured the juice from the **blender** into the **cocktail glass** until _ was full. **Correct Filler**: **the glass** **Incorrect Filler**: **the blender** **FR Context**: Le barman versa le jus du **mixeur** dans le **verre** à cocktail jusqu'à ce que _ soit plein. **Correct Filler**: **le verre** **Incorrect Filler**: **le mixeur**
EN-RU	**EN Context:** The man took off the **tank top** and put on the **t-shirt**, because _ was sweaty. **Correct Filler:** **the tank top** **Incorrect Filler:** **the t-shirt** **RU Context:** Мужчина снял **майку** и надел **футболку**, потому что _ была потной. **Correct Filler:** **майка** **Incorrect Filler:** **футболка**

Table 4.2: *LM-Wino-X* examples. Highlighting signifies coreference.

4.2.2 From Monolingual to Multilingual

We find that not all *WinoGrande* samples are suitable for the inclusion in *Wino-X*, as replacing the *gap* with *it* can yield ungrammatical or disfluent sequences. To obtain grammatical sentences after this replacement operation, we exclude *WinoGrande* samples from *Wino-X* if:

- Either referent is animate (e.g. *teacher*, *baker*)
- The *gap* token is part of a compound noun or a noun phrase
- Either referent is a plural noun
- The gap token is modified by an adjective

To improve the quality of our constructed cross-lingual and multilingual schemas, we aim to reduce potential sources of noise by furthermore excluding samples if:

- The translated *it* or gap-filler is not in the nominative case
- Either antecedent denotes an activity (e.g. *singing* or *playing the piano*) (due to issues it presents to morphological analyzers)

Additionally, we use a grammar checker[2] to ensure that the insertion of *it* does not introduce grammatical errors. We furthermore ignore samples where the *gap* is not located in the same sentence as its antecedents, to allow for a fair evaluation of models trained on sentence-level data. To reduce dataset artifacts in *Wino-X*, both instances of a schema are removed if a single one of them is filtered-out.

To obtain contrastive translations, the *gap* token is replaced with one of its fillers (which serve as the antecedents of *it*) before passing the sample through a translation engine. For all target languages, translations are obtained via the Google Translate API[3], due to its relative domain generality. Afterwards, the previously inserted filler is replaced with a pronoun of the same grammatical gender, yielding the final contrastive translation included in *MT-Wino-X*. For *LM-Wino-X* samples, the inserted filler is replaced with the *gap* token.

Following the translation step, we remove *MT-Wino-X* samples where the translated *it* has the **same gender** in both translations, resulting in an undecidable sample.[4] In contrast, for EN-FR and EN-RU portions of *LM-Wino-X*, we only remove samples where translations of both fillers have a **different gender**, as models could otherwise exploit gender agreement of verbs and adjectives to identify the correct filler. Table

[2]LanguageTool: https://pypi.org/project/language-tool-python/

[3]https://cloud.google.com/translate

[4]We use Stanza (Qi et al., 2020): https://stanfordnlp.github.io/stanza/index.html, for the linguistic analysis.

4.3 reports the primary properties of the final dataset, whereas Table 4.4 provides fine-grained, sequence-level statistics.

	MT-Wino-X			***LM-Wino-X***		
	EN-DE	EN-FR	EN-RU	EN-DE	EN-FR	EN-RU
# Schemas	1,887	1,499	1,119	2,917	1,396	743
# Samples	3,774	2,988	2,238	5,834	2,792	1,486

Table 4.3: Composition of the final *Wino-X* dataset.

	MT-Wino-X		***LM-Wino-X***			
	Mean Sentence Length	Mean Trans. Length	Mean EN Context Length	Mean X Context Length	Mean EN Filler Length	Mean X Filler Length
EN-DE	17.8 (2.86)	17.15 (3.1)	17.84 (2.86)	17.16 (3.11)	2.04 (0.19)	2 (0.0)
EN-FR	17.85 (2.9)	20 (3.87)	18.01 (2.86)	20.24 (3.74)	2.02 (0.13)	2 (0.0)
EN-RU	17.73 (2.87)	14.86 (2.99)	18.06 (2.97)	15.34 (3.07)	2.02 (0.14)	2 (0.0)

Table 4.4: Dataset statistics. *X* stands for the language aligned with English for each language pair (DE: German, FR: French, RU: Russian)
. Length is computed in tokens based on Moses-tokenized sentences. Values in parentheses denote standard deviation.

To estimate whether the constructed samples are solvable by humans, we recruited two bilingual raters for each language pair and asked them to select correct translations for a randomly drawn subset of 100 *MT-Wino-X* samples. For EN-DE, mean rater accuracy was 0.84, 0.88 for EN-FR, and 0.87 for EN-RU. Inter-rater agreement was 0.69, 0.75, and 0.77 respectively, according to Cohen's Kappa (Cohen, 1960). For readability, rater instructions are included at the end of this chapter. We note that since the construction of *Wino-X* relies on automated translation and linguistic analysis, the dataset is not completely free of noise. However, its impact on human performance remains within limits.

Like monolingual Winograd schemas, samples included in *Wino-X* represent particularly challenging instances of the CoR problem. However, how models handle such examples is indicative of their general language understanding capabilities. For a computational model to achieve true human parity on the translation task, it must be robust to high levels of semantic ambiguity, given that it poses little difficulty to human raters.

Next, we leverage *Wino-X* for the evaluation of coreference robustness in NMT models and of commonsense knowledge transfer in MLLMs.

4.3 Testing CoR in NMT with Cross-Lingual Schemas

To probe whether NMT models can accurately identify coreference in cases requiring commonsense knowledge, contrastive translations are scored according to perplexity assigned to them by the evaluated model, as in Eqn. 4.1, where X is the source sequence and Y is the candidate translation:

$$PPL(Y|X) = \exp(-\tfrac{1}{|Y|}\textstyle\sum_{i=1}^{|Y|} log_{\phi}(y_i|y_{<i};X)) \tag{4.1}$$

Accuracy is based on the number of instances in which the correct translation is assigned the lower perplexity score.

4.3.1 Experimental Setup

Our evaluation focuses on transformer NMT models (Vaswani et al., 2017), due to their current dominance in the field. For a comprehensive examination of the relationship between model quality and CoR accuracy, we examine three model categories for each language pair: 1. **transformer-BASE** (BASE), 2. **transformer-BIG** (BIG) models distributed as part of the fairseq library[5], and 3. **mBART50**, a multilingual translation model built on top of a pre-trained mBART[6] (Tang et al., 2020). The inclusion of mBART50 follows the assumption that extensive pre-training may endow models with commonsense knowledge, as previously indicated for large-scale monolingual LMs (Bhagavatula et al., 2019; Huang et al., 2019; Sakaguchi et al., 2020).

BASE models are randomly initialized and trained on the WMT news training data[7]. As can be seen from Table 4.5, models differ noticeably in their size, amount of training data, and translation quality.[8] EN-DE and EN-RU models are trained on the concatenation of WMT20 news task data, with *newstest2019* used for development

[5]We use single-best models in place of ensembles for the WMT19 models: https://github.com/pytorch/fairseq/tree/master/examples/translation.

[6]We use the `mbart-large-50-one-to-many-mmt` checkpoint distributed as part of the Transformers library (Wolf et al., 2020).

[7]http://www.statmt.org/wmt[14,20]/translation-task.html

[8]Notably, the EN-FR BIG model had not been trained on back-translated data, unlike its EN-DE and EN-RU counterparts. We elected to tolerate this to allow for easy replication of our experiments using the same openly available, pre-trained NMT models, as well as to reduce the computational overhead and environmental impact incurred by our study.

	EN-DE			EN-FR			EN-RU		
	BASE	BIG	mBART	BASE	BIG	mBART	BASE	BIG	mBART
# Parameters (M)	65.5	363.5	610.9	67.7	313.1	610.9	72.5	317.9	610.9
# Training pairs (M)	39.7	538.7*	42.6	140.6	36	36.8	34.3	162*	13.9
Test BLEU	29.9	36.2	25.6	40.2	41.1	36	21.3	25.5	20.6

Table 4.5: Overview of the evaluated NMT models. Training size estimates were taken from corresponding publications (Ott et al., 2018; Ng et al., 2019; Tang et al., 2020). * denotes inclusion of back-translated parallel data. For mBART50, training size does not include monolingual data used in pre-training. BLEU scores were computed with SacreBLEU (Post, 2018).

and *newstest2020* serving as the text set. For EN-DE, we exclude the *Wiki Titles v2* corpus. The EN-FR model, on the other hand, is trained on the WMT14 news task data, augmented with *ParaCrawl v8*[9]. We use *newstest2013* as the development set and test on *newstest2014*. All data is cleaned by removing sentence pairs with a source-to-target length ratio exceeding 2 or identified as belonging to unrelated languages by langid[10]. We tokenize all datasets using Moses scripts[11] and employ the subword-nmt library[12] (Sennrich et al., 2016) to segment words. Subword segmentation used 32k merge operations and a vocabulary threshold of 50.

Hyper-parameter settings for all the models are provided in Table 4.6. The only exception is the use of tied embeddings for EN-DE and EN-FR, but not EN-RU, as recommended in (Ng et al., 2019). Parameters specific to the transformer architecture (e.g. layer size, number of attention heads) correspond to the BASE configuration in (Vaswani et al., 2017). Other hyper-parameters not covered in Table 4.6 use the default fairseq settings for the 'transformer' architecture. All models were trained on NVIDIA RTX 2080 Ti cards until convergence according to early stopping ($\sim$20 hours each).

4.3.2 Results and Discussion

The results of the contrastive evaluation on the full *MT-Wino-X* dataset are summarized in Table 4.7. All models perform at chance level (a randomly guessing model would

[9] https://paracrawl.eu/
[10] https://github.com/saffsd/langid.py
[11] https://github.com/moses-smt/mosesdecoder
[12] https://github.com/rsennrich/subword-nmt

Hyper-parameter	**Value**
LR	7e-4
LR schedule	*inverse_sqrt*
Batch size	4,096 tokens
# Gradient accumulation steps	6
Optimizer	Adam
Adam betas	0.9, 0.98
Dropout p	0.1
Warm-up updates	4k
Max # Epochs	1k
Validation frequency	5k updates
Early stopping patience	3
Random seed	42

Table 4.6: Hyper-parameters for training **BASE** models.

be 50% accurate), without any observable effect of language pair, model size, training data, or monolingual pre-training.

	EN-DE			**EN-FR**			**EN-RU**		
	BASE	BIG	mBART	BASE	BIG	mBART	BASE	BIG	mBART
Accuracy	0.5032	**0.5093**	0.5048	0.4960	**0.5107**	0.5030	0.4973	0.5009	**0.5049**

Table 4.7: Model performance on the full *MT-Wino-X* dataset. Best results per language pair are in **bold**.

One likely explanation is that models fall back on exploiting surface-level patterns when trying to identify the antecedent of *it*, rather than engaging in deeper language understanding. Such undesirable behaviour is facilitated by dataset biases that models are exposed to during training (Emelin et al., 2020). In their study of coreference, (Stojanovski et al., 2020) indicate that gender and positional biases can influence model behavior. To verify whether this is the case for cross-lingual Winograd schemas, we examine how strongly pronoun gender and the relative antecedent position correlates with model preference. In contrast to prior work, we quantify model bias explicitly as the *absolute effect size of the observed correlation* (i.e. its 'magnitude'), allowing us to directly compare between individual models and language pairs. Correlation significance is computed according to the Mann-Whitney U test (Mann and Whitney, 1947), whereas the effect size is estimated as the Rank Biserial Correlaton (RBC) score[13] (Cureton, 1956).

	EN-DE			EN-FR			EN-RU		
Bias Type	BASE	BIG	mBART	BASE	BIG	mBART	BASE	BIG	mBART
Gender (|RBC|)	<u>0.33</u>	<u>0.27</u>	**0.37**	<u>0.24</u>	0.05	0.05	<u>0.31</u>	**0.48**	**0.44**
Positional (|RBC|)	0.16	0.17	0.14	0.05	0.07	0.15	0.07	0.06	0.05

Table 4.8: Model bias identified for *MT-Wino-X* samples. Higher values indicate a stronger correlation between antecedent features and model choice, and thus a greater bias. All values are statistically significant ($p < .05$). **Bold** values denote a large effect / bias size, <u>underlined</u> values a medium one.

By construction, *Wino-X* is free of gender or positional bias, since the translated *it* is guaranteed to agree with each antecedent in exactly one instance per schema, depending on the trigger word. Thus, each gender and positional category corresponds to the correct choice just as frequently as to the wrong one. As such, preferences of an unbiased NMT system should show no correlation with either property, corresponding to an $|RBC|$ score of 0. As Table 4.8 shows, this is not the case for the evaluated models, as we observe moderate to strong gender bias for EN-DE and EN-RU, but not EN-FR, as well as a trivial, but statistically significant positional bias.

Based on these observations, we can draw several conclusions: 1. While both bias types influence model behaviour, gender bias usually dominates positional bias, 2. Neither extensive pre-training nor multilingual training result in bias reduction for individual language pairs, and 3. The magnitude of biases in CoR is closer associated with training data properties than model properties. We verify the last point by examining the frequency with which different pronoun forms occur in the training data of our BASE models, finding that gender preferences exhibited when scoring *MT-Wino-X* mirror the pronoun gender distribution in the training data. Specifically, for EN-DE, our BASE model strongly favours neutral antecedents, preferring them over the alternative in $\sim$48% of samples, while they represent the correct choice in just $\sim$31% of the dataset. Looking at the training data, we find that translations of *it* are 4.5-12 times

[13] As implemented in the pingouin library (Vallat, 2018).

more likely to have the neutral gender than female and male, respectively. A similar trend can be observed for EN-FR, where *it* is translated as male in ∼63% of samples favoured by the model (which is correct in ∼50% of the dataset), with translations into the male gender being 3.2 times more likely than female in the training data. Male gender is even more dominant for EN-RU, where it is preferred by the model in ∼79% of instances (and correct in just ∼40% of the dataset).

Importantly, the likelihood of *it* being translated as male or female in the EN-RU training data is roughly equal, with translation into male being 1.05 times more likely, yet the absolute frequency of the male pronoun is roughly twice as high compared to the female form. A similar picture emerges for the EN-FR data, where the male pronoun is 3.6 times more frequent than its female analogue, overall. It is difficult to estimate the absolute frequency of the German female pronoun, as it is highly polysemous. Table 4.9 summarizes the corresponding statistics.

	EN-DE			EN-FR		EN-RU		
	Masc.	Fem.	Neut.	Masc.	Fem.	Masc.	Fem.	Neut.
# preferred by model	982 (26.02%)	985 (26.1%)	1,807 (47.88%)	1,893 (63.35%)	1,095 (36.65%)	1,674 (79.37%)	311 (14.75%)	124 (5.88%)
# reference *it* translations	175.6k (6.39%)	473.9k (17.24%)	2.1M (76.38%)	4.7M (77.05%)	1.4M (22.95%)	236.4k (41.9%)	225.6k (40%)	102.2k (18.1%)
# abs. reference occurrences	1.8M	-*	4.5M	21.7M (78.34%)	6M (21.66%)	1M (59.76%)	508.1k (30.36%)	165.4k (9.88%)

Table 4.9: Pronoun frequencies in *MT-Wino-X* translations preferred by BASE models and found in the training data; Fractions of respective total counts are given in parentheses, where possible.

*The German *sie* is highly polysemous and, as such, not included in the absolute counts, since disambiguation via linguistic analysis of ∼10M candidate sentences (e.g. with Stanza) was computationally prohibitive.

Importantly, *absolute* pronoun form frequencies appear to matter more than the likelihood of *it* being translated into a particular gender. This suggests that the frequency prior underlying the models' gender bias is surprisingly simple and, at least partly, based on raw occurrence statistics.

While model reliance on surface-level patterns provides one possible explanation for the challenging nature of *MT-Wino-X*, we also investigate whether models consider trigger terms to be especially salient when translating ambiguous pronouns.

4.3.2.1 Statistical Analysis

To estimate the statistical significance of the correlation between the gender of the translated *it* and model preference, the Mann-Whitney U test combines translations preferred by the model (i.e. those assigned the lower PPL) and those rejected by the model and ranks them according to the numerical ID that corresponds to the gender of the *it* translation (i.e. 1=*masculine*, 2=*feminine*, 3=*neutral*). Subsequently, the U-value is computed according to Eqn. 4.2 - 4.4, where R_1 denotes the sum of ranks of translations preferred by the model and n_1 their total count, while R_2 denotes the sum of ranks of translations rejected by the model and n_2 their respective total count.

$$U = min(U_1, U_2) \tag{4.2}$$

$$U_1 = R_1 - \frac{n_1(n_1+1)}{2} \tag{4.3}$$

$$U_2 = R_2 - \frac{n_2(n_2+1)}{2} \tag{4.4}$$

To obtain the p-values, U-values are subjected to tie correction and normal approximation. Significance of the positional bias is computed following the same procedure, with ranking taking place according to the relative antecedent location.

In order to compute the RBC values, test sentences are divided into two groups - one containing translations that are preferred by the model and another comprised of the rejected translations. Next, all possible pairs are constructed between the two groups, pairing together each translation from one group with all translations in the other. The proportion of pairs f where the pronoun ID of the preferred translation is greater than that of the rejected translation is computed, as well as the proportion of pairs u where the opposite relation holds. The RBC value is obtained according to Eqn. 4.5.

$$RBC = f - u \tag{4.5}$$

As we are only interested in the effect size and not in the direction of the effect, we take its absolute value to signify bias strength. Positional bias is estimated in the same manner.

A common practice for interpreting effect size strength is the adoption of Cohen's benchmark (Cohen, 2013), which posits that the effect size d is large if $d >= 0.8$, medium if $d >= 0.5$, and small if $d >= 0.2$. It is, however, not inherently applicable to the interpretation of RBC, due to its insensitivity to the *base rate* - the size ratio between the two groups denoted by the dichotomous variable, i.e. whether a translation is preferred or rejected by the model. For a detailed discussion, see (McGrath and Meyer, 2006). To apply the aforementioned thresholds to RBC, we use the conversion formula in Eqn. 4.6 (McGrath and Meyer, 2006), where $p1$ and $p2$ represent the proportions of groups described by the dichotomous variable, with $p_1 = p_1 = 0.5$. Within the contrastive evaluation setting, the base rate is guaranteed to equal 1, since for each sample, one translation will be preferred by the model while the other one is rejected.

$$threshold = \frac{d}{\sqrt{d^2 + \frac{1}{p_1 p_2}}} \tag{4.6}$$

The adjusted effect size thresholds are, therefore, as follows: *small* if $d >= 0.1$, *medium* if $d >= 0.24$, and *large* if $d >= 0.37$.

