Karlsruhe Institute of Technology

Detecting Ambiguity in Neural Machine Translation Models by Inspecting Diversity in Translation

Master's Thesis of

Violina Zhekova

at the Department of Informatics
Artificial Intelligence for Language Technologies (AI4LT)

Reviewer:	Prof. Dr. Jan Niehues
Second reviewer:	Prof. Dr. Alexander Waibel
Advisor:	M. Sc. Tu Anh Dinh

01. May 2023 – 02. November 2023

I declare that I have developed and written the enclosed thesis completely by myself. I have submitted neither parts of nor the complete thesis as an examination elsewhere. I have not used any other than the aids that I have mentioned. I have marked all parts of the thesis that I have included from referenced literature, either in their original wording or paraphrasing their contents. This also applies to figures, sketches, images and similar depictions, as well as sources from the internet.

PLACE, DATE

Karlsruhe, 31.10.2023

(Violina Zhekova)

Violina Zhekova

Abstract

In the era of globalization, the need for effective communication across different languages is paramount. Neural Machine Translation (NMT), a subfield of Machine Learning, has made significant strides in this direction by using deep learning to translate directly from an input source language to an output target language. However, these models often perpetuate social constructs and stereotypes present in the training data, leading to biased translations.

This thesis presents a novel method for detecting ambiguous words in text that could lead to bias by inspecting the diversity of translation. We hypothesize that sentences containing ambiguous words, which have one version in the source language and multiple versions in the target language, produce less diverse backtranslations than sentences without ambiguity. To test this hypothesis, we compare a dataset containing ambiguous sentences with such where the ambiguous word is replaced with a non-ambiguous equivalent. Our approach reveals patterns in translation diversity that could indicate ambiguity, providing a new framework for detecting ambiguity in text devoid of contextual information regarding the ambiguity. The results show that replacing an ambiguous word with its disambiguated equivalent leads to more diverse translations in some scenarios, partially proving the hypothesis. Furthermore, we tested our approach in a real-world experiment using natural spoken sentences containing ambiguity. The experiment revealed the potential of the approach to detect ambiguous words, while also uncovering specific limitations of the method.

This research has significant implications for the field of NMT, offering a potential solution to the problem of unjustified assumptions about ambiguous words leading to bias in machine translations. By uncovering ambiguity, we can work towards more accurate and fair translations, fostering better cross-cultural communication and understanding.

Zusammenfassung

In der Ära der Globalisierung ist die Notwendigkeit einer effektiven Kommunikation in verschiedenen Sprachen von größter Bedeutung. Die neuronale maschinelle Übersetzung (NMT), ein Teilbereich des maschinellen Lernens, hat in diese Richtung bedeutende Fortschritte gemacht, indem sie Deep Learning verwendet, um direkt von einer Eingabesprache in eine Ausgabesprache zu übersetzen. Diese Modelle perpetuieren jedoch oft soziale Konstrukte und Stereotypen, die in den Trainingsdaten vorhanden sind, was zu voreingenommenen Übersetzungen führt.

Diese Arbeit stellt eine neuartige Methode zur Erkennung mehrdeutiger Wörter in Texten vor, die zu Vorurteilen führen könnten, indem die Vielfalt der Übersetzung untersucht wird. Wir stellen die Hypothese auf, dass Sätze, die mehrdeutige Wörter enthalten, die in der Quellsprache eine Version und in der Zielsprache mehrere Versionen haben, weniger vielfältige Rückübersetzungen erzeugen als Sätze ohne Mehrdeutigkeit. Um diese Hypothese zu testen, vergleichen wir einen Datensatz mit mehrdeutigen Sätzen mit einem solchen, bei dem das mehrdeutige Wort durch ein nicht mehrdeutiges Äquivalent ersetzt wird. Unser Ansatz offenbart Muster in der Übersetzungsdiversität, die auf Mehrdeutigkeit hindeuten könnten und bietet einen neuen Rahmen zur Erkennung von Mehrdeutigkeit in Texten, die keine kontextuellen Informationen zur Mehrdeutigkeit enthalten. Die Ergebnisse zeigen, dass das Ersetzen eines mehrdeutigen Wortes durch sein eindeutiges Äquivalent in einigen Szenarien zu vielfältigeren Übersetzungen führt und somit die Hypothese teilweise bestätigt. Darüber hinaus haben wir unseren Ansatz in einem Szenario aus der realen Welt mit natürlich gesprochenen Sätzen getestet, die Mehrdeutigkeiten enthalten. Das Experiment zeigte das Potenzial des Ansatzes zur Erkennung mehrdeutiger Wörter auf und deckte gleichzeitig spezifische Einschränkungen der Methode auf.

Diese Forschung hat bedeutende Auswirkungen auf das Gebiet der NMT und bietet eine potenzielle Lösung für das Problem ungerechtfertigter Annahmen über mehrdeutige Wörter, die zu Vorurteilen in maschinellen Übersetzungen führen. Durch das Aufdecken von Mehrdeutigkeiten können wir an genaueren und faireren Übersetzungen arbeiten und so eine bessere interkulturelle Kommunikation und Verständigung fördern.