4.3.3 Do Models Recognize Coreference Trigger Words?

For the estimation of salience of individual source words for the translation of *it*, we adopt the *prediction difference* (PD) technique (Li et al., 2019), shown to provide informative explanations of model behaviour by (Li et al., 2020). To apply PD to the study of coreference, we compare the probabilities assigned by the model to the correct *it* translation (w) conditioned on 1. the full source sentence (X) and 2. the source sentence without the trigger term ($X \backslash t$). To 'remove' a trigger word, its embedding is replaced with a zero vector of equal size. Salience is computed according to Eqn. 4.7, as the difference between the two probabilities.[14]

$$Salience(t; w, X) = P(w|X) - P(w|X \backslash t) \tag{4.7}$$

In order to quantify the overall *relative importance* of trigger tokens compared to non-trigger words per model, we compute **importance scores**, defined as the standardised difference between the means of salience score distributions assigned to trigger tokens and words present in both contrastive translations (i.e. non-triggers). Formally, we compute Cohen's D effect size measure, by subtracting the means of the compared distributions μ_T and μ_{NT} and dividing the result by the pooled standard deviation s, as in Eqn. 4.8. Table 4.10 reports the results.

$$D = \frac{\mu_T - \mu_{NT}}{s} \tag{4.8}$$

[14] We average the salience of constituent sub-words for segmented words.

	EN-DE			EN-FR			EN-RU		
	BASE	BIG	mBART	BASE	BIG	mBART	BASE	BIG	mBART
Scores	0.03	0.11*	0.12*	0.16*	0.01	0.2*	0.02	0.35*	0.08*

Table 4.10: Trigger importance. * denotes statistically significant differences according to paired t-tests ($p < .05$).

Across all models and language pairs, importance scores remain low[15] with the difference between salience scores lacking statistical significance in several cases. On the sentence level, this corresponds to models failing to identify trigger words required to establish coreference, as illustrated in Figure 4.2 for the BIG EN-DE model.

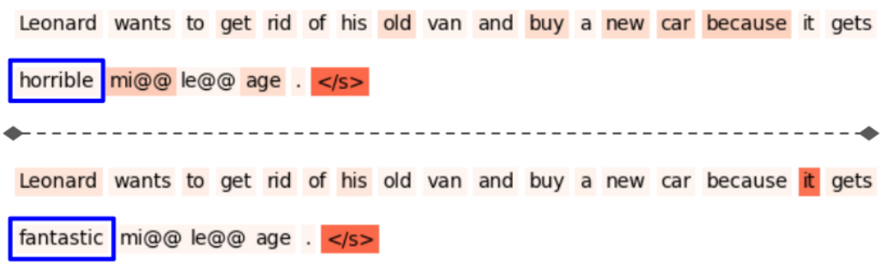

Figure 4.2: Salience maps for two *MT-Wino-X* samples (DE side is omitted for clarity). Words that are more salient for the translation of *it* are highlighted in a deeper shade of orange. Blue frames indicate trigger words that resolve coreference ambiguity.

Therefore, the failure of models to perform well on the *MT-Wino-X* benchmark can be partially attributed to their inherent inability to identify information relevant for establishing coreference.

4.3.4 Improving CoR by Reducing Biases and Enhancing Model Awareness

Finally, we set out to improve coreference resolution in NMT models by addressing undesirable biases and enhancing their ability to detect disambiguating information. Since *MT-Wino-X* is constructed to be unbiased towards antecedent gender, a straightforward way to mitigate model bias is to fine-tune models on a fraction of the dataset, building upon the methodology proposed in (Saunders and Byrne, 2020). Given its limited size, extensive fine-tuning on *MT-Wino-X* is not feasible. However, to investi-

[15] Cohen's D values < 0.5 are considered to be trivial to small (Cohen, 2013).

gate whether bias reduction alone is sufficient to improve CoR that presupposes commonsense knowledge, we conduct a series of few-shot fine-tuning experiments.

For this purpose, we split language-specific *MT-Wino-X* datasets into training, development, and test sets, taking care that both instances belonging to the same schema are assigned to the same split. For all experiments, development and test sets are fixed, containing 200 and 1k samples, respectively. Training set size is varied in increments of 500 up to 2k for EN-DE, 1.5k for EN-FR, and 1k for EN-RU. All models are fine-tuned until convergence as determined by early-stopping. We focus on the BIG models, measuring the effect of increased training size on accuracy and translation quality.

To fine-tune the BASE and BIG NMT models, we use the same settings as provided in 4.6, but set the learning rate to 1e-7, reduce the total batch size to 8 sentence pairs, and forego any warm-up steps. Models are fine-tuned to convergence according to early-stopping, with patience set to 3 validation steps. Validation takes place after each completed training epoch. The optimal LR was determined via grid search over [1e-5, 1e-6, 1e-7].

Settings for fine-tuning mBART are summarized in table 4.11. Hyper-parameters not covered use the default setting in HuggingFace Transformers.

Hyper-parameter	**Value**
LR	1e-5
# Gradient accumulation steps	1
Batch size	16 sentence pairs
Max # Epochs	1k
Validation frequency	1 epoch
Early stopping patience	3
Random seed	42

Table 4.11: Settings used to fine-tune **mBART50** on the *MT-Wino-X* data.

As shown in Figure 4.3, fine-tuning yields slight improvements in accuracy for all language pairs, up to 3.2% for EN-RU. In parallel, we observe a substantial reduction in gender bias in fine-tuned models, using the methodology from §4.3.2. Exposing translation models to 2.5k samples for EN-DE and 1k for EN-RU reduces gender bias by **71%** and **73%**, respectively, from 0.24 to 0.07 and from 0.49 to 0.13.[16] **Still, debiasing alone is not sufficient to substantially increase CoR accuracy**.

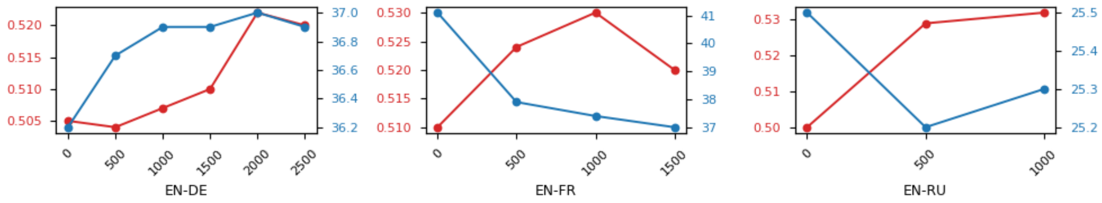

Figure 4.3: Few-shot fine-tuning results on *MT-Wino-X*. Red lines denote accuracy, blue lines correspond to BLEU.

We also note that fine-tuning has a mixed effect on test BLEU which increases for EN-DE but degrades for EN-FR and, to a lesser extent, EN-RU. An analysis of EN-DE test translations before and after fine-tuning shows an increased pronoun coverage for the fine-tuned model, with most pronounced improvements detected for masculine and feminine pronoun forms (Table 4.12), corroborating the quantitative reduction in gender bias.

Source	Feminine	Masculine	Neutral
Reference	340	476	380
Pre-trained	270	410	321
Fine-tuned	**290**	**420**	**326**

Table 4.12: Pronoun frequencies in BIG EN-DE translations, compared to the *newstest2020* reference.

Since bias reduction alone does not suffice to address the unique challenges presented by *MT-Wino-X*, we additionally experiment with equipping translation models with an inductive bias that facilitates accurate pronoun translation. To accomplish this, we define the *Pronoun Penalty* (PP) objective that actively penalizes translation models for assigning higher probability to an incorrect pronoun form during training.[17], so as to encourage models to better utilize trigger words. The objective is defined in Eqn. 4.9, where *CE* is the smoothed cross-entropy loss, λ is the scaling factor, $r \in R$ are correct target pronouns found in the reference translation, and $a \in A$ are alternative,

[16]Initial gender biase values (i.e. 0.24 and 0.49) are recomputed on test sets used in the few-shot experiments. Given the low initial gender bias in EN-FR BIG (0.024), fine-tuning has no noticeable effect.

[17]For simplicity, we only consider singular pronouns in the nominative case, e.g. [*er*, *sie*, *es*] for DE.

incorrect pronoun forms for each correct pronoun (e.g. [*er*, *es*] if the correct German pronoun is *sie*).

$$L(S) = CE(S) + \lambda \sum_{r=1}^{|R|} PP(r) \tag{4.9}$$

$$PP(r) = 1 - \frac{P(r)}{N} + \frac{\max_{a \in A} P(a)}{N} \tag{4.10}$$

$$N = P(r) + \sum_{i=1}^{|A|} P(a_i) \tag{4.11}$$

We fine-tune the BIG models on the largest training set for each language pair with this enhanced objective, and present the results in Table 4.13.[18] The new objective substantially improves accuracy for EN-DE and EN-FR, by 4-7%, while no noticeable difference can be observed for EN-RU. Crucially, the observed improvements correlate with an increase in trigger word importance. Reusing the method introduced in §4.3.3, we find trigger importance increase by a factor of **1.5** for EN-DE and **4.25** for EN-FR compared to models fine-tuned without PP, from 0.12 to 0.18 and 0.04 to 0.17.[19]

Regime	EN-DE	EN-FR	EN-RU
Pre-trained	0.51 (36.2)	0.51 (41.1)	0.5 (25.5)
Fine-tuned	0.52 (36.9)	0.52 (37)	**0.53** (25.3)
+ PP	**0.56** (36.6)	**0.59** (39.4)	**0.53** (25.3)

Table 4.13: *MT-Wino-X* accuracy of models with different training regimes. Test BLEU in parentheses.

Overall, our findings indicate that coreference remains an unsolved challenge in machine translation, especially in cases requiring commonsense knowledge. **While debiasing models leads to improved CoR accuracy, inductive biases that enable models to detect disambiguating information can be more important still.**

4.4 Testing Cross-Lingual Transfer in MLLMs

Having thus probed the capacity and limitations of NMT models for solving cross-lingual *Wino-X* samples, we now turn to MLLMs.

[18] $\lambda = 100$ for all language pairs.

[19] As with bias values, initial trigger importance scores are re-computed on test sets used in few-shot experiments. Fine-tuning has a limited effect on EN-RU which had the highest initial importance scores.

4.4.1 Experimental Setup

Our investigation seeks to answer two questions: 1. *To what extent can MLLMs solve Winograd schemas in different languages?* and 2. *Does commonsense knowledge actively transfer across languages?* Should the latter be the case, it could substantially reduce the need for language-specific commonsense knowledge bases that usually require significant human effort to construct and expand (Speer et al., 2017). Our experiments focus on the XLM-RoBERTa (XLM-R) model introduced in (Conneau et al., 2020). Structurally similar to the decoder of a transformer NMT model, XLM-R is trained on monolingual as well as parallel data covering 100 diverse languages, to induce language-agnostic representations in a shared semantic space. Intuitively, sharing representations across languages should facilitate commonsense knowledge transfer, although it is yet unclear to what extent this holds true for Winograd schemas.

Analogous to our evaluation of NMT models, MLLMs are examined in the contrastive setting. As input, models receive a schema instance containing a *gap*, as depicted in Figure 4.1 (bottom half), which is replaced with a model-specific *<MASK>* token used during pre-training. Conditioned on this input, we compute sentence-level pseudo-perplexities (PPPL) (Salazar et al., 2020) for two completions of the input sequence, each with a different filler that replaces the *<MASK>* token. The completion assigned the lowest PPPL indicates the model's preference towards a specific gap-filler, which informs model accuracy.

4.4.2 Results

As a first step, we measure the zero-shot performance of XLM-R BASE (∼270M parameters) and LARGE (∼550M parameters) models[20] on the full *LM-Wino-X* datasets, summarizing the results in Table 4.14. Accuracy remains comparatively low across the board, with the BASE model scoring close to chance level. On the other hand, the XLM-R LARGE variant substantially outperforms its BASE analogue and demonstrates roughly comparable performance across all examined languages.

4.4.3 Is Monolingual Data Enough for Multilingual CSR?

Of central interest to our investigation is whether fine-tuning models on schema instances in a *primary language*, e.g. EN, also improves CSR in a *transfer language*, e.g.

[20] We use the Transformers library.

	EN-DE		EN-FR		EN-RU	
	EN	DE	EN	FR	EN	RU
BASE	0.53	0.53	0.54	0.53	0.52	0.52
LARGE	**0.62**	**0.61**	**0.63**	**0.6**	**0.62**	**0.59**

Table 4.14: XLM-R accuracy on *LM-Wino-X*. Since dataset composition and size differs between language pairs as detailed in §4.2.2, for EN-X, EN denotes model performance on the EN side of the pair-specific dataset, and X on the aligned non-EN language.

DE, and how this improvement compares to directly fine-tuning the model on the latter. We conduct a series of few-shot experiments to answer this question, while exploring the relationship between cross-lingual commonsense knowledge transfer and the amount of fine-tuning data. Due to its greater efficiency, our investigation is focused on XLM-R BASE[21]. Analogous to experiments in §4.3.4, we split the *LM-Wino-X* data into training, development, and test sets, keeping development and test sizes fixed at 200 and 1k samples, while varying the size of the training set in increments of 500. Instances derived from the same schema are assigned to the same set.

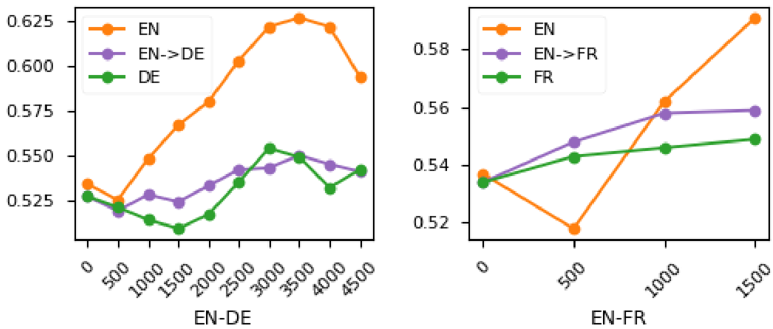

Figure 4.4: Few-shot fine-tuning results on *LM-Wino-X*. EN->X denotes zero-shot knowledge transfer to language X after training the model on EN samples only.

To adopt XLM-R to the studied task, it is fine-tuned on target sequences containing the correct gap-filler with the masked language modeling objective. Models are trained until convergence as determined by early-stopping. We provide the fine-tuning hyper-parameters in Table 4.15. As before, settings not covered in the table correspond to their default values in Huggingface Transformers. Same settings are used for all language pairs. The optimal LR was determined via grid search over [1e-5, 1e-6,

[21]We were unable to train XLM-R LARGE as our hardware could not accommodate its significant size outside of inference.

1e-7].

Hyper-parameter	Value
LR	1e-7
# Gradient accumulation steps	1
Batch size	16 sentence pairs
Max # Epochs	1k
Validation frequency	1 epoch
Early stopping patience	3
Random seed	42

Table 4.15: Settings used to fine-tune **XLM-R** on the LM-Wino-X data.

We treat EN as the primary language and evaluate knowledge transfer toward DE and FR[22], summarizing the results in Figure 4.4. Improved accuracy is observed for all models. However, fine-tuning benefits EN models most as the amount of training samples increases, which may be linked to EN being the dominant language in the XLM-R pre-training corpus (Conneau et al., 2020). More importantly, we can observe a substantial transfer of commonsense knowledge between languages. Models fine-tuned on EN and evaluated on DE / FR often achieve higher accuracy than models directly fine-tuned on the transfer language.

To shed light on commonsense knowledge transfer beyond the few-shot setting, we additionally fine-tune instances of XLM-R on the entirety of WinoGrande and evaluate them on the few-shot test sets.[23] As can be seen from Table 4.16, commonsense knowledge transfer benefits from the increase in training data, with improvements in the transfer languages being roughly half of those observed for the primary language. **This indicates that large-scale, monolingual commonsense resources can significantly contribute towards building models capable of CSR in a wide variety of languages**.

[22] Due to its limited size, EN-RU data is excluded from the few-shot evaluation.
[23] Excluding samples found in each test set from training.

	EN-DE		EN-FR		EN-RU	
	EN	DE	EN	FR	EN	RU
Accuracy FT	0.67	0.60	0.67	0.59	0.65	0.57
Accuracy Δ	14%	7%	13%	6%	13%	5%

Table 4.16: Test accuracy of XLM-R BASE fine-tuned on *WinoGrande* (English only). DE, FR, and RU are transfer languages not seen during fine-tuning. Δ denotes the accuracy increase compared to the first row of Table 4.14.

4.5 Conclusion and Outlook

In this work, we introduced *Wino-X*, a dataset containing cross-lingual and multilingual Winograd schemas. Based on this resource, we showed that NMT models struggle to correctly resolve coreference that presupposes commonsense knowledge, due to over-reliance on dataset artifacts and general inability to detect disambiguating information. We defined methods to quantify biases and trigger word importance in a principled way, and proposed strategies for reducing the former while increasing the latter. For MLLMs, we presented evidence of commonsense knowledge transfer, showing that transferring knowledge from English to another language can lead to similar (or greater) improvements as directly fine-tuning on transfer languages. Overall, our study identifies existing difficulties in cross-lingual CoR and CSR, discusses potential causes, and offers initial ways to mitigate them.

In future work, we intend to further improve the handling of coreference in NMT by reducing undesirable biases and introducing useful ones. For MLLMs, future efforts can be directed towards identifying categories of knowledge that do not benefit from cross-lingual transfer to effectively guide data collection in lower-resourced languages. This category includes commonsense knowledge that varies across cultures and geographic regions, e.g. wedding and funeral customs, where existing work such as (Yin et al., 2022) offers valuable pointers for data acquisition and curation.

Ethical Considerations

Since our work introduces a novel resource, we include a Data Statement (Bender and Friedman, 2018) as a concise overview of its provenance and construction. We hope this will motivate the research community to adopt the dataset for projects relating to cross-lingual natural language understanding by increasing transparency.

A. CURATION RATIONALE: We discuss the filtering criteria applied to *Wino-Grande* samples and their translations in §4.2.2. In enforcing conservative selection criteria, our aim is to ensure grammaticality of the semi-automatically constructed samples and to minimize the percentage of undecidable or disfluent instances.

B. LANGUAGE VARIETY: The collected dataset contains English, German, French, and Russian sentences. English sentences were authored by human crowd-workers, while translations into other languages were obtained from an online translation service. Since (Sakaguchi et al., 2020) do not provide demographics of workers involved in data collection, we cannot report on the dominant variety of English. Due to their origin, translations into DE, FR, and RU are likely to exhibit features of neural translationese (Graham et al., 2020).

C. SPEAKER DEMOGRAPHIC: N/A

D. ANNOTATOR DEMOGRAPHIC: We appropriate this section to summarize the demographics of raters involved in evaluating the dataset quality, as detailed in §4.2.2. Of the 6 annotators involved (two per language pair), all were bilingual speakers with native or native-like proficiency in both English and German / French / Russian. All six were of European origin, between 25-35 years of age, and held a graduate degree. Four of the raters identified as female and two as male.

E. SPEECH SITUATION: The dataset was constructed semi-automatically using scripts distributed in the project's repository. Raters submitted their judgments in the course of a single week and had the opportunity to contact the primary author with clarifying questions.

F. TEXT CHARACTERISTICS: *Wino-X* contains a collection of cross-lingual and multilingual Winograd schemas for the study of coreference resolution and commonsense reasoning in NMT models and MLLMs. Due to the relative simplicity of scenarios described by the schemas, it is highly unlikely for the dataset to have significant ethical implications.

G. RECORDING QUALITY: N/A

H. OTHER: N/A

I. PROVENANCE APPENDIX: According to (Sakaguchi et al., 2020), *WinoGrande* was collected through the Amazon Mechanical Turk (AMT) platform. Workers had to meet a minimum qualification that required 99% approval rate and 5k AMT approvals in total. For composing twin sentences corresponding to a single schema, workers were awarded $0.4. Each collected sample was subsequently validated by three other crowd-workers, with 68% of samples deemed to be valid. For each sentence validation, workers were reimbursed with $0.03. See (Sakaguchi et al., 2020) for a more extensive discussion of *WinoGrande*.

Rater instructions

Once you open the form you were given a link to, you will see a sheet containing ∼100 rows, with each row representing an individual sample for you to annotate. Each row is subdivided into 4 fields: SENTENCE, TRANSLATION_1, TRANSLATION_2, and WHICH TRANSLATION IS BETTER?

Please begin the annotation of each row by first reading the sentence given in the SENTENCE field. Each SENTENCE should contain the English pronoun "it" as well as several nouns. One of the nouns should be identifiable as the referent of "it", i.e. as denoting the object or entity that "it" clearly refers to. For instance, given the SENTENCE "The trophy does not fit into the suitcase because ***it*** *is too small", the bolded* ***it*** *clearly refers to suitcase rather than trophy, since a suitcase can be too small to fit a trophy, but a trophy cannot be too small to fit inside a suitcase.*

TRANSLATION_1 and TRANSLATION_2 provide two alternative, minimally different translations of SENTENCE. The primary difference between both translations is the gender of the pronoun representing the translation of the ambiguous "it" in SENTENCE. Continuing with our running example, TRANSLATION_1 could be "Die Trophäe passt nicht in den Koffer, weil er zu klein ist", while TRANSLATION_2 could be "Die Trophäe passt nicht in den Koffer, weil sie zu klein ist". In TRANSLATION_1, "it" has been translated as the German pronoun er that unambiguously refers to Koffer (corresponding to the English "suitcase"), as both are masculine in gender. On the other hand, in TRANSLATION_2, "it" is translated as the German pronoun sie that unambiguously refers to Trophäe (corresponding to the English "trophy"), as both are feminine in gender. Given that things cannot usually be too small to fit into receptacles, TRANSLATION_1 should be judged as correct, rather than TRANSLATION_2.

When annotating each example, please select the most appropriate option from the drop-down menu in the WHICH TRANSLATION IS BETTER? column. If you think that TRANSLATION_1 is accurate or have a preference towards it (e.g. based on your world knowledge / common sense), please choose "1". If you think that TRANSLATION_2 is accurate or have a preference towards it, please choose "2". If both translations are perfectly equally likely, please choose "BOTH". If the translation quality *is insufficient for you to make a confident judgment, please select "BAD SAMPLE".*

Since the translations were machine-generated, we ask you to be lenient towards translation errors that do not affect the pronoun disambiguation. If the translation is not perfect, e.g. containing odd structure or mistranslated words, but you're still able to identify the correct pronoun translation, please indicate your translation choice, rather than marking the sample as bad.

TRANSLATION_1 and TRANSLATION_2 will always differ as to how "it" is translated, but may have other surface-level differences, as well. As long as both translations convey similar content, we encourage you to ignore any differences other than the translation of "it" for the purpose of your judgments.

4.6 Post-Publication Comments

As discussed in Section 3.5, modern LLMs can be increasingly considered as a valid alternative to traditional NMT models, achieving competitive translation quality. Given their exposure to large quantities of data across a variety of domains and languages, they might additionally be more adept at performing cross-lingual commonsense reasoning within the translation setting, since LLMs have previously been shown to perform well on a variety of monolingual commonsense knowledge benchmarks (Srivastava et al., 2022; Li et al., 2022). To establish if this is indeed the case, we examined whether the proposed *MT-Wino-X* benchmark presents a challenge for the current generation of LLMs. Additionally, we utilized the *LM-Wino-X* portion of the dataset to probe whether more recent LLMs demonstrate an improved ability to resolve coreference ambiguity within a monolingual setting as compared to the XLM model tested in Section 4.4.3. Similarly to the WSD study described in Section 3.5, we utilize the ChatGPT API in all of the conducted experiments.

By utilizing an API, we no longer have access to the LLM's internal representations such as logits and token-wise probabilities. Consequently, we modify the format used to query the model with *MT-Wino-X* samples, re-framing the task as a multiple choice question, as opposed to the original method of comparing model perplexity assigned to contrastive translations. Specifically, we use a prompt containing the task description to prime the model, followed by a total of six training samples corresponding to three distinct Winograd schemas. The prompt is shown in Table 4.17, accompanied by one of the examples. All prompting examples were selected from their respective datasets at random, making sure that each two samples belong to the same schema.

U: Your task is to pick the correct [German / French / Russian] translation of an English sentence out of two alternatives presented to you. The correct translation should fully and accurately capture the meaning of the English source sentence. Please identify the preferred translation in your response without providing any additional information. Here are some examples of what this should look like: English sentence: "The woman looked for a different vase for the bouquet because it was too large." Translation 1: "Die Frau suchte nach einer anderen Vase für den Blumenstrauß, weil sie zu groß war." Translation 2: "Die Frau suchte nach einer anderen Vase für den Blumenstrauß, weil er zu groß war." Which is the better translation?"
S: Translation 2

Table 4.17: Basic prompt used to elicit ChatGPT's judgments regarding the correctness of contrastive Winograd schema instance translations sampled from *MT-Wino-X*, without intermediate reasoning steps. The LLM was prompted with four examples in total. U = User, S = System.

A similar approach was taken for the evaluation of *LM-Wino-X*, where the choice presented to the model is between two possible gap fillers, rather than two potential translations. Again, the model was provided with a prompt and six examples representing three different schemas to prime it for the task, as shown in Table 4.18. While the context sentence and gap fillers could be either English, German, French, or Russian, depending on the subset of *LM-Wino-X*, the instruction text remained English throughout to ensure consistency.

Here, too, it must be mentioned that valid model replies could not be obtained for all queries due to issues such as API timeouts and model restrictions, as had been the case for WSD queries in Section 3.5. The fraction of successful queries for each subset of *Wino-X* is documented in Table 4.19. It is important to note that the exclusion of samples used for prompting and those for which no model judgment could be obtained precludes the observed LLM accuracy from being directly compared to the NMT model results in Table 4.8. Nonetheless, given that model replies were collected for the vast majority of the challenge set, the conducted evaluation offers useful insights into the LLM's CoR capabilities and limitations.

As Table 4.20 indicates, ChatGPT scores poorly on the *MT-Wino-X* task, falling slightly behind the significantly smaller, debiased NMT models described in Section

U: Your task is to pick the correct [English / German / French / Russian] term out of two alternatives presented to you in order to fill the gap in the provided context sentence. The gap to be filled is represented by the word <GAP>. The completed sentence should be grammatical, fluent, and make sense within the context of an everyday situation. Please identify the preferred gap filler in your response without providing any additional information. Here are some examples of what this should look like: Context sentence: "The woman looked for a different vase for the bouquet because <GAP> was too large." Filler 1: "the vase" Filler 2: "the bouquet" Which is the more appropriate gap filler?"
S: Filler 2

Table 4.18: Basic prompt used to elicit ChatGPT's judgments regarding the correctness of contrastive Winograd schema instance gap fillers sampled from *LM-Wino-X*, without intermediate reasoning steps. The LLM was prompted with four examples in total. U = User, S = System.

4.3.4. By contrast, the LLM substantially outperforms XLM-LARGE on the *LM-Wino-X* task, with improvements on non-English languages ranging from 5.56% for the Russian side of the EN-RU split to 16.9% for the French side of the EN-FR split. On the English parts of the challenge set, accuracy gains of up to 19.22% can be observed. This discrepancy indicates that ChatGPT is, in principle, capable of performing commonsense reasoning required to solve cross-lingual Winograd schemas, but is unable to leverage this competency for *MT-Wino-X*[24].

CoT prompting, previously discussed in 3.5, is one potential pathway towards facilitating the required reasoning steps within the *MT-Wino-X* setting. To investigate if this prompting strategy improves the LLM's performance on the benchmark, a CoT prompt and four prompting examples corresponding to two Winograd schemas were manually constructed for each translation direction, based on randomly selected schemas that were subsequently excluded from the evaluation. Table 4.21 reproduces the prompt accompanied by one of the prompting examples (each API call to the model included all four examples).

As before, we report the fraction of valid replies obtained via he API for each

[24]While the generally good performance of ChatGPT on the *LM-Wino-X* challenge sets suggests that *Wino-X* may have been included in the LLM's training data, its subpar accuracy on the *MT-Wino-X* split suggests otherwise. Unfortunately, as ChatGPT's training data is not not publicly available, it is impossible to verify whether *Wino-X* samples were seen by the model during the pre-training stage.

Dataset	# API failures (%)	# Invalid replies (%)	# Valid replies (%)
MT-Wino-X EN-DE	17 (0.5%)	20 (0.5%)	3,750 (99%)
MT-Wino-X EN-FR	20 (0.65%)	20 (0.65%)	2,944 (98.7%)
MT-Wino-X EN-RU	6 (0.3%)	10 (0.5%)	2,228 (99.2%)
LM-Wino-X EN | DE	83 (1.4%) | 95 (1.6%)	5 (0.09%) | 5 (0.09%)	5,747 (98.58%) | 5,734 (98.35%)
LM-Wino-X EN | FR	17 (0.61%) | 37 (1.33%)	2 (0.07%) | 7 (0.25%)	2,774 (99.46%) | 2,749 (98.57%)
LM-Wino-X EN | RU	23 (1.55%) | 0 (0%)	0 (0%) | 1 (0.07%)	1,448 (97.64%) | 1,470 (99.12%)

Table 4.19: Fractions of successful and unsuccessful *Wino-X* queries addressed to ChatGPT via the web API. *Invalid replies* refers to cases where the model did not produce a reply that corresponded to one of the two possible choices.

MT-Wino-X EN-DE	MT-Wino-X EN-FR	MT-Wino-X EN-RU
55.63%	56.93%	52.71%
LM-Wino-X EN | DE	**LM-Wino-X EN | FR**	**LM-Wino-X EN | RU**
79.55% | 75.93%	80.03% | 76.90%	81.22% | 64.56%

Table 4.20: ChatGPT accuracy for the different subsets of the *Wino-X* benchmark.

dataset in Table 4.22, thus ascertaining that the amount of rejected samples remains small. Similar to the mixed results on the WSD task, CoT prompting does not improve ChatGPT's accuracy on *MT-Wino-X*, as can be seen from the results summarized in Table 4.23. Specifically, using the CoT prompt causes a substantial drop in accuracy, of 3.39%, for the EN-DE direction, which suggests that the extended prompt is not only ineffective, but can be actively detrimental to the LLM's cross-lingual commonsense CoR ability. It is possible that the inclusion of target language terms into the prompt confounds the model or, alternatively, the close integration of the CoR and translation steps within the prompt is complicating the reasoning steps, rendering them ineffective. To investigate whether this is indeed the case, an alternative prompting strategy is introduced, motivated by the high accuracy achieved by the LLM on *LM-Wino-X*, which we refer to as Chain-of-Tasks (CoTA).

In contrast to the CoT prompt, the CoTA prompt introduces monolingual CoR on the (English) source side as an explicit intermediate step to be completed by the model prior to the selection of the preferred translation. The so constructed prompt is shown in Table 4.24, followed by one of the four prompting examples that were created analogously to the CoT examples. In essence, the prompt converts the English

User: Your task is to pick the correct [German / French / Russian] translation of an English sentence out of two alternatives presented to you. Prior to making your selection, you should reason about the differences in the presented translations and why one of them is preferable to the other. The correct translation should fully and accurately capture the meaning of the English source sentence. Here are some examples of what this should look like: English sentence: "The woman looked for a different vase for the bouquet because it was too small.", Translation 1: "Die Frau suchte nach einer anderen Vase für den Blumenstrauß, weil sie zu groß war." Translation 2: "Die Frau suchte nach einer anderen Vase für den Blumenstrauß, weil er zu groß war." Which is the better translation?"
System: In translation 1, the English pronoun "it" is translated as "sie" which agrees in its grammatical gender with the referent noun "Vase" that corresponds to the English noun "vase" as both are feminine in German. On the other hand, in translation 2 "it" is translated as "er" which agrees in its grammatical gender with the masculine German noun "Blumenstrauß" corresponding to the English noun "bouquet". It is more likely that the woman would look for a different vase if the vase does not fit the bouquet due to the small size of the vase. Therefore, it follows that the correct translation is translation 1.

Table 4.21: CoT prompt used to elicit ChatGPT's judgments regarding the correctness of contrastive Winograd schema instance translations sampled from *MT-Wino-X*. Four-shot prompting was used for all samples.

side of the *MT-Wino-X* samples into the *LM-Wino-X* format, since ChatGPT obtained high accuracies on the latter. The decision to situate the CoR step on the English side was motivated by the consistently higher English CoR accuracy values observed in the *LM-Wino-X* experiments (see Table 4.20). A reasonable assumption is that this is due to the vast majority of ChatGPT's training data being English (OpenAI, 2023).

The inclusion of an intermediate task into the model prompt leads to notable improvements over the basic prompt and CoT prompting alike, as evidenced by the results presented in Table 4.26. The corresponding fraction of successful API calls is reported in Table 4.25. The extent of the observed improvements varies between the evaluated translation directions, with EN-RU scoring lower than EN-DE and EN-FR, similarly to previous experiments, which may be in part due to the difference in target language script (Hendy et al., 2023). The accuracy gains furthermore suggest that CoTA prompting is able to leverage some of the commonsense reasoning processes employed by the LLM to resolve lexical ambiguity in the *LM-Wino-X* samples, although not fully, as the

Dataset	# API failures (%)	# Invalid replies (%)	# Valid replies (%)
MT-Wino-X EN-DE	38 (0.98%)	74 (1.96%)	3,658 (97.03%)
MT-Wino-X EN-FR	36 (1.21%)	38 (1.27%)	2,909 (97.52%)
MT-Wino-X EN-RU	38 (1.7%)	30 (1.34%)	2,169 (96.96%)

Table 4.22: Fractions of successful and unsuccessful *Wino-X* CoT queries addressed to ChatGPT via the web API.

MT-Wino-X EN-DE	MT-Wino-X EN-FR	MT-Wino-X EN-RU
52.24% (-3.39)	57.3% (+0.37)	52.88% (+0.17)

Table 4.23: ChatGPT accuracy for the different subsets of the *Wino-X* benchmark with CoT prompting. Values in parentheses denote changes compared to prompting without CoT.

MT-Wino-X task still appears to pose a greater challenge to ChatGPT. Overall, CoTA prompting, or more generally the dissection of a complex cross-lingual task into simpler, partly monolingual sub-tasks constitutes a valid strategy for obtaining meaningful translations that are consistent with commonsense assumptions about the world.

In order to illustrate the effectiveness as well as shortcomings of CoTA prompting, Tables 4.27 - 4.29 show model responses across all considered prompting methods for selected *MT-Wino-X* samples. While CoTA prompting is, on average, more effective than basic and CoT prompting, cases can be identified for all three strategies where one is successful while the other two fail. As such, Table 4.27 shows instances where the basic prompt alone elicits the expected model judgment, while Table 4.28 does the same for CoT prompting and Table 4.29 for CoTA prompts. One interesting takeaway from these examples is that additional reasoning steps are not always beneficial and may, in fact, confuse the model. Furthermore, the reasoning presented by the model may be logically sound but grounded in an incorrect reading of the input sentence, thus leading to CoR errors. In other cases, the reasoning itself is faulty, ultimately causing the model to draw incorrect conclusions about the referent of the ambiguous pronoun. On the whole, the introduction of explicit reasoning steps can demonstrably improve model performance and utility, but only if done in a manner that can be effectively leveraged by the LLM. As such, one promising avenue for future work could be investigating the types of information that are either beneficial or detrimental to

User: Your task is to pick the correct [German / French / Russian] translation of an English sentence out of two alternatives presented to you. Prior to making your selection, you should identify the ambiguous pronoun in the English sentence that can refer to different entities, and reason about which of the possible referents is the correct one, as that will determine the choice of the correct translation. Here are some examples of what this should look like:
English sentence:
"The woman looked for a different vase for the bouquet because it was too small.",
Translation 1:
"Die Frau suchte nach einer anderen Vase für den Blumenstrauß, weil sie zu groß war."
Translation 2:
"Die Frau suchte nach einer anderen Vase für den Blumenstrauß, weil er zu groß war."
Which is the better translation?"

System: If we replace the ambiguous pronoun "it" in the English sentence with a gap, then the two possible gap fillers would be ether "vase" or "bouquet". The correct filler in this case is "vase", since a vase that is too small would not be able to fit a bouquet. Since the German translation of "vase" is the feminine Noun "Vase" that agrees grammatically in its gender with the German pronoun "sie", it follows that the correct translation is translation 1.

Table 4.24: CoTA prompt used to elicit ChatGPT's judgments regarding the correctness of contrastive Winograd schema instance translations sampled from *MT-Wino-X*. Four-shot prompting was used for all samples.

Dataset	# API failures (%)	# Invalid replies (%)	# Valid replies (%)
MT-Wino-X EN-DE	30 (0.8%)	3 (0.08%)	3,740 (99.13%)
MT-Wino-X EN-FR	24 (0.8%)	2 (0.07%)	2,961 (99.13%)
MT-Wino-X EN-RU	18 (0.8%)	2 (0.09%)	2,217 (99.11%)

Table 4.25: Fractions of successful and unsuccessful *Wino-X* CoTA queries addressed to ChatGPT via the web API.

MT-Wino-X EN-DE	MT-Wino-X EN-FR	MT-Wino-X EN-RU
60.72% (+5.09 / +8.48)	65.05% (+8.12 / +7.75)	57.92% (+5.21 / +5.04)

Table 4.26: ChatGPT accuracy for the different subsets of the *Wino-X* benchmark with CoTA prompting. Values in parentheses denote changes compared to basic prompting / CoT prompting.

LLM behaviour if incorporated into prompts, especially for complex multi-lingual or multi-modal tasks.

In light of the findings presented in this section, it is clear that *Wino-X* remains a valuable and useful resource following the advent of LLM-powered machine translation. The cross-lingual *MT-Wino-X* portion of the challenge set, in particular, appears to pose a veritable challenge to contemporary LLMs, since it requires complex multi-stage reasoning that can, in part, be facilitated by prompts targeting the relevant reasoning steps. Nonetheless. despite demonstrable improvements in model accuracy enabled by CoTA prompting, LLMs – as represented by ChatGPT – fall short of human-like performance on translations that incorporate commonsense knowledge.

Follow-Up Work

Findings and resources presented in (Emelin and Sennrich, 2021) have found fruitful application in subsequent research. Particularly worth noting is the extensive use of the *Wino-X* benchmark in (Yehudai et al., 2023) for the purpose of comparing the efficacy of monolingual coreference resolvers with that of NMT models. As part of their investigation, the authors find CoR that presupposes commonsense reasoning to be especially difficult for both model classes, thereby showcasing the enduring utility and broad applicability of the benchmark. The central finding of the work, partly enabled by the evaluation conducted on *Wino-X*, is that integrating labels obtained from monolingual resolvers into the NMT pipeline as a form of direct supervision markedly improves CoR in the generated translations. A contrasting application of the *Wino-X* data is its inclusion in ACES, a compehensive evaluation suite for machine translation quality metrics (Amrhein et al., 2022). In particular, the included samples were either directly adopted or derived from *Wino-X* and used to estimate the ability of evaluated metrics to adequately capture the capacity of machine translation models for commonsense CoR, thereby widening the utility of the original dataset.

Source / Translation **1** / Translation **2** / **Basic Prompt** reply / **CoT** reply / **CoTA** reply
S: The syrup irritated his throat, but the pill made it feel better, because it was so soothing. **T1**: Der Sirup reizte seinen Hals, aber durch die Pille fühlte es sich besser an, weil er so beruhigend war. **T2**: Der Sirup reizte seinen Hals, aber durch die Pille fühlte es sich besser an, weil sie so beruhigend war. **BP**: Translation 2 is better because "sie" refers to "Pille" which is feminine. **CoT**: In translation 1, the English pronoun "it" is translated as "es" which agrees in its grammatical gender with the referent noun "Hals" (throat) that corresponds to the English word "throat" as both are masculine in German. On the other hand, in translation 2 "it" is translated as "sie" which agrees in its grammatical gender with the feminine noun "Pille" (pill) corresponding to "pill" in English. Since the pronoun "it" refers to the throat (Hals) here, which is masculine in German, translation 1 is the better choice. **CoTA**: If we replace the ambiguous pronoun "it" in the English sentence with a gap, then the two possible gap fillers would be either "his throat" or "the syrup". To determine which is the correct filler, we need to consider the logic of the sentence. The syrup is mentioned as the cause of irritation, so it makes sense that the pill would make "his throat" feel better. Therefore, "his throat" is the correct interpretation of the pronoun "it" in this context. Since the German translation of "his throat" is "seinen Hals", which is masculine, and the German pronoun "er" agrees grammatically with masculine nouns, the correct translation is Translation 1.
S: Bob went to the vehicle dealer to trade his truck for a motorcycle, because it was new. **T1**: Bob ging zum Fahrzeughändler, um seinen LKW gegen ein Motorrad zu tauschen, weil es neu war. **T2**: Bob ging zum Fahrzeughändler, um seinen LKW gegen ein Motorrad zu tauschen, weil er neu war. **BP**: Translation 1 is better. **CoT**: In translation 2, the English pronoun "it" is translated as "er" which agrees in its grammatical gender with the masculine German noun "LKW" corresponding to the English noun "truck". On the other hand, in translation 1 "it" is translated as "es" which agrees in its grammatical gender with the neuter German noun "Motorrad" corresponding to the English noun "motorcycle". Since the referent noun in the source sentence is the truck, it is more appropriate to use translation 1 where "it" is translated as "er". Therefore, the correct translation is Translation 2. **CoTA**: In this sentence, the ambiguous pronoun "it" refers to either "his truck" or "a motorcycle" in the English sentence. To determine which translation is better, letś analyze the meaning of the sentence. If we interpret "it" as referring to "his truck," the sentence means that Bob went to the vehicle dealer to trade his truck for a motorcycle because his truck was new. In this case, the truck is the focus of the sentence, and translation 1 would be the better option because the pronoun "er" (masculine pronoun) agrees with the German word "LKW" (masculine noun) which means "truck." However, if we interpret "it" as referring to "a motorcycle," the sentence means that Bob went to the vehicle dealer to trade his truck for a motorcycle because the motorcycle was new. In this case, the focus shifts to the motorcycle, and translation 1 would be the better option because the pronoun "es" (neuter pronoun) agrees with the German word "Motorrad" (neuter noun) which means "motorcycle." Based on the context, it is more likely that the sentence is emphasizing the condition of Bobś truck. Therefore, translation 2 is the better translation.
S: The afternoon breeze was less dangerous than the wind because it was a lot stronger. **T1**: Die Nachmittagsbrise war weniger gefährlich als der Wind, weil sie viel stärker war. **T2**: Die Nachmittagsbrise war weniger gefährlich als der Wind, weil er viel stärker war. **BP**: Translation 2 is the better translation. **CoT**: In this case, both translations are grammatically correct and convey the same meaning. However, when it comes to the pronoun "it" referring to the afternoon breeze, translation 1 ("sie") is more preferable because it agrees in gender with the feminine German noun "Nachmittagsbrise," whereas translation 2 ("er") uses a masculine pronoun that does not align with the gender of the noun. Therefore, the better translation is translation 1. **CoTA**: In this sentence, the ambiguous pronoun is "it". If we replace "it" with a gap, the two possible gap fillers are either "afternoon breeze" or "wind". To determine the correct referent, we need to consider the comparison made in the sentence. The sentence states that the afternoon breeze was less dangerous than the wind because it was a lot stronger. This implies that the afternoon breeze is the one that is a lot stronger. Therefore, the correct translation is Translation 1.

Table 4.27: Examples of Chat-GPT replies elicited through different prompting methods. Here, the basic prompt produced the correct response, as opposed to CoT and CoTA. Correct CoR steps are coloured **blue**, incorrect ones are colored **red**.

Source / Translation 1 / Translation 2 / Basic Prompt reply / CoT reply / CoTA reply
S: The piercing on the tongue got infected but the ear didn't because it was dirtier. **T1**: Das Piercing auf der Zunge wurde infiziert, das Ohr jedoch nicht, weil es schmutziger war. **T2**: Das Piercing auf der Zunge wurde infiziert, das Ohr jedoch nicht, weil sie schmutziger war. **BP**: Translation 1 is better. **CoT**: In translation 1, the English pronoun "it" is translated as "es" which agrees in its grammatical gender with the neuter noun "Ohr" (ear) and not with the feminine noun "Zunge" (tongue). This does not correctly reflect the meaning of the English sentence. In translation 2, the pronoun "it" is translated as "sie" which agrees in gender with the feminine noun "Zunge" (tongue). This correctly reflects the meaning of the English sentence, stating that the tongue was dirtier and therefore the piercing on the tongue got infected while the ear remained unaffected. Therefore, the better translation is Translation 2. **CoTA**: In this sentence, the ambiguous pronoun "it" can refer to either "ear" or "piercing." Since the ear is the subject of the second clause ("the ear didnt́ [get infected]"), it would be grammatically incorrect for the pronoun to refer back to the ear in the first clause. Therefore, the pronoun "it" refers to the piercing, which got infected. Based on this reasoning, the better translation would be Translation 1.
S: After abandoning the hostage, the criminal hid in the alley after leaving the basement, because it was safe. **T1**: Nachdem er die Geisel verlassen hatte, versteckte sich der Verbrecher in der Gasse, nachdem er den Keller verlassen hatte, weil sie sicher war. **T2**: Nachdem er die Geisel verlassen hatte, versteckte sich der Verbrecher in der Gasse, nachdem er den Keller verlassen hatte, weil er sicher war. **BP**: Translation 2 is better. **CoT**: In this sentence, the English pronoun "it" refers to the location "the alley". In translation 1, "it" is translated as "sie" which agrees with the feminine noun "Gasse" (alley). In translation 2, "it" is translated as "er" which agrees with the masculine noun "Keller" (basement). However, the intended meaning is that the alley was safe, not the basement. Therefore, the correct translation is translation 1. **CoTA**: If we replace the ambiguous pronoun "it" in the English sentence with a gap, then the two possible gap fillers would be either "alley" or "basement". To determine the correct translation, we need to consider which location is being referred to as safe. The sentence states that the criminal hid in the alley after leaving the basement because "it" was safe. In this context, "safe" refers to the alley, not the basement. Translation 1 incorrectly uses the feminine pronoun "sie" to refer to the alley, while Translation 2 correctly uses the masculine pronoun "er" to refer to the alley. Therefore, the better translation is Translation 2.
S: The man decided to immigrate from his homeland to the foreign country, because it was experiencing a war. **T1**: Der Mann beschloss, aus seiner Heimat ins Ausland auszuwandern, weil sie einen Krieg erlebte. **T2**: Der Mann beschloss, aus seiner Heimat ins Ausland auszuwandern, weil es einen Krieg erlebte. **BP**: Translation 2 is better. **CoT**: Both translations are grammatically correct and convey the general meaning of the English sentence. However, when referring to the reason for the manś decision to immigrate, translation 1 uses the pronoun "sie" which refers to "Heimat" (homeland), while translation 2 uses the pronoun "es" which refers to "Ausland" (foreign country). Considering the context of the sentence, the pronoun "sie" in translation 1 seems more appropriate because the homeland is the entity experiencing the war, not the foreign country. Therefore, the better translation is Translation 1. **CoTA**: If we replace the ambiguous pronoun "it" in the English sentence with a gap, then the two possible gap fillers would be ether "homeland" or "foreign country". The correct filler in this case is "foreign country", since the manś decision to immigrate is based on the fact that the foreign country is experiencing a war, not his own homeland. Since the German translation of "foreign country" is the neuter noun "Ausland" that agrees grammatically in its gender with the German pronoun "es", it follows that the correct translation is translation 2.

Table 4.28: Examples of Chat-GPT replies elicited through different prompting methods. Here, the CoT strategy produced the correct response, as opposed to the basic prompt and CoTA. Correct CoR steps are coloured **blue**, incorrect ones are colored **red**.

Source / Translation **1** / Translation **2** / **Basic Prompt** reply / **CoT** reply / **CoTA** reply
S: Per the contract, the business had to pay the bank thousands of dollars, as it had lost the lawsuit. **T1**: Laut Vertrag musste das Unternehmen der Bank Tausende von Dollar zahlen, da sie die Klage verloren hatte. **T2**: Laut Vertrag musste das Unternehmen der Bank Tausende von Dollar zahlen, da es die Klage verloren hatte. **BP**: Translation 1 is correct. **CoT**: In this case, both translations are quite similar. The only difference is the choice of the pronoun for "it". In Translation 1, "it" is translated as "sie", which refers to the feminine noun "das Unternehmen" (the company). In Translation 2, "it" is translated as "es", which refers to the neuter noun "Vertrag" (contract). Considering the context of the sentence, it is more logical for "it" to refer to the company rather than the contract. The company is the subject of the sentence and it is the one that lost the lawsuit. Therefore, the better translation in this case is Translation 1. **CoTA**: The ambiguous pronoun in the English sentence is "it". If we replace the pronoun with a gap, the two possible gap fillers would be "business" or "bank". To determine the correct filler, we need to identify the subject that wlost the lawsuit. In this case, it is the business that has lost the lawsuit, not the bank. Therefore, the correct translation is Translation 2.
S: James cannot write the essay within the time given to him because it is short. **T1**: James kann den Aufsatz nicht innerhalb der ihm vorgegebenen Zeit schreiben, da er kurz ist. **T2**: James kann den Aufsatz nicht innerhalb der ihm gegebenen Zeit schreiben, weil sie kurz ist. **BP**: Both translations are acceptable, but Translation 1 is slightly better. **CoT**: In this instance, both translations are grammatically correct and convey the intended meaning. However, there is a subtle difference between the two translations in terms of the cause-effect relationship. Translation 1 uses the conjunction "da" which implies a cause-and-effect relationship. It suggests that James cannot write the essay within the given time because the essay itself is short. Translation 2 uses the conjunction "weil" which also indicates a cause-and-effect relationship. However, it suggests that James cannot write the essay within the given time because the time itself is short. Considering the context and intended meaning of the sentence, it is more likely that the intended meaning is that James cannot write the essay within the given time because the essay itself is short. Therefore, Translation 1 is the better translation. **CoTA**: The ambiguous pronoun in the English sentence is "it". If we replace "it" with a gap, the two possible gap fillers are "essay" and "time". In this case, the correct filler is "time" since James cannot write the essay within the given time. Since the German translation for "time" is the feminine noun "Zeit" that agrees grammatically with the German pronoun "sie", the correct translation is Translation 2.
S: The man threw away the cheese and only ate the bread because it was moldy. **T1**: Der Mann warf den Käse weg und aß nur das Brot, weil er schimmelig war. **T2**: Der Mann warf den Käse weg und aß nur das Brot, weil es schimmelig war. **BP**: Both translations are grammatically correct, but Translation 2 is slightly better as it uses a more common phrasing "because it was moldy." **CoT**: Both translations correctly convey the meaning of the English sentence. However, when translating the pronoun "it" referring to the bread being moldy, translation 2 uses the neuter pronoun "es," while translation 1 uses the masculine pronoun "er." In this case, the bread is a neuter noun in German, which would make translation 2 more accurate. Therefore, the better translation is translation 2. **CoTA**: If we replace the ambiguous pronoun "it" in the English sentence with a gap, then the two possible gap fillers would be either "cheese" or "bread". The correct filler in this case is "cheese", since it is the cheese that was moldy and not the bread. Since the German translation of "cheese" is the masculine noun "Käse" that agrees grammatically in its gender with the German pronoun "er", it follows that the correct translation is translation 1.

Table 4.29: Examples of Chat-GPT replies elicited through different prompting methods. Here, the CoTA strategy produced the correct response, as opposed to the basic prompt and CoT. Correct CoR steps are coloured **blue**, incorrect ones are colored **red**.

Chapter 5

Direct Access to Embedding Features Improves Lexical Reasoning

To get the right word in the right place is a rare achievement.

Mark Twain, *Mark Twain's Letters, Volume 2*

Abstract: The transformer is a state-of-the-art neural translation model that uses attention to **iteratively refine lexical representations with information drawn from the surrounding context**. Lexical features are fed into the first layer and propagated through a deep network of hidden layers. We argue that **the need to represent and propagate lexical features in each layer limits the model's capacity for learning and representing other information relevant to the task**. To alleviate this bottleneck, **we introduce gated shortcut connections** between the embedding layer and each subsequent layer within the encoder and decoder. This enables the model to access relevant lexical content dynamically, without expending limited resources on storing it within intermediate states. We show that the proposed modification yields **consistent improvements over a baseline transformer** on standard WMT translation tasks in 5 translation directions (0.9 BLEU on average) and **reduces the amount of lexical information passed along the hidden layers**. We furthermore evaluate different ways to integrate lexical connections into the transformer architecture and present ablation experiments exploring the effect of proposed shortcuts on model behavior.[1]

[1]This section is based on work previously published at WMT 2019 (Emelin et al., 2019). Experimental codebase is available at https://github.com/demelin/transformer_lexical_shortcuts.

5.1 Introduction

Since it was first proposed, the transformer model (Vaswani et al., 2017) has quickly established itself as a popular choice for neural machine translation, where it has been found to deliver state-of-the-art results on various translation tasks (Bojar et al., 2018). Its success can be attributed to the model's high parallelizability allowing for significantly faster training compared to recurrent neural networks (Chen et al., 2018), superior ability to perform lexical disambiguation, and capacity for capturing long-distance dependencies on par with existing alternatives (Tang et al., 2018).

Recently, several studies have investigated the nature of features encoded within individual layers of neural translation models (Belinkov et al., 2017a,b). One central finding reported in this body of work is that, in recurrent architectures, different layers prioritize different information types. As such, lower layers appear to predominantly perform morphological and syntactic processing, whereas semantic features reach their highest concentration towards the top of the layer stack. One necessary consequence of this distributed learning is that different types of information encoded within input representations received by the translation model have to be transported to the layers specialized in exploiting them.

Within the transformer encoder and decoder alike, information exchange proceeds in a strictly sequential manner, whereby each layer attends over the output of the immediately preceding layer, complemented by a shallow residual connection. For input features to be successfully propagated to the uppermost layers, the translation model must therefore store them in its intermediate representations until they can be processed. By retaining lexical content, the model is unable to leverage its full representational capacity for learning new information from other sources, such as the surrounding sentence context. We refer to this limitation as the *representation bottleneck*.

To alleviate this bottleneck, we propose extending the standard transformer architecture with lexical shortcuts which connect the embedding layer with each subsequent self-attention sub-layer in both encoder and decoder. The shortcuts are defined as gated skip connections, allowing the model to access relevant lexical information at any point, instead of propagating it upwards from the embedding layer along the hidden states.

We evaluate the resulting model's performance on multiple language pairs and varying corpus sizes, showing a consistent improvement in translation quality over the unmodified transformer baseline. Moreover, we examine the distribution of lexical information across the hidden layers of the transformer model in its standard configuration and with added shortcut connections. The presented experiments provide quantitative evidence for the presence of a representation bottleneck in the standard transformer and its reduction following the integration of lexical shortcuts.

While our experimental efforts are centered around the transformer, the proposed components are compatible with other multi-layer NMT architectures.

The contributions of our work are as follows:

1. We propose the use of **lexical shortcuts** as a simple strategy for **alleviating the representation bottleneck** in NMT models.

2. We demonstrate **significant improvements in translation quality** across multiple language pairs as a result of equipping the transformer with lexical shortcut connections.

3. We conduct a series of **ablation studies**, showing that shortcuts are best applied to the self-attention mechanism in both encoder and decoder.

4. We report a positive impact of our modification on the model's ability to perform **word sense disambiguation**.

5.2 Proposed Method

5.2.1 Background: The transformer

As defined in (Vaswani et al., 2017), the transformer is comprised of two sub-networks, the encoder and the decoder. The encoder converts the received source language sentence into a sequence of continuous representations containing translation-relevant features. The decoder, on the other hand, generates the target language sequence, whereby each translation step is conditioned on the encoder's output as well as the translation prefix produced up to that point.

Both encoder and decoder are composed of a series of identical layers. Each encoder layer contains two sub-layers: A self-attention mechanism and a position-wise fully connected feed-forward network. Within the decoder, each layer is extended with a third sub-layer responsible for attending over the encoder's output. In each case, the attention mechanism is implemented as multi-head, scaled dot-product attention,

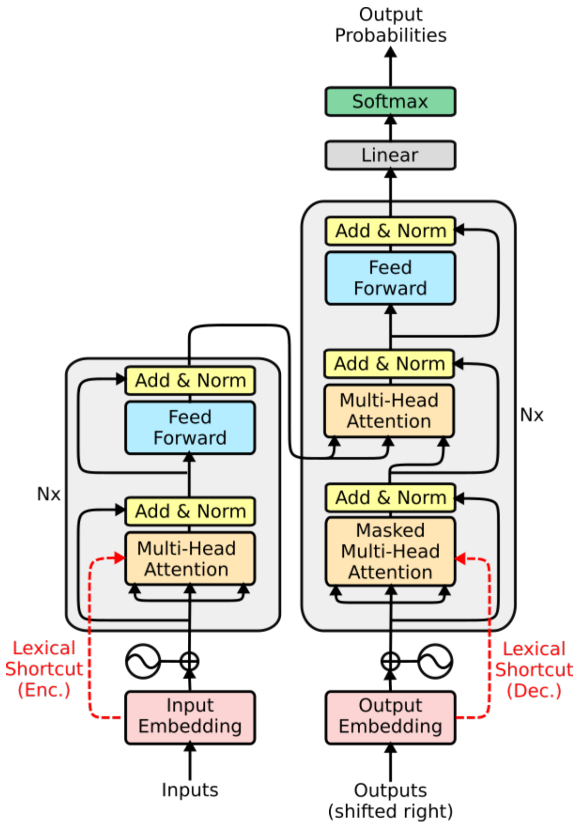

Figure 5.1: Integration of lexical shortcut connections into the overall transformer architecture.

which allows the model to simultaneously consider different context sub-spaces. Additionally, residual connections between layer inputs and outputs are employed to aid with signal propagation.

In order for the dot-product attention mechanism to be effective, its inputs first have to be projected into a common representation sub-space. This is accomplished by multiplying the input arrays H^S and H^T by one of the three weight matrices K, V, and Q, as shown in Eqn. 5.1 - 5.3, producing attention keys, values, and queries, respectively. In case of multi-head attention, each head is assigned its own set of keys, values, and queries with the associated learned projection weights.

$$Q = W^Q H^S \tag{5.1}$$

$$K = W^K H^T \tag{5.2}$$

$$V = W^V H^T \tag{5.3}$$

In case of encoder-to-decoder attention, H^T corresponds to the final encoder states, whereas H^S is the context vector generated by the preceding self-attention sub-layer.

For self-attention, on the other hand, all three operations are given the output of the preceding layer as their input. Eqn. 5.4 defines attention as a function over the projected representations.

$$Attention(Q,K,V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V \tag{5.4}$$

To prevent the magnitude of the pre-softmax dot-product from becoming too large, it is divided by the square root of the total key dimensionality d_k. Finally, the translated sequence is obtained by feeding the output of the decoder through a softmax activation function and sampling from the produced distribution over target language tokens.

5.2.2 Lexical shortcuts

Given that the attention mechanism represents the primary means of establishing parameterized connections between the different layers within the transformer, it is well suited for the re-introduction of lexical content. We achieve this by adding gated connections between the embedding layer and each subsequent self-attention sub-layer within the encoder and the decoder, as shown in Figure 5.1.

To ensure that lexical features are compatible with the learned hidden representations, the retrieved embeddings are projected into the appropriate latent space, by multiplying them with the layer-specific weight matrices $W_l^{K^{SC}}$ and $W_l^{V^{SC}}$. We account for the potentially variable importance of lexical features by equipping each added connection with a binary gate inspired by the Gated Recurrent Unit (Cho et al., 2014). Functionally, our lexical shortcuts are equivalent to highway connections of (Srivastava et al., 2015) that span an arbitrary number of intermediate layers.

$$K_l^{SC} = W_l^{K^{SC}} E \tag{5.5}$$

$$V_l^{SC} = W_l^{V^{SC}} E \tag{5.6}$$

$$K_l = W_l^K H_{l-1} \tag{5.7}$$

$$V_l = W_l^V H_{l-1} \tag{5.8}$$

$$r_l^K = \text{sigmoid}(K_l^{SC} + K_l + b_l^K) \tag{5.9}$$

$$r_l^V = \text{sigmoid}(V_l^{SC} + V_l + b_l^V) \tag{5.10}$$

$$K_l' = r_l^K \odot K_l^{SC} + (1 - r_l^K) \odot K_l \tag{5.11}$$

$$V_l' = r_l^V \odot V_l^{SC} + (1 - r_l^V) \odot V_l \tag{5.12}$$

After situating the outputs of the immediately preceding layer H_{l-1} and the embeddings E within a shared representation space (Eqn. 5.5 - 5.8), the relevance of lexical information for the current attention step is estimated by comparing lexical and latent features, followed by the addition of a bias term b (Eqn. 5.9 - 5.10). The respective attention key arrays are denoted as K_l^{SC} and K_l, while V_l^{SC} and V_l represent the corresponding value arrays. The result is fed through a sigmoid function to obtain the lexical relevance weight r, used to calculate the weighted sum of the two sets of features (Eqn. 5.11 - 5.12), where $\odot$ denotes element-wise multiplication. Next, the key and value arrays K_l' and V_l' are passed to the multi-head attention function as defined in Eqn. 5.4, replacing the original K_l and V_l.

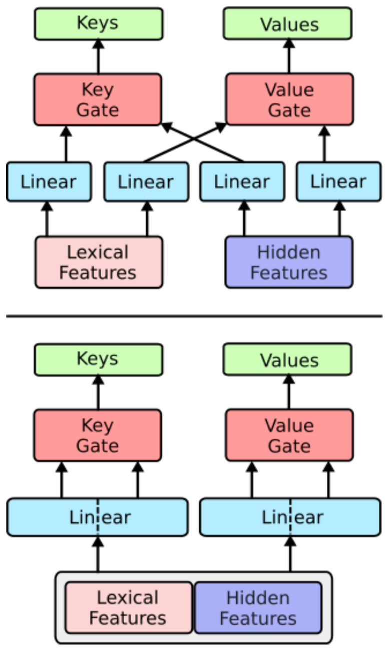

Figure 5.2: Modified attention inputs. Top: lexical shortcuts, bottom: lexical shortcuts + feature-fusion. Dashed lines denote splits along the feature dimension.

In an alternative formulation of the model, referred to as *feature-fusion* from here on, we concatenate E and H_{l-1} before the initial linear projection, splitting the result in two halves along the feature dimension and leaving the rest of the shortcut definition unchanged[2]. This reduces Eqn. 5.5 - 5.8 to Eqn. 5.13 - 5.14, and enables the model to select relevant information by directly inter-relating lexical and hidden features. As such, both K_l^{SC} and K_l encode a mixture of embedding and hidden features, as do the corresponding value arrays. While this arguably diminishes the contribution of

[2]The feature-fusion mechanism is therefore based on the same principle as the Gated Linear Unit (Dauphin et al., 2017), while utilizing a more expressive gating function.

the gating mechanism towards feature selection, preliminary experiments have shown that replacing gated shortcuts with gate-less residual connections (He et al., 2016) produces models that fail to converge, characterized by poor training and validation performance. For illustration purposes, figure 5.2 depicts the modified computation path of the lexically-enriched attention key and value vectors.

$$K_l^{SC}, K_l = W_l^K [E; H_{l-1}] \tag{5.13}$$

$$V_l^{SC}, V_l = W_l^V [E; H_{l-1}] \tag{5.14}$$

Other than the immediate accessibility of lexical information, one potential benefit afforded by the introduced shortcuts is the improved gradient flow during back-propagation. As noted in (Huang et al., 2017), the addition of skip connections between individual layers of a deep neural network results in an implicit *deep supervision* effect (Lee et al., 2015), which aids the training process. In case of our modified transformer, this corresponds to the embedding layer receiving its learning signal from the model's overall optimization objective as well as from each layer it is connected to, making the model easier to train.

5.3 Experiments

5.3.1 Training

To evaluate the efficacy of the proposed approach, we re-implement the transformer model and extend it by applying lexical shortcuts to each self-attention layer in the encoder and decoder. The majority of our experiments is conducted using the transformer-BASE configuration, with the number of encoder and decoder layers set to 6 each, embedding and attention dimensionality to 512, number of attention heads to 8, and feed-forward sub-layer dimensionality to 2,048. We tie the encoder embedding table with the decoder embedding table and the pre-softmax projection matrix to speed up training, following (Press and Wolf, 2017). All trained models are optimized using Adam (Kingma and Ba, 2014) adhering to the learning rate schedule described in (Vaswani et al., 2017). We set the number of warm-up steps to 4,000 for the baseline model, increasing it to 6,000 and 8,000 when adding lexical shortcuts and feature-fusion, respectively, so as to accommodate the increase in parameter size.

We also evaluate the effect of lexical shortcuts on the larger transformer-BIG model, limiting this set of experiments to EN→DE due to computational constraints. Here, the baseline model employs 16 attention heads, with attention, embedding, and feed-forward dimensions doubled to 1,024, 1,024, and 4,096. Warm-up period for all big models is 16,000 steps. For our probing experiments, the classifiers used are simple feed-forward networks with a single hidden layer consisting of 512 units, dropout (Srivastava et al., 2014) with $p = 0.5$, and a ReLU non-linearity. In all presented experiments, we employ beam search during decoding, with beam size set to 16.

Model	**# Parameters**	**Words / sec.**
transformer-BASE	65,166k	29,698
+ lexical shortcuts	71,470k	26,423
+ feature-fusion	84,053k	23,601
transformer-BIG	218,413k	10,215
+ feature-fusion	293,935k	6,769

Table 5.1: Model size and training speed of the compared transformer variants.

All models are trained concurrently on four Nvidia P100 Tesla GPUs using synchronous data parallelization. Delayed optimization (Saunders et al., 2018) is employed to simulate batch sizes of 25,000 tokens, to be consistent with (Vaswani et al., 2017). Each transformer-BASE model is trained for a total of 150,000 updates, while our transformer-BIG experiments are stopped after 300,000 updates. Validation is performed every 4,000 steps, as is check-pointing. Training base models takes $\sim$43 hours, while the addition of shortcut connections increases training time up to $\sim$46 hours ($\sim$50 hours with feature-fusion). Table 5.1 details the differences in parameter size and training speed for the different transformer configurations. Parameters are given in thousands, while speed is averaged over the entire training duration.

Validation-BLEU is calculated on a reference which we pre- and post-process following the same steps as for the models' inputs and outputs. All reported test-BLEU scores were obtained by averaging the final 5 checkpoints for transformer-BASE and final 16 for transformer-BIG.

5.3.2 Data

We investigate the potential benefits of lexical shortcuts on 5 WMT translation tasks: German $\rightarrow$ English (DE$\rightarrow$EN), English $\rightarrow$ German (EN$\rightarrow$DE), English $\rightarrow$ Russian (EN$\rightarrow$RU), English $\rightarrow$ Czech (EN$\rightarrow$CS), and English $\rightarrow$ Finnish (EN$\rightarrow$FI). Our choice is motivated by the differences in size of the training corpora as well as by the typological diversity of the target languages.

To make our findings comparable to related work, we train EN↔DE models on the WMT14 news translation data which encompasses ∼4.5M sentence pairs. EN→RU models are trained on the WMT17 version of the news translation task, consisting of ∼24.8M sentence pairs. For EN→CS and EN→FI, we use the respective WMT18 parallel training corpora, with the former containing ∼50.4M and the latter ∼3.2M sentence pairs. We do not employ backtranslated data in any of our experiments to further facilitate comparisons to existing work.

We tokenize, clean, and truecase each training corpus using scripts from the Moses toolkit[3], and apply byte-pair encoding (Sennrich et al., 2016) to counteract the open vocabulary issue. Cleaning is skipped for validation and test sets. For EN↔DE and EN→RU we limit the number of BPE merge operations to 32,000 and set the vocabulary threshold to 50. For EN→CS and EN→FI, the number of merge operations is set to 89,500 with a vocabulary threshold of 50, following (Haddow et al., 2018)[4]. In each case, the BPE vocabulary is learned jointly over the source and target language, which necessitated an additional transliteration step for the pre-processing of Russian data[5].

Throughout training, model performance is validated on *newstest2013* for EN↔DE, *newstest2016* for EN→RU, and on *newstest2017* for EN→CS and EN→FI. Final model performance is reported on multiple tests sets from the news domain for each direction.

5.3.3 Translation performance

The results of our translation experiments are summarized in Tables 5.2-5.3. To ensure their comparability, we evaluate translation quality using sacreBLEU (Post, 2018). As such, our baseline performance diverges from that reported in (Vaswani et al., 2017). We address this by evaluating our EN→DE models using the scoring script from the *tensor2tensor* toolkit[6] (Vaswani et al., 2018) on the tokenized model output, and list the corresponding BLEU scores in the first column of Table 5.2.

Our evaluation shows that the introduction of lexical shortcuts consistently improves translation quality of the transformer model across different test-sets and lan-

[3] https://github.com/moses-smt/mosesdecoder
[4] We do not use synthetic data, which makes our results not directly comparable to theirs.
[5] We used *Lingua Translit* for this purpose: https://metacpan.org/release/Lingua-Translit
[6] https://github.com/tensorflow/tensor2tensor/blob/master/tensor2tensor/utils/get_ende_bleu.sh

Model	nt2014 tokenized	sacreBLEU nt2014	nt2015	nt2016	nt2017	nt2018	test mean
transformer-BASE	27.3	25.8	28.5	33.2	27.3	40.4	31.0
+ lexical shortcuts	27.6	26.1	29.5	33.3	27.5	41.1	31.5
+ feature-fusion	**28.3**	**26.8**	**29.9**	**34.0**	**27.7**	**41.6**	**32.0**
transformer-BIG	28.7	27.2	30.1	**34.0**	28.1	41.3	32.1
+ lexical shortcuts + feature-fusion	**29.4**	**27.8**	**30.3**	33.2	**28.4**	41.3	**32.2**

Table 5.2: BLEU scores for the EN→DE news translation task; *nt* = *newstest*.

Model	DE→EN		EN→RU		EN→CS		EN→FI	
	nt2014	nt2017	nt2017	nt2018	nt2015	nt2018	nt2015	nt2018
transformer-BASE	31.1	32.3	27.9	24.2	23.4	21.1	18.7	14.0
+ lexical shortcuts	31.3	32.3	28.4	24.9	24.1	21.4	19.5	14.5
+ feature-fusion	**31.7**	**32.9**	**28.9**	**25.3**	**24.3**	**21.6**	**19.8**	**14.8**

Table 5.3: Effect of lexical shortcuts on translation quality for different language pairs, as measured by sacreBLEU; *nt* = *newstest*.

guage pairs, outperforming transformer-BASE by 0.5 BLEU on average. With feature-fusion, we see even stronger improvements, yielding total performance gains over transformer-BASE of up to 1.4 BLEU for EN→DE (averaging to 1.0), and 0.8 BLEU on average for the other 4 translation directions. We furthermore observe that the relative improvements from the addition of lexical shortcuts are substantially smaller for transformer-BIG compared to transformer-BASE. One potential explanation for this drop in efficacy is the increased size of latent representations the wider model is able to learn, which we discuss in section 5.4.1.

Furthermore, it is worth noting that transformer-BASE equipped with lexical connections performs comparably to the standard transformer-BIG, despite containing fewer than half of its parameters and being only marginally slower to train than our unmodified transformer-BASE implementation.

Concerning the examined language pairs, the average increase in BLEU is highest for EN→RU (1.1 BLEU) and lowest for DE→EN (0.6 BLEU). A potential explanation for why this is the case could be the difference in language topology. Of all target languages we consider, English has the least complex morphological system where individual words carry little inflectional information, which stands in stark contrast to a highly inflectional language with a flexible word order such as Russian. It is plausible that lexical shortcuts are especially important for translation directions where the target language is morphologically rich and where the surrounding context is essential to accurately predicting a word's case and agreement. With the proposed shortcuts in place, the transformer has more capacity for modeling such context information.

To investigate the role of lexical connections within the transformer, we conduct a thorough examination of our models' internal representations and learning behaviour. The following analysis is based on models utilizing lexical shortcuts with feature-fusion, due to its superior performance.

5.4 Analysis

5.4.1 Representation bottleneck

The proposed approach is motivated by the hypothesis that the transformer retains lexical features within its individual layers, which limits its capacity for learning and representing other types of relevant information. Direct connections to the embedding layer alleviate this by providing the model with access to lexical features at each processing step, reducing the need for propagating them along hidden states. To investigate whether this is indeed the case, we perform a probing study, where we estimate the amount of lexical content present within each encoder and decoder state.

We examine the internal representations learned by our models by modifying the probing technique introduced in (Belinkov et al., 2017a). Specifically, we train a separate lexical classifier for each layer of a frozen translation model. Each classifier receives hidden states extracted from the respective transformer layer[7] and is tasked with reconstructing the sub-word corresponding to the position of each hidden state. Encoder-specific classifiers learn to reconstruct sub-words in the source sentence, whereas classifiers trained on decoder states are trained to reconstruct target sub-words.

The accuracy of each layer-specific classifier on a withheld test set is assumed to be indicative of the lexical content encoded by the corresponding transformer layer. We

[7]We treat the output of the feed-forward sub-layer as that layer's hidden state.

expect classification accuracy to be high if the evaluated representations predominantly store information propagated upwards from the embeddings at the same position and to decrease proportionally to the amount of information drawn from the surrounding sentence context. Figure 5.3 offers side-by-side comparisons of the accuracy scores obtained for each layer of the base transformer and its variant equipped with lexical shortcut connections.

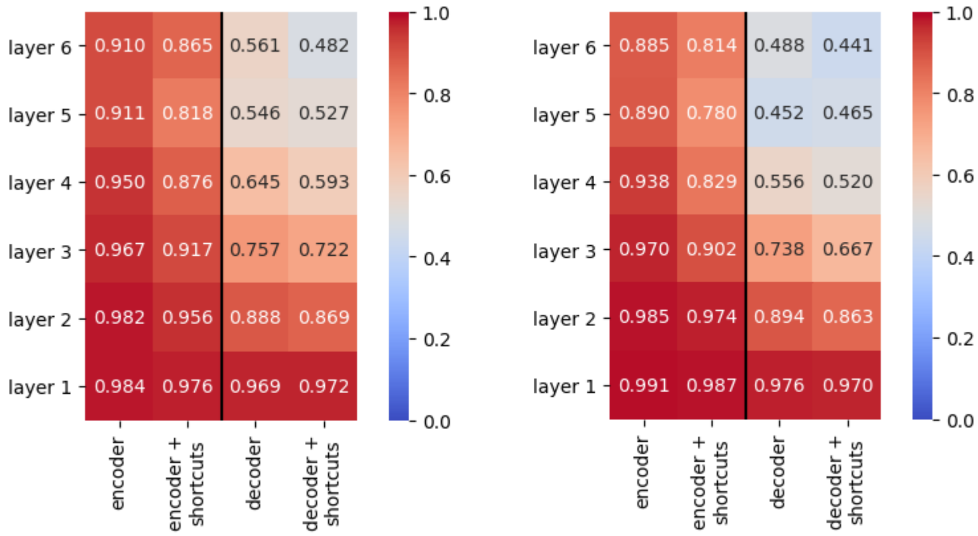

Figure 5.3: **Left**: Layer-wise lexical probe accuracy measured on transformer-BASE for EN→DE (*newstest2014*); **Right**: Layer-wise lexical probe accuracy measured on transformer-BASE for EN→RU (*newstest2017*)

Based on the observed classification results, it appears that immediate access to lexical information does indeed alleviate the representation bottleneck by reducing the extent to which (sub-)word-level content is retained across encoder and decoder layers. By introducing shortcut connections, we effectively reduce the amount of lexical information the model retains within its intermediate states, thereby increasing its capacity for exploiting sentence context. The effect is consistent across multiple language pairs, supporting its generality. Additionally, to examine whether lexical retention depends on the specific properties of the input tokens, we track classification accuracy conditioned on part-of-speech tags and sub-word frequencies. Our working hypotheses are that 1.) processing of low-frequency tokens may benefit more from the introduction of lexical shortcuts due to the model being less practiced at integrating their embedding features into its latent representations, and 2.) processing of tokens belonging to different POS categories may be impacted by the presence of lexical shortcuts to varying degree according to differences in their relative, average semantic complexity.

For the former feature category, we first parse our test-sets with *TreeTagger* (Schmid, 1999), projecting tags onto the constituent sub-words of each annotated word. For frequency-based evaluation, we divide sub-words into ten equally-sized frequency bins, with bin 1 containing the least frequent sub-words and bin 10 containing the most frequent ones. We do not observe any immediately obvious, significant effects of either POS or frequency on the retention of lexical features, and are thus unable to establish any empirical support for either of the aforementioned hypotheses. While classification accuracy is notably low for infrequent sub-words, this can be attributed to the limited occurrence of the corresponding transformer states in the classifier's training data. Evaluation for EN→DE models is done on *newstest2014*, while *newstest2017* is used for EN→RU models. Figures 5.6 and 5.7 illustrate representative results for the frequency-based classification. Examples of accuracy scores conditioned on POS tags are visualized in Figure 5.8.[8] For readability, the figures are included at the end of this chapter.

We also investigated the activation patterns of the lexical shortcut gates. However, despite their essential status for the successful training of transformer variants equipped with lexical connections, we were unable to discern any distinct patterns in the activations of the individual gates, which tend to prioritize lexical and hidden features to an equal degree regardless of training progress or (sub-)word characteristics.

Another observation arising from the probing analysis is that the decoder retains fewer lexical features beyond its initial layers than the encoder. This may be due to the decoder having to represent information it receives from the encoder in addition to target-side content, necessitating a lower rate of lexical feature retention. Even so, by adding shortcut connections we can increase the dissimilarity between the embedding layer and the subsequent layers of the decoder, indicating a noticeable reduction in the retention and propagation of lexical features along the decoder's hidden states.

A similar trend can be observed when evaluating layer similarity directly, which we accomplish by calculating the cosine similarity between the embeddings and the hidden states of each transformer layer. Echoing our findings so far, the addition of lexical shortcuts reduces layer similarity relative to the baseline transformer for both encoder and decoder. Cosine similarity scores between the embedding layer and each successive layer in transformer-BASE and its variant equipped with lexical shortcuts are summarized in Figure 5.5, provided at the end of this chapter.

[8] Additional plots reporting results for other model states as well as the EN→RU translation direction can be found in Emelin et al. (2019).

Overall, the presented analysis supports the existence of a representation bottleneck in NMT models as one potential explanation for the efficacy of the proposed lexical shortcut connections.

5.4.2 Model size

Model	newstest 2017	newstest 2018	test mean
transformer-SMALL	25.2	37.0	28.6
+ lexical shortcuts	25.7	38.0	29.3
+ feature-fusion	**25.7**	**38.5**	**29.6**

Table 5.4: sacreBLEU scores for small EN→DE models; *test mean* denotes the average of all test sets in table 5.2.

Next, we investigate the interaction between the number of model parameters and improvements in translation quality afforded by the proposed lexical connections. Following up on findings presented in section 5.3.1, we hypothesize that the benefit of lexical shortcuts diminishes once the model's capacity is sufficiently large. To establish whether this decline in effectiveness is gradual, we scale down the standard transformer, halving the size of its embeddings, hidden states, and feed-forward sublayers. Table 5.4 shows that, on average, quality improvements are comparable for the small and standard transformer (1.0 BLEU for both), which is in contrast to our observations for transformer-BIG. One explanation is that given sufficient capacity, the model is capable of accommodating the upward propagation of lexical features without having to neglect other sources of information. However, as long as the model's representational capacity is within certain limits, the effect of lexical shortcuts remains comparable across a range of model sizes. With this in mind, the exact interaction between model scale and the types of information encoded in its hidden states remains to be fully explored. We leave a more fine-grained examination of this relationship to future research.

5.4.3 Shortcut variants

Until now, we focused on applying shortcuts to self-attention as a natural re-entry point for lexical content. However, previous studies suggest that providing the decoder with direct access to source sentences can improve translation adequacy, by conditioning decoding on relevant source tokens (Kuang et al., 2018; Nguyen and Chiang, 2018).

To investigate whether the proposed method can confer a similar benefit to the transformer, we apply lexical shortcuts to decoder-to-encoder attention, replacing or adding to shortcuts feeding into self-attention. Formally, this equates to fixing E to E^{enc} in Eqn. 5.5 - 5.6 and can be regarded as a variant of source-side bridging proposed by (Kuang et al., 2018). As Table 5.5 shows, while integrating shortcut connections into the decoder-to-encoder attention improves upon the base transformer, the improvement is smaller than when we modify self-attention. Interestingly, combining both methods yields worse translation quality than either one does in isolation, indicating that the decoder is unable to effectively consolidate information from both source and target embeddings, which negatively impacts its learned latent representations. We therefore conclude that lexical shortcuts are most beneficial to self-attention.

Model	**newstest 2017**	**newstest 2018**	**test mean**
transformer-BASE	27.3	40.4	31.0
+ self-attn. shortcuts	**27.7**	**41.6**	**32.0**
dec-to-enc shortcuts	27.6	40.7	31.5
+ self-attn. shortcuts	27.7	40.5	31.4
non-lexical shortcuts	27.1	40.6	31.3

Table 5.5: sacreBLEU for shortcut variants of EN→DE models; *test mean* denotes the average of all test sets in table 5.2.

A related question is whether the encoder and decoder benefit from the addition of lexical shortcuts to self-attention equally. We explore this by disabling shortcuts in either sub-network and comparing the so obtained translation models to one with intact connections. Figure 5.4 illustrates that best translation performance is obtained by enabling shortcuts in both encoder and decoder. This also improves training stability, as compared to the decoder-only ablated model. The latter may be explained by our use of tied embeddings which receive a stronger training signal from shortcut connections due to *deep supervision*, as this may bias learned embeddings against the sub-network lacking improved lexical connectivity.

While adding shortcuts improves translation quality, it is not obvious whether this

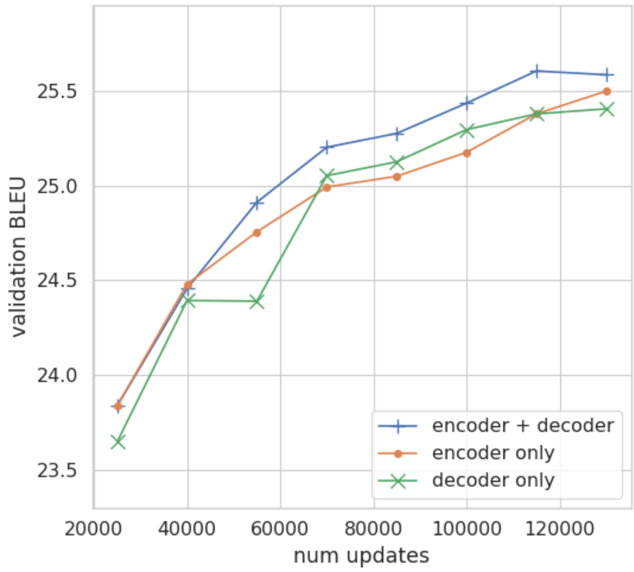

Figure 5.4: Effect of disabling shortcuts in either sub-network on validation BLEU.

is predominantly due to improved accessibility of lexical content, rather than increased connectivity between network layers, as suggested in (Dou et al., 2018). To isolate the importance of lexical information, we equip the transformer with non-lexical shortcuts connecting each layer n to layer $n-2$, e.g. layer 6 to layer 4.[9] As a result, the number of added connections and parameters is kept identical to lexical shortcuts, whereas lexical accessibility is disabled, allowing for minimal comparison between the two configurations. Test BLEU reported in Table 5.5 suggests that while non-lexical shortcuts improve over the baseline model, they perform noticeably worse than lexical connections. Therefore, the increase in translation quality associated with lexical shortcuts is not solely attributable to a better signal flow or the increased number of trainable parameters.

5.4.4 Word-sense disambiguation

Beyond the effects of lexical shortcuts on the transformer's learning dynamics, we are interested in how widening the representation bottleneck affects the properties of the produced translations. One challenging problem in translation which intuitively should benefit from the model's increased capacity for learning information drawn from sentence context is word-sense disambiguation.

We examine whether the addition of lexical shortcuts aids disambiguation by evaluating our trained DE→EN models on the *ContraWSD* corpus (Rios et al., 2017). The

[9] The first layer is connected to the embedding layer, as there is no further antecedent.

contrastive dataset is constructed by paring source sentences with multiple translations, varying the translated sense of selected source nouns between translation candidates. A competent model is expected to assign a higher probability to the translation hypothesis containing the appropriate word-sense.

While the standard transformer offers a strong baseline for the disambiguation task, we nonetheless observe improvements after adding direct connections to the embedding layers. Specifically, our baseline model reaches an accuracy of 88.8%, which improves to **89.5%** with lexical shortcuts.

5.5 Conclusion

In this chapter, we have proposed a simple yet effective method for widening the representation bottleneck in the transformer by introducing lexical shortcuts. Our modified models achieve up to 1.4 BLEU (0.9 BLEU on average) improvement on 5 standard WMT datasets, at a small cost in computing time and model size. Our analysis suggests that lexical connections are useful to both encoder and decoder, and remain effective when included in smaller models. Moreover, the addition of shortcuts noticeably reduces the similarity of hidden states to the initial embeddings, indicating that dynamic lexical access aids the network in learning novel, diverse information. We also performed ablation studies comparing different shortcut variants and demonstrated that one effect of lexical shortcuts is an improved WSD capability.

The presented findings offer new insights into the nature of information encoded by the transformer layers, supporting the iterative refinement view of feature learning. In future work, we intend to explore other ways to better our understanding of the refinement process and to help translation models learn more diverse and meaningful internal representations.

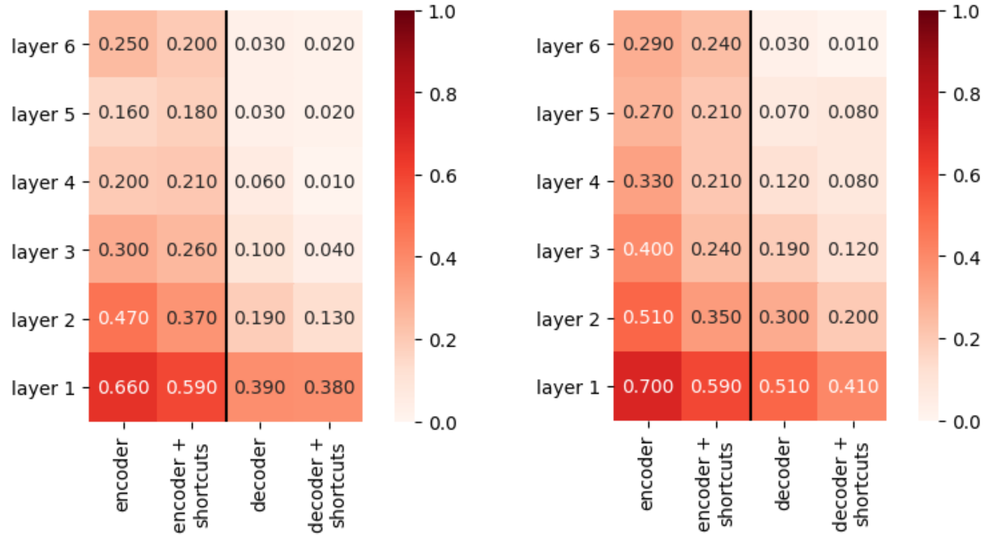

Figure 5.5: **Left**: Cosine similarity measured on transformer-BASE for EN→DE (newstest2014); **Right**: Cosine similarity measured on transformer-BASE for EN→RU (newstest2017)

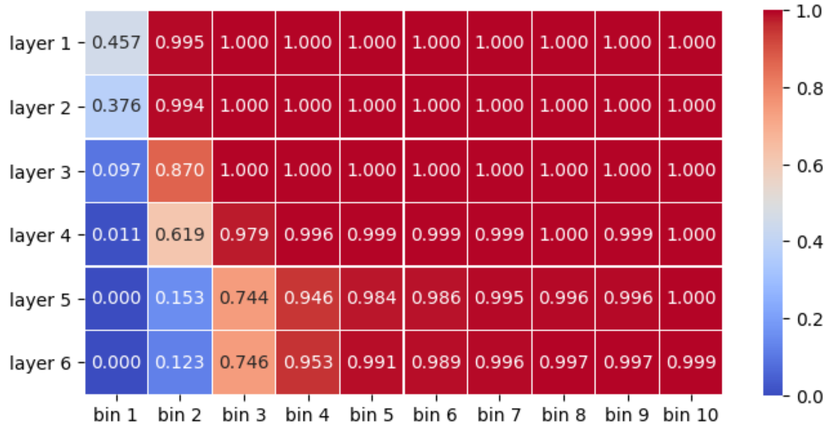

Figure 5.6: Frequency-based classification accuracy on states from the EN→DE encoder.

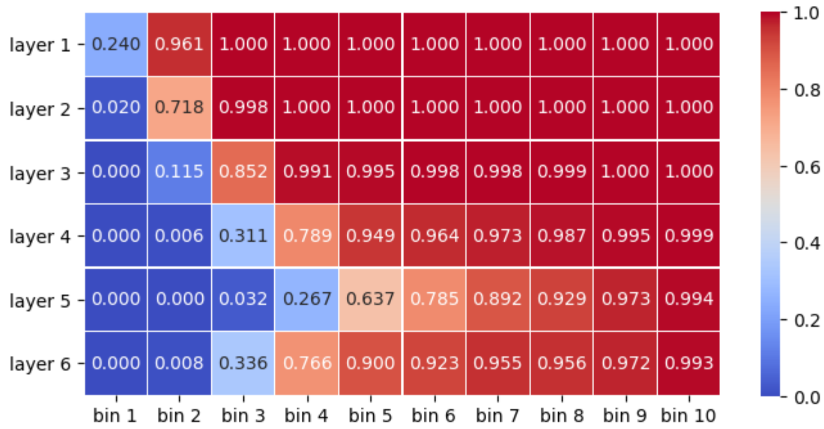

Figure 5.7: Frequency-based classification accuracy on states from the EN→DE encoder + lexical shortcuts.

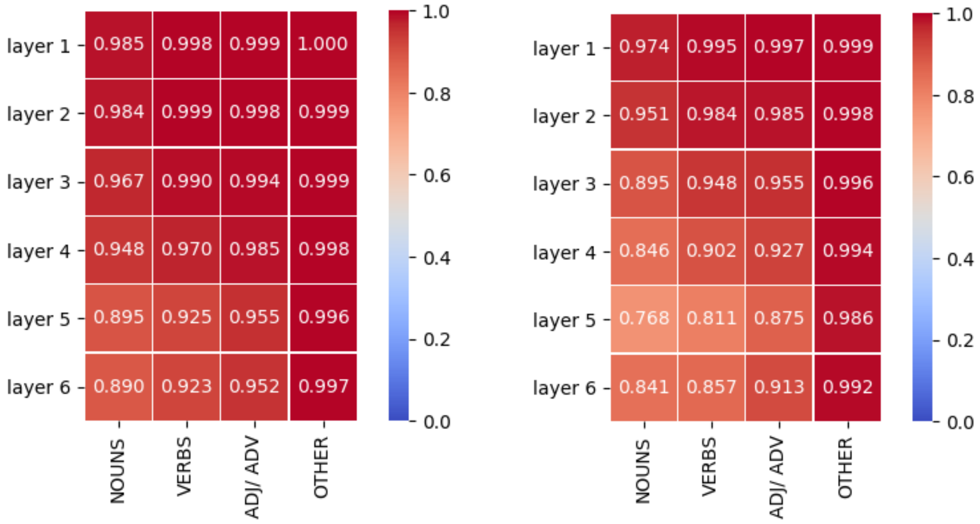

Figure 5.8: **Left**: POS-based classification accuracy on states from the EN→DE encoder; **Right**: POS-based classification accuracy on states from the EN→DE encoder + lexical shortcuts

5.6 Post-Publication Comments

In light of the demonstrated improvements to the translation quality afforded by the addition of lexical shortcut connections to the original transformer architecture, we elected to conduct a more thorough examination of the shortcuts' impact on lexical understanding after the publication of the conference paper presented in this chapter. Specifically, we evaluated the WSD capabilities of the lexical shortcut transformer

(LST) in a setting that is more challenging than the ContraWSD benchmark, by following the methodology outlined in Chapter 3, as well as its capacity for commonsense CoR using the benchmark introduced in Chapter 4. We discuss our findings in this section, offering further evidence for the strengths and limitations of the LST and the utility of the evaluation methods developed in the context of the PhD candidacy.

Model	Test SacreBLEU	% WSD Error	% Attack Success
WMT19 transformer	36.0	17.65	3.96
+ lexical shortcuts	**37.0 (+1.0)**	**15.25 (-2.4)**	**3.17 (-0.79)**
OS18 transformer	29.0	24.17	17.32
+ lexical shortcuts	**29.2 (+0.2)**	**22.46 (-1.71)**	**16.6 (-0.72)**

Table 5.6: Results of the WSD evaluation of the LST model, following the methodology discussed in Chapter 3. Bold numbers represent improvements over the baseline.

The results of the WSD evaluation are summarized in Table 5.6. All evaluated models were implemented using the *fairseq* toolkit. This included a re-implementation of the LST which we found to perform on-par with the original models used in (Emelin et al., 2019) on the WMT14 data[10]. For the WSD study, models are trained and evaluated on the same datasets as in Sections 3.2 and 3.3.1. Hyperparameters were set to be identical to those provided in Table 3.5.

Overall, we observe that the introduction of lexical shortcuts yields notable improvements in test BLEU and greater resilience to adversarial WSD attacks. Likewise, the frequency of WSD errors is substantially reduced in the LST compared to baseline models. Of note are the differences in the extent of the observed performance gains between the two examined domains – *news* (WMT19) and *subtitles* (OS18) – with the former benefiting more from the addition of shortcuts. One possible explanation for this contrast is that individual sentences are, on average, significantly longer in the news domain than in the subtitles domain (19.31 vs. 6.69 tokens per sentence). Consequently, more contextual information must be compressed into latent representations by the WMT transformer in order to encode the source sentence and model the target sequence, leading to increased competition for representational capacity of the hidden states with the propagated embedding features. As the latter can be accessed directly via lexical shortcuts in the LST, the model is able to allocate more space for learn-

[10]https://github.com/demelin/sct_fairseq

ing more informative representations of sentence context, thereby achieving greater translation quality.

Interestingly, we observe minor differences in the behaviour of our baseline models compared to (Emelin et al., 2020). In particular, the WMT19 baseline transformer was reported to achieve 38.2 BLEU on *test2019*, the same test set as used in the current evaluation, i.e. superceding results in 5.6 by 2.2 BLEU. Additionally, the OS18 baseline transformer in (Emelin et al., 2020) is more susceptible to adversarial attacks, with 24.39% of challenge set attacks eliciting flipped translations. The reimplementation of the same model evaluated in Table 5.6 is comparatively more resilient, successfully defending against 6.74% more attacks. While all referenced models were implemented using the same toolkit and evaluated following the same evaluation protocol, the *fairseq* version in (Emelin et al., 2020) precedes the one used in the current set of experiments, suggesting that the observed differences may stem from changes in the underlying codebase. Another potential explanation is that the randomness of model initialization has caused the final model parameters in the different training runs to substantially diverge. The difference in model robustness to adversarial attacks in the OS18 domain, on the other hand, can potentially be explained by the impact of model initialization on the success of individual adversarial attacks, as noted in Section 3.3.4 of Chapter 3.

Model	**Test BLEU EN-DE**	**Test BLEU EN-FR**	**Test BLEU EN-RU**
WMT20 transformer	31.88	39.27	22.28
+ lexical shortcuts	**32.9 (+ 1.02)**	37.93 (-1.34)	22.24 (-0.04)
	Wino-X EN-DE Accuracy	**Wino-X EN-FR Accuracy**	**Wino-X EN-RU Accuracy**
WMT20 transformer	50.32%	49.97%	50.27%
+ lexical shortcuts	50.21% (-0.11%)	49.93% (-0.04%)	48.39% (-1.88%)

Table 5.7: Results of the CoR evaluation of the LST model on the *Wino-X* benchmark, following the methodology discussed in Chapter 4. Bold numbers represent improvements over the baseline.

Table 5.7 reports the findings of our post-publication evaluation of the LST on the *MT-Wino-X* benchmark. All CoR experiments used the same *fairseq* model implementations as the WSD experiments. Training and evaluation followed the same protocol as described in Section 4.3, using the same data and hyperparameters (see Table 4.6).

Perhaps unsurprisingly, lexical shortcuts do not improve model performance on *Wino-X*. This can be potentially explained by *Wino-X* probing for CoR capabilities that explicitly presuppose commonsense reasoning, which is generally observed to emerge in large-scale neural models trained on extraordinary amounts of data, such as LLMs, rather than due to altered neural architecture design (Brown et al., 2020). Interestingly, improvements to translation quality as denoted by test BLEU are inconsistent, differing between the examined translation directions. While a substantial improvement can be observed for EN-DE as in 5.6, BLEU drops noticeably for EN-FR and remains largely unchanged for EN-RU. Together with the results presented in Table 5.3, this suggests that while the addition of lexical shortcuts, generally speaking, positively impacts translation performance, the specifics of their contribution are contingent on a variety of factors, such as the target language or training data size, and are best verified empirically.

Taking this additional evaluation of the proposed LST into account, we can conclude that lexical shortcuts represent a computationally lightweight, easy to implement, and predominantly beneficial addition to the base transformer architecture. While their efficacy varies when evaluated on different aspects of language understanding as well as between target languages and amounts of training data, the overall improvements in translation quality and WSD accuracy provide compelling evidence for their deployment in sufficiently small neural translation models.

Chapter 6

Conclusion

> Alles hat ein Ende, nur die Wurst hat zwei.
>
> German proverb

Taken as a whole, the work presented in this thesis had been conducted with the aim to explore, identify, and quantify some of the current challenges in NMT, while also proposing strategies for alleviating them via changes to model architectures and training regimes. In doing so, the focus of the individual studies remained on lexical language understanding as represented by the WSD and CoR tasks. Unambiguously identifying the intended meaning of every lexical unit within a text requires strong competency in both of these areas. Thus, they should be considered integral to understanding the semantics of natural language and, consequently, to faithfully mapping the meaning of text authored in one language into another language, in other words, to the process of translation. Findings summarized in Chapters 3 and 4 indicate that, despite major advances in deep learning, conventional NMT systems are incapable of consistently reasoning about lexical semantics, instead relying on shallow heuristics that are informed by undesirable biases to perform WSD and CoR alike. At the same time, recently popularized LLMs demonstrate a superior performance on both tasks owing to their immense scale, which suggests a future where LLMs either fully supplant or complement conventional NMT models.

With respect to challenges in WSD, research discussed in Chapter 3 showed that NMT models suffer from *disambiguation bias* that significantly informs their choices when disambiguating polysemous terms. Rather than engaging in deeper reasoning about the intended sense of a homograph by jointly leveraging all available context cues, models instead rely on shallow heuristics such as co-occurrence counts of word senses with specific context words. Once parameterized, model-specific DB can guide the discovery of naturally occurring source language sentences which are likely to elicit WSD errors from evaluated translation models. Furthermore, DB can be effectively exploited to craft synthetic adversarial attacks intended to trigger WSD errors through minimal perturbations of sentences containing ambiguous terms. These observations have been validated across different model architectures and translation domains, suggesting that DB is a pathology that is intrinsic to neural translation models and likely constitutes a side effect of the distributional hypothesis that drives the learning of lexical representations that inform WSD choices.

Similarly, work presented in Chapter 4 found that conventional NMT systems are unable to successfully identify referents of ambiguous pronouns as part of the translation process in cases where the CoR step requires commonsense reasoning. Rather than taking into consideration properties of entities denoted by the linguistic expressions and their relations to each other, NMT models are guided in their CoR decisions by superficial biases that arise from artifacts present in their training distribution. Experimental results show that among the identified bias sources, grammatical gender of potential referents informs model behavior to a greater extent than their position within the translated sentence. At the same time, strategies aiming to reduce such biases are shown to improve CoR accuracy. Given the exploratory nature of this research, it is likely that future work may be able to uncover additional CoR biases in NMT systems, complementing these initial findings. Compared to NMT models, MLLMs pre-trained on data from a diverse set of languages stand out as the superior coreference resolvers, exhibiting a highly promising capacity for cross-lingual commonsense knowledge transfer.

As a potential avenue for improving lexical reasoning in NMT models, research documented in Chapter 5 explored a simple yet effective extension to the popular transformer architecture. Gated shortcut connections introduced between the embedding layer and each subsequent layer within the encoder and the decoder allow lexical features to be accessed dynamically for the computation of individual layer activations, instead of having to sequentially traverse the layer stack to be accessible across the model. Freeing up the representational capacity of transformer layers in this manner enables the translation model to better capture the contextual information within the source and target sentences, resulting in improved translation quality. Experimental findings furthermore indicate greater lexical understanding in models equipped with lexical shortcuts, based on their superior WSD capabilities. Post-publication experiments further support these findings by demonstrating that the incorporation of lexical shortcuts into the transformer architecture reduces WSD errors attributable to DB and improves models' resilience to adversarial WSD attacks.

Following the completion of studies comprising the main chapters of this thesis, additional experiments were conducted so as to evaluate whether the proposed challenge sets and evaluation methodologies remain effective if applied to LLMs, which have recently came to prominence as highly capable language compreheders and generators. Overall, LLMs are found to offer superior lexical reasoning capabilities compared to conventional NMT models, especially when used in combination with advanced prompting techniques, such as the proposed chain-of-tasks prompting. In particular, explicit deliberation of the intended homograph sense in the source sentence can improve the robustness of LLMs to adversarial WSD attacks, albeit not consistently. Similarly, a positive trend can be observed when tasking LLMs with translations that require commonsense reasoning to successfully resolve coreference. While basic prompting yields no substantial improvements over much smaller, conventional NMT systems, CoTA prompting can be effectively leveraged to raise CoR accuracy. Of note is the impressive performance of LLMs on the *LM-Wino-X* challenge sets intended to evaluate monolingual CoR, which appears to be far less challenging than CoR within the translation context, as represented by the *MT-Wino-X* challenge sets, despite containing many of the same samples albeit in a different format. This suggests that while CoTA prompting is an effective tool for eliciting semantically consistent translations from LLMs, it fails to fully leverage the lexical understanding capabilities of LLMs. Ultimately, despite improving upon NMT models, LLMs remain prone to WSD and CoR errors with much room left for improvement on both tasks, which substantiates the utility of methods and resources introduced in this thesis.

As with any scientific endeavor, the work presented here is not without its limitations. One of them is that uncovering the DB of a model is a labor-intensive, inherently noisy process that relies on several tools and resources that are imperfect and may introduce errors. Similarly, constructing adversarial WSD attacks that are natural-sounding and coherent remains a challenge, with the proposed methodology occasionally producing sub-par samples that are of limited utility for model evaluation. Commonsense reasoning as part of CoR, on the other hand, has been shown to benefit from model scaling and may pose less of a challenge for sufficiently large NMT models, which our study was unable to verify due to unavailability of such resources at the time. Efficient integration of LLMs into NMT systems is likewise an approach that could potentially alleviate some of the challenges uncovered in this thesis. Finally, while lexical shortcuts demonstrably improve translation quality and lexical reasoning in smaller transformer models, current NLP research is becoming increasingly dominated by massive models with billions of parameters. For models of this scope, limited representational capacity is unlikely to be an issue, which limits the utility of the proposed LST architecture. However, in settings where space and compute are constrained, such as embedded devices, lexical shortcuts may find a fruitful practical application.

A potential guideline for future research motivated by the central findings presented in this thesis, therefore, emphasizes the creation, extension, and maintenance of challenging evaluation benchmarks that target specific linguistic capabilities of translation models. Ensuring that translation systems are capable of successfully completing the individual sub-steps involved in the translation process – including, but not limited to, lexical comprehension – is essential for attaining human-like translation quality through iterative changes to model design and learning objectives.

Ultimately, WSD and CoR represent unsolved challenges in machine translation and will remain such until models are capable of language understanding that is not defacto shortcut learning in disguise. In order to ensure that our evaluations of model capabilities remain objective and accurate, more challenging and comprehensive benchmarks are needed. Work introduced in Chapters 3 and 4 represents such evaluation efforts and offers valuable insights into potential methodologies for the construction of future cross-lingual benchmarks, adaptation of existing monolingual benchmarks for the multi-lingual use case, and strategies for the targeted evaluation of the word-level language processing capabilities of NMT models and LLMs. While the methods discussed in Chapter 3 are challenging to apply to models trained on vast quantities of data or on proprietary data sources inaccessible to the public, they can nevertheless be effectively leveraged for small- and medium-scale fine-tuning data frequently employed to adapt a foundation model to a target domain or application. As such – jointly with the model-agnostic benchmarks presented in Chapter 4 – they remain a valuable analysis tool within the ever-evolving AI landscape. Conclusions drawn from evaluations based on these and similar resources can in turn motivate model development and refinement.

Undoubtedly, the field of machine translation will continue to advance, incorporating and spurring future innovations in the field of NLP. The newly available LLMs, in particular, represent a powerful tool to be considered for inclusion into the translation pipeline. As such, it is paramount to keep identifying the weaknesses of prominent models and finding ways to address them, in order to ensure that this technology functions as a reliable bridge between human communities, enabling a free and unburdened exchange of ideas by bridging linguistic divides.

Bibliography

Alzantot, M., Sharma, Y. S., Elgohary, A., Ho, B.-J., Srivastava, M., and Chang, K.-W. (2018). Generating natural language adversarial examples. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.

Amrhein, C., Moghe, N., and Guillou, L. (2022). Aces: Translation accuracy challenge sets for evaluating machine translation metrics. *arXiv preprint arXiv:2210.15615*.

Amsili, P. and Seminck, O. (2017). A google-proof collection of french winograd schemas. In *Proceedings of the 2nd Workshop on Coreference Resolution Beyond OntoNotes (CORBON 2017)*, pages 24–29.

Anil, R., Dai, A. M., Firat, O., Johnson, M., Lepikhin, D., Passos, A., Shakeri, S., Taropa, E., Bailey, P., Chen, Z., et al. (2023). Palm 2 technical report. *arXiv preprint arXiv:2305.10403*.

Bapna, A., Chen, M. X., Firat, O., Cao, Y., and Wu, Y. (2018). Training deeper neural machine translation models with transparent attention. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3028–3033.

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

Belinkov, Y., Durrani, N., Dalvi, F., Sajjad, H., and Glass, J. (2017a). What do neural machine translation models learn about morphology? In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 861–872.

Belinkov, Y., Màrquez, L., Sajjad, H., Durrani, N., Dalvi, F., and Glass, J. (2017b). Evaluating layers of representation in neural machine translation on part-of-speech and semantic tagging tasks. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1–10.

Belinkov, Y., Poliak, A., Shieber, S., Van Durme, B., and Rush, A. S. (2019). On adversarial removal of hypothesis-only bias in natural language inference. In *Proceedings of the Joint Conference on Lexical and Computational Semantics*.

Bender, E. M. and Friedman, B. (2018). Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Bernard, T. and Han, T. (2020). Mandarinograd: A chinese collection of winograd schemas. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 21–26.

Bhagavatula, C., Bras, R. L., Malaviya, C., Sakaguchi, K., Holtzman, A., Rashkin, H., Downey, D., Yih, S. W.-t., and Choi, Y. (2019). Abductive commonsense reasoning. *arXiv preprint arXiv:1908.05739*.

Biderman, S., Schoelkopf, H., Anthony, Q. G., Bradley, H., O'Brien, K., Hallahan, E., Khan, M. A., Purohit, S., Prashanth, U. S., Raff, E., et al. (2023). Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, pages 2397–2430. PMLR.

Bisazza, A. and Tump, C. (2018). The lazy encoder: A fine-grained analysis of the role of morphology in neural machine translation. In *Conference on Empirical Methods in Natural Language Processing*.

Bojar, O., Federmann, C., Fishel, M., Graham, Y., Haddow, B., Huck, M., Koehn, P., and Monz, C. (2018). Findings of the 2018 conference on machine translation (wmt18). In *Proceedings of the Third Conference on Machine Translation*, pages 272–307, Belgium, Brussels. Association for Computational Linguistics.

Bovi, C. D., Telesca, L., and Navigli, R. (2015). Large-scale information extraction from textual definitions through deep syntactic and semantic analysis. *Transactions of the Association for Computational Linguistics*, 3:529–543.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901.

Campolungo, N., Martelli, F., Saina, F., and Navigli, R. (2022a). Dibimt: A novel benchmark for measuring word sense disambiguation biases in machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4331–4352.

Campolungo, N., Pasini, T., Emelin, D., and Navigli, R. (2022b). Reducing disambiguation biases in nmt by leveraging explicit word sense information. In *North American Chapter of the Association for Computational Linguistics*.

Chen, M. X., Firat, O., Bapna, A., Johnson, M., Macherey, W., Foster, G., Jones, L., Schuster, M., Shazeer, N., Parmar, N., et al. (2018). The best of both worlds: Combining recent advances in neural machine translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 76–86.

Cheng, M., Yi, J., Chen, P.-Y., Zhang, H., and Hsieh, C.-J. (2020). Seq2sick: Evaluating the robustness of sequence-to-sequence models with adversarial examples. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 3601–3608.

Cheng, Y., Jiang, L., and Macherey, W. (2019). Robust neural machine translation with doubly adversarial inputs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4324–4333.

Cho, K., van Merrienboer, B., Gülçehre, , Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014). Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734.

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S., et al. (2022). Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Cohen, J. (2013). *Statistical Power Analysis for the Behavioral Sciences*. Academic press.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, É., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.

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

Cureton, E. E. (1956). Rank-biserial correlation. *Psychometrika*, 21(3):287–290.

Dankers, V., Lucas, C., and Titov, I. (2022). Can transformer be too compositional? analysing idiom processing in neural machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3608–3626.

Dauphin, Y. N., Fan, A., Auli, M., and Grangier, D. (2017). Language modeling with gated convolutional networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 933–941. JMLR. org.

Dehghani, M., Gouws, S., Vinyals, O., Uszkoreit, J., and Kaiser, Ł. (2018). Universal transformers. *arXiv preprint arXiv:1807.03819*.

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

Dhar, P. and Bisazza, A. (2020). The effect of vocabulary overlap on linguistic probing tasks for neural language models.

Dinu, G., Mathur, P., Federico, M., and Al-Onaizan, Y. (2019). Training neural machine translation to apply terminology constraints. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3063–3068.

Dou, Z.-Y., Tu, Z., Wang, X., Shi, S., and Zhang, T. (2018). Exploiting deep representations for neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4253–4262.

Dyer, C., Chahuneau, V., and Smith, N. A. (2013). A simple, fast, and effective reparameterization of ibm model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648.

Emelin, D., Le Bras, R., Hwang, J. D., Forbes, M., and Choi, Y. (2021). Moral stories: Situated reasoning about norms, intents, actions, and their consequences. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 698–718.

Emelin, D. and Sennrich, R. (2021). Wino-x: Multilingual winograd schemas for commonsense reasoning and coreference resolution. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8517–8532.

Emelin, D., Titov, I., and Sennrich, R. (2019). Widening the representation bottleneck in neural machine translation with lexical shortcuts. In *ACL 2019 Fourth Conference on Machine Translation*, pages 102–115. Association for Computational Linguistics (ACL).

Emelin, D., Titov, I., and Sennrich, R. (2020). Detecting word sense disambiguation biases in machine translation for model-agnostic adversarial attacks. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7635–7653.

Futeral, M., Schmid, C., Laptev, I., Sagot, B., and Bawden, R. (2022). Tackling ambiguity with images: Improved multimodal machine translation and contrastive evaluation. *arXiv preprint arXiv:2212.10140*.

Gardner, M., Artzi, Y., Basmov, V., Berant, J., Bogin, B., Chen, S., Dasigi, P., Dua, D., Elazar, Y., Gottumukkala, A., et al. (2020). Evaluating models' local decision boundaries via contrast sets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 1307–1323.

Gehring, J., Auli, M., Grangier, D., and Dauphin, Y. (2017). A convolutional encoder model for neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 123–135.

Glockner, M., Shwartz, V., and Goldberg, Y. (2018). Breaking nli systems with sentences that require simple lexical inferences. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 650–655.

Graham, Y., Haddow, B., and Koehn, P. (2020). Translationese in machine translation evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 72–81.

Greff, K., Srivastava, R. K., and Schmidhuber, J. (2016). Highway and residual networks learn unrolled iterative estimation. *arXiv preprint arXiv:1612.07771*.

Guerreiro, N. M., Alves, D., Waldendorf, J., Haddow, B., Birch, A., Colombo, P., and Martins, A. (2023). Hallucinations in large multilingual translation models. *ArXiv*, abs/2303.16104.

Guillou, L. and Hardmeier, C. (2016). Protest: A test suite for evaluating pronouns in machine translation. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 636–643.

Gururangan, S., Swayamdipta, S., Levy, O., Schwartz, R., Bowman, S., and Smith, N. A. (2018). Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112.

Haddow, B., Bogoychev, N., Emelin, D., Germann, U., Grundkiewicz, R., Heafield, K., Miceli Barone, A. V., and Sennrich, R. (2018). The university of edinburgh's submissions to the wmt18 news translation task. In *Proceedings of the Third Conference on Machine Translation*, pages 403–413, Belgium, Brussels. Association for Computational Linguistics.

Hassan, H., Aue, A., Chen, C., Chowdhary, V., Clark, J., Federmann, C., Huang, X., Junczys-Dowmunt, M., Lewis, W., Li, M., et al. (2018). Achieving human parity on automatic chinese to english news translation. *arXiv preprint arXiv:1803.05567*.

He, J., Wang, T., Xiong, D., and Liu, Q. (2020). The box is in the pen: Evaluating commonsense reasoning in neural machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 3662–3672.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778.

Hendy, A., Abdelrehim, M., Sharaf, A., Raunak, V., Gabr, M., Matsushita, H., Kim, Y. J., Afify, M., and Awadalla, H. H. (2023). How good are gpt models at machine translation? a comprehensive evaluation. *arXiv preprint arXiv:2302.09210*.

Herold, C. and Ney, H. (2023). Improving long context document-level machine translation. *arXiv preprint arXiv:2306.05183*.

Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., Casas, D. d. L., Hendricks, L. A., Welbl, J., Clark, A., et al. (2022). Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*.

Honnibal, M. and Montani, I. (2017). spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. *To appear*, 7(1).

Hu, J., Ruder, S., Siddhant, A., Neubig, G., Firat, O., and Johnson, M. (2020). Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *International Conference on Machine Learning*, pages 4411–4421. PMLR.

Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. (2017). Densely connected convolutional networks. In *CVPR*, number 2, page 3.

Huang, L., Le Bras, R., Bhagavatula, C., and Choi, Y. (2019). Cosmos qa: Machine reading comprehension with contextual commonsense reasoning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2391–2401.

Isaak, N. and Michael, L. (2019). Winoflexi: A crowdsourcing platform for the development of winograd schemas. In *Australasian Joint Conference on Artificial Intelligence*, pages 289–302. Springer.

Iyer, V., Chen, P., and Birch, A. (2023). Towards effective disambiguation for machine translation with large language models. In *Proceedings of the Eighth Conference on Machine Translation*, pages 482–495.

Jean, S., Lauly, S., Firat, O., and Cho, K. (2017). Neural machine translation for cross-lingual pronoun prediction. In *Proceedings of the third workshop on discourse in machine translation*, pages 54–57.

Jin, D., Jin, Z., Zhou, J. T., and Szolovits, P. (2020). Is bert really robust? a strong baseline for natural language attack on text classification and entailment. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 8018–8025.

Karpathy, A., Johnson, J., and Fei-Fei, L. (2015). Visualizing and understanding recurrent networks. *arXiv preprint arXiv:1506.02078*.

Kim, Y. and Rush, A. M. (2016). Sequence-level knowledge distillation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1317–1327.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kocijan, V., Lukasiewicz, T., Davis, E., Marcus, G., and Morgenstern, L. (2020). A review of winograd schema challenge datasets and approaches. *arXiv preprint arXiv:2004.13831*.

Kocmi, T., Bawden, R., Bojar, O., Dvorkovich, A., Federmann, C., Fishel, M., Gowda, T., Graham, Y., Grundkiewicz, R., Haddow, B., et al. (2022). Findings of the 2022 conference on machine translation (wmt22). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 1–45.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., et al. (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180.

Kuang, S., Li, J., Branco, A., Luo, W., and Xiong, D. (2018). Attention focusing for neural machine translation by bridging source and target embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1767–1776.

Kuleshov, V., Thakoor, S., Lau, T., and Ermon, S. (2018). Adversarial examples for natural language classification problems.

Le Bras, R., Swayamdipta, S., Bhagavatula, C., Zellers, R., Peters, M., Sabharwal, A., and Choi, Y. (2020). Adversarial filters of dataset biases. In *International Conference on Machine Learning*, pages 1078–1088. PMLR.

Lee, C.-Y., Xie, S., Gallagher, P., Zhang, Z., and Tu, Z. (2015). Deeply-supervised nets. In *Artificial Intelligence and Statistics*, pages 562–570.

Lefever, E. and Hoste, V. (2013). Semeval-2013 task 10: Cross-lingual word sense disambiguation. In *Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 158–166.

Levesque, H., Davis, E., and Morgenstern, L. (2012). The winograd schema challenge. In *Thirteenth International Conference on the Principles of Knowledge Representation and Reasoning*. Citeseer.

Li, J., Liu, L., Li, H., Li, G., Huang, G., and Shi, S. (2020). Evaluating explanation methods for neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 365–375.

Li, X., Li, G., Liu, L., Meng, M., and Shi, S. (2019). On the word alignment from neural machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1293–1303.

Li, X. L., Kuncoro, A., Hoffmann, J., de Masson d'Autume, C., Blunsom, P., and Nematzadeh, A. (2022). A systematic investigation of commonsense knowledge in large language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11838–11855.

Li, Y. and Vasconcelos, N. (2019). Repair: Removing representation bias by dataset resampling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9572–9581.

Liang, Y., Duan, N., Gong, Y., Wu, N., Guo, F., Qi, W., Gong, M., Shou, L., Jiang, D., Cao, G., et al. (2020). Xglue: A new benchmark datasetfor cross-lingual pre-training, understanding and generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6008–6018.

Lin, B. Y., Lee, S., Qiao, X., and Ren, X. (2021). Common sense beyond english: Evaluating and improving multilingual language models for common sense reasoning. In *To appear.*

Lison, P., Tiedemann, J., Kouylekov, M., et al. (2019). Open subtitles 2018: Statistical rescoring of sentence alignments in large, noisy parallel corpora. In *LREC 2018, Eleventh International Conference on Language Resources and Evaluation*. European Language Resources Association (ELRA).

Liu, F., Lu, H., and Neubig, G. (2018). Handling homographs in neural machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1336–1345.

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., and Neubig, G. (2023). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35.

Liu, Y., Gu, J., Goyal, N., Li, X., Edunov, S., Ghazvininejad, M., Lewis, M., and Zettlemoyer, L. (2020). Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Luong, M.-T., Pham, H., and Manning, C. D. (2015). Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421.

Mann, H. B. and Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, pages 50–60.

Marvin, R. and Koehn, P. (2018). Exploring word sense disambiguation abilities of neural machine translation systems (non-archival extended abstract). In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Papers)*, pages 125–131.

McCoy, T., Pavlick, E., and Linzen, T. (2019). Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448.

McGrath, R. E. and Meyer, G. J. (2006). When effect sizes disagree: The case of r and d. *Psychological methods*, 11(4):386.

Melo, G., Imaizumi, V., and Cozman, F. (2019). Winograd schemas in portuguese. In *Anais do XVI Encontro Nacional de Inteligência Artificial e Computacional*, pages 787–798. SBC.

Michel, P., Li, X., Neubig, G., and Pino, J. M. (2019). On evaluation of adversarial perturbations for sequence-to-sequence models. In *Proceedings of NAACL-HLT*, pages 3103–3114.

Morris, J., Lifland, E., Lanchantin, J., Ji, Y., and Qi, Y. (2020). Reevaluating adversarial examples in natural language. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3829–3839.

Moryossef, A., Aharoni, R., and Goldberg, Y. (2019). Filling gender & number gaps in neural machine translation with black-box context injection. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 49–54.

Müller, M., Gonzales, A. R., Voita, E., and Sennrich, R. (2018). A large-scale test set for the evaluation of context-aware pronoun translation in neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 61–72.

Müller, M., Rios, A., and Sennrich, R. (2020). Domain robustness in neural machine translation. In *Proceedings of the 14th Conference of the Association for Machine Translation in the Americas*, pages 151–164. Association for Machine Translation in the Americas.

Murray, K. and Chiang, D. (2018). Correcting length bias in neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 212–223.

Navigli, R. and Ponzetto, S. P. (2010). Babelnet: Building a very large multilingual semantic network. In *Proceedings of the 48th Annual Meeting of the Association* *for Computational Linguistics*, pages 216–225. Association for Computational Linguistics.

Newman, B., Hewitt, J., Liang, P., and Manning, C. D. (2020). The EOS decision and length extrapolation. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 276–291, Online. Association for Computational Linguistics.

Ng, N., Yee, K., Baevski, A., Ott, M., Auli, M., and Edunov, S. (2019). Facebook fair's wmt19 news translation task submission. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 314–319.

Nguyen, T. Q. and Chiang, D. (2018). Improving lexical choice in neural machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 334–343.

OpenAI (2022). Chatgpt: Optimizing language models for dialogue.

OpenAI (2023). Gpt-4 technical report. *ArXiv*, abs/2303.08774.

Ott, M., Edunov, S., Baevski, A., Fan, A., Gross, S., Ng, N., Grangier, D., and Auli, M. (2019). fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*.

Ott, M., Edunov, S., Grangier, D., and Auli, M. (2018). Scaling neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 1–9.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). 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.

Post, M. (2018). A call for clarity in reporting bleu scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191.

Press, O. and Wolf, L. (2017). Using the output embedding to improve language models. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 157–163.

Pu, X., Pappas, N., Henderson, J., and Popescu-Belis, A. (2018). Integrating weakly supervised word sense disambiguation into neural machine translation. *Transactions of the Association for Computational Linguistics*, 6:635–649.

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., and Manning, C. D. (2020). Stanza: A python natural language processing toolkit for many human languages. *arXiv preprint arXiv:2003.07082*.

Qian, P., Qiu, X., and Huang, X. (2016). Analyzing linguistic knowledge in sequential model of sentence. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 826–835.

Radford, A., Narasimhan, K., Salimans, T., and Sutskever, I. (2018). Improving language understanding by generative pre-training.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). Language models are unsupervised multitask learners.

Raganato, A., Scherrer, Y., and Tiedemann, J. (2019). The mucow test suite at wmt 2019: Automatically harvested multilingual contrastive word sense disambiguation test sets for machine translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 470–480.

Raunak, V., Menezes, A., and Junczys-Dowmunt, M. (2021). The curious case of hallucinations in neural machine translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1172–1183.

Renduchintala, A., Díaz, D., Heafield, K., Li, X., and Diab, M. (2021). Gender bias amplification during speed-quality optimization in neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 99–109.

Rios, A., Mascarell, L., and Sennrich, R. (2017). Improving word sense disambiguation in neural machine translation with sense embeddings. In *Proceedings of the 2nd Conference on Machine Translation, Copenhagen, Denmark*.

Roth, T., Gao, Y., Abuadbba, A., Nepal, S., and Liu, W. (2021). Token-modification adversarial attacks for natural language processing: A survey. *arXiv preprint arXiv:2103.00676*.

Ruscio, J. (2008). A probability-based measure of effect size: Robustness to base rates and other factors. *Psychological Methods*, 13(1):19.

Sakaguchi, K., Le Bras, R., Bhagavatula, C., and Choi, Y. (2020). Winogrande: An adversarial winograd schema challenge at scale. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8732–8740.

Salazar, J., Liang, D., Nguyen, T. Q., and Kirchhoff, K. (2020). Masked language model scoring. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2699–2712.

Samanta, S. and Mehta, S. (2017). Towards crafting text adversarial samples. *arXiv preprint arXiv:1707.02812*.

Sarti, G., Feldhus, N., Sickert, L., van der Wal, O., Nissim, M., and Bisazza, A. (2023). Inseq: An interpretability toolkit for sequence generation models. *ArXiv*, abs/2302.13942.

Saunders, D. and Byrne, B. (2020). Reducing gender bias in neural machine translation as a domain adaptation problem. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7724–7736.

Saunders, D., Stahlberg, F., de Gispert, A., and Byrne, B. (2018). Multi-representation ensembles and delayed sgd updates improve syntax-based nmt. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 319–325.

Scao, T. L., Fan, A., Akiki, C., Pavlick, E., Ilić, S., Hesslow, D., Castagné, R., Luccioni, A. S., Yvon, F., Gallé, M., et al. (2022). Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Schmid, H. (1999). Improvements in part-of-speech tagging with an application to german. In *Natural Language Processing Using Very Large Corpora*, pages 13–25. Springer.

Sennrich, R., Haddow, B., and Birch, A. (2016). Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725.

Sennrich, R., Vamvas, J., and Mohammadshahi, A. (2023). Mitigating hallucinations and off-target machine translation with source-contrastive and language-contrastive decoding. *arXiv preprint arXiv:2309.07098*.

Shayegani, E., Mamun, M. A. A., Fu, Y., Zaree, P., Dong, Y., and Abu-Ghazaleh, N. (2023). Survey of vulnerabilities in large language models revealed by adversarial attacks. *arXiv preprint arXiv:2310.10844*.

Shi, X., Padhi, I., and Knight, K. (2016). Does string-based neural mt learn source syntax? In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1526–1534.

Shimura, K., Li, J., and Fukumoto, F. (2019). Text categorization by learning predominant sense of words as auxiliary task. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1109–1119.

Skianis, K., Briand, Y., and Desgrippes, F. (2020). Evaluation of machine translation methods applied to medical terminologies. In *Proceedings of the 11th International Workshop on Health Text Mining and Information Analysis*, pages 59–69.

Speer, R., Chin, J., and Havasi, C. (2017). Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31.

Srivastava, A., Rastogi, A., Rao, A., Shoeb, A. A. M., Abid, A., Fisch, A., Brown, A. R., Santoro, A., Gupta, A., Garriga-Alonso, A., et al. (2022). Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958.

Srivastava, R. K., Greff, K., and Schmidhuber, J. (2015). Highway networks. *arXiv preprint arXiv:1505.00387*.

Stanovsky, G., Smith, N. A., and Zettlemoyer, L. (2019). Evaluating gender bias in machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1679–1684.

Stojanovski, D. and Fraser, A. (2019). Improving anaphora resolution in neural machine translation using curriculum learning. In *Proceedings of Machine Translation Summit XVII: Research Track*, pages 140–150.

Stojanovski, D., Krojer, B., Peskov, D., and Fraser, A. (2020). Contracat: Contrastive coreference analytical templates for machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4732–4749.

Tang, G., Müller, M., Rios, A., and Sennrich, R. (2018). Why self-attention?: A targeted evaluation of neural machine translation architectures. In *Conference on Empirical Methods in Natural Language Processing, October 31–November 4 Brussels, Belgium, 2018*, pages 4263–4272.

Tang, G., Sennrich, R., and Nivre, J. (2019). Encoders help you disambiguate word senses in neural machine translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1429–1435.

Tang, Y., Tran, C., Li, X., Chen, P.-J., Goyal, N., Chaudhary, V., Gu, J., and Fan, A. (2020). Multilingual translation with extensible multilingual pretraining and finetuning. *arXiv preprint arXiv:2008.00401*.

Tiedemann, J. and Scherrer, Y. (2017). Neural machine translation with extended context. In *Proceedings of the Third Workshop on Discourse in Machine Translation*, pages 82–92.

Toral, A. and Way, A. (2018). *What Level of Quality can Neural Machine Translation Attain on Literary Text?* Springer.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. (2023). Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Vallat, R. (2018). Pingouin: Statistics in python. *The Journal of Open Source Software*, 3(31):1026.

Vamvas, J. and Sennrich, R. (2021). Contrastive conditioning for assessing disambiguation in mt: A case study of distilled bias. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10246–10265.

Vanmassenhove, E., Hardmeier, C., and Way, A. (2018). Getting gender right in neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3003–3008.

Vanmassenhove, E., Shterionov, D., and Gwilliam, M. (2021). Machine translationese: Effects of algorithmic bias on linguistic complexity in machine translation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2203–2213.

Vaswani, A., Bengio, S., Brevdo, E., Chollet, F., Gomez, A., Gouws, S., Jones, L., Kaiser, Ł., Kalchbrenner, N., Parmar, N., et al. (2018). Tensor2tensor for neural machine translation. In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 193–199.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

Vickrey, D., Biewald, L., Teyssier, M., and Koller, D. (2005). Word-sense disambiguation for machine translation. In *Proceedings of human language technology conference and conference on empirical methods in natural language processing*, pages 771–778.

Voita, E., Sennrich, R., and Titov, I. (2019a). Context-aware monolingual repair for neural machine translation. In *2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing*, pages 876–885. Association for Computational Linguistics (ACL).

Voita, E., Sennrich, R., and Titov, I. (2019b). When a good translation is wrong in context: Context-aware machine translation improves on deixis, ellipsis, and lexical cohesion. In *57th Annual Meeting of the Association for Computational Linguistics*, pages 1198–1212. ACL Anthology.

Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. (2019). Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in Neural Information Processing Systems*, 32.

Wang, C. and Sennrich, R. (2020). On exposure bias, hallucination and domain shift in neural machine translation. In *2020 Annual Conference of the Association for Computational Linguistics*, pages 3544–3552. Association for Computational Linguistics (ACL).

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., et al. (2022a). Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., Zhou, D., et al. (2022b). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Werlen, L. M., Ram, D., Pappas, N., and Henderson, J. (2018). Document-level neural machine translation with hierarchical attention networks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2947–2954.

White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., Elnashar, A., Spencer-Smith, J., and Schmidt, D. C. (2023). A prompt pattern catalog to enhance prompt engineering with chatgpt. *arXiv preprint arXiv:2302.11382*.

Winograd, T. (1972). Understanding natural language. *Cognitive Psychology*, 3(1):1–191.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., et al. (2020). 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.

Wu, L., Tian, F., Zhao, L., Lai, J., and Liu, T.-Y. (2018). Word attention for sequence to sequence text understanding. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Łukasz Kaiser, Gouws, S., Kato, Y., Kudo, T., Kazawa, H., Stevens, K., Kurian, G., Patil, N., Wang, W., Young, C., Smith, J., Riesa, J., Rudnick, A., Vinyals, O., Corrado, G., Hughes, M., and Dean, J. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, abs/1609.08144.

Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., and Raffel, C. (2021). mt5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498.

Yehudai, A., Cattan, A., Abend, O., and Stanovsky, G. (2023). Evaluating and improving the coreference capabilities of machine translation models. *arXiv preprint arXiv:2302.08464*.

Yin, D., Bansal, H., Monajatipoor, M., Li, L. H., and Chang, K.-W. (2022). Geomlama: Geo-diverse commonsense probing on multilingual pre-trained language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2039–2055.

Zacharopoulos, C., Desbordes, T., and Sablé-Meyer, M. (2023). Assessing the influence of attractor-verb distance on grammatical agreement in humans and language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16081–16090.

Zeng, Z. and Xiong, D. (2021). An empirical study on adversarial attack on nmt: Languages and positions matter. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 454–460.

Zhang, B., Haddow, B., and Birch, A. (2023). Prompting large language model for machine translation: A case study. *arXiv preprint arXiv:2301.07069*.

Zhang, H., Zhou, H., Miao, N., and Li, L. (2019). Generating fluent adversarial examples for natural languages. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5564–5569.

Zhang, T., Ye, W., Yang, B., Zhang, L., Ren, X., Liu, D., Sun, J., Zhang, S., Zhang, H., and Zhao, W. (2022). Frequency-aware contrastive learning for neural machine translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11712–11720.

Zhang, W. E., Sheng, Q. Z., Alhazmi, A., and Li, C. (2020). Adversarial attacks on deep-learning models in natural language processing: A survey. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 11(3):1–41.

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., et al. (2023). A survey of large language models. *arXiv preprint arXiv:2303.18223*.