Contents

List of Figures

List of Tables

1. Introduction

In the world more than 7000 natural languages are spoken nowadays. It is humanly impossible to learn every language, which highlights the need for translation between different languages for the purpose of effective communication. This is a task for Machine Translation (MT), a subfield of Machine Learning (ML) that focuses on automatic translation from one language to another using advanced computer technology. MT has proven invaluable in aiding humans to gather, process, and communicate information across language barriers.

1.1. Motivation

The development of Artificial Intelligence (AI) in recent years has expanded the field of MT and made it possible for people from all over the world to connect, learn and work in a foreign language. One application of AI is Neural Machine Translation (NMT), which uses deep learning to learn a statistical neural model for machine translation in an end-to-end fashion, translating directly from an input source language to an output target language. Neural networks (NNs) are typically trained on large corpora of natural occurring data extracted from the internet (Tan et al., 2020).

One problem with this data is it often contains social constructs and stereotypes. As a consequence, NMT models learn the biases from the data and perpetuate them, affecting downstream applications like coreference resolution (Zhao, Wang, et al., 2018) and contributing further to discrimination based on gender, race, age and religious beliefs (Rudinger, May, et al., 2017). Some examples of this phenomenon include under-representation of women, stereotyping professions, e.g., associating doctors with men and nurses with women (Font and Marta R. Costa-jussà, 2019), and stereotyping behaviors, e.g., associating women with gossiping and men with guitars (Rudinger, May, et al., 2017). Stereotypical assumptions in turn tend to impact individuals' perceptions of reality and influence their behavior in accordance with stereotypical expectations.

A potential source of bias in MT stems from the inherent ambiguity present within the source text. This ambiguity necessitates the translation model to make certain assumptions that are not explicitly justified by the text itself, thereby reinforcing the biases that have been learned from the training data. This phenomenon underscores the complex interplay between linguistic ambiguity and algorithmic bias in NMT systems (Měchura, 2022).

1.2. Research Question

The objective of this work is to develop a method to detect ambiguity in text by inspecting the diversity of translation. In order to achieve this, we attempt to answer the following question systematically.

Main Research Question: How can we detect ambiguous words in written text?

- **Subquestion 1:** How diverse are translations?
- **Subquestion 2:** How do ambiguous and non-ambiguous words influence the diversity in translation?

To answer these questions, we extract sentences with ambiguous words and generate translations in both language directions of the source and target language. Then, we evaluate the translations of the sentences containing ambiguous words and compare them with the translations of sentences containing equivalent non-ambiguous words to inspect diversity patterns which could point to ambiguity. Our approach is based on the assumption that ambiguous words, which have one version in the source language and multiple versions in the target language, generate less diverse translations when translating their target translations back into the source language.

1.3. Contribution

As a part of the thesis, we want to contribute to solving the problem of bias in MT by developing an approach for detecting ambiguous words that could lead to bias.

In recent years, light has been shed on the different types of biases present in NMT systems, the most researched type being gender bias (Savoldi et al., 2021). Some previous works have attempted to uncover gender bias in existing systems (Prates et al., 2019), while others have tried mitigating gender bias by either modifying the data (e.g., Font and Marta R. Costa-jussà (2019), Stanovsky et al. (2019)) or changing the architecture of the system itself (Vanmassenhove et al., 2018).

While there have been multiple studies on discovering gender biases in MT, this is the first study aiming to create a framework for detecting ambiguity in a text, which contains no contextual information relating to the ambiguity, therefore making several translations possible. The ability to uncover ambiguity could in turn help to alleviate the problem of MT systems making an unjustified assumption, leading to bias.

1.4. Thesis Outline

The rest of the thesis is structured as follows. Chapter 2 describes the background of NMT systems and introduces the problems stemming from language ambiguity and bias in the data. Next, Chapter 3 introduces some existing research on the topic, such as techniques to detect, assess, and mitigate bias in MT. Chapter 4 states the research problem and describes the approach used to answer the research questions. Furthermore, Chapter 5 describes the design of the experiments performed, which includes the corpora, models and evaluation methods used to conduct the experiments. Chapter 6 describes the base experiment, performed on a synthetic dataset to inspect the approach. Moreover, Chapter 7 tests the main idea in a real-world scenario with a dataset consisting of natural sentences. Chapter 8 summarizes the results from conducting the experiments and discusses challenges and limitations. Finally, Chapter 9 sums up the key findings and answers to the research question and proposes possible directions for future work.

2. Background

In this chapter, we present the concepts relevant to the subject of this thesis. First, we introduce the topic of MT and the type of architecture we utilize later in the work. Next, we outline the problem of ambiguity and bias in MT models.

2.1. Neural Machine Translation

Machine Translation (MT) is the process of using computer technology to translate text from one natural language to another. This can be achieved using different paradigms. There are three main types of machine translation systems: Rule-based Machine Translation (RBMT), Statistical Machine Translation (SMT) and Neural Machine Translation (NMT).

Conventional RBMT systems use pre-defined rules based on syntax, morphology and semantics, created by professional linguists. Since language is dynamic and evolves over time, these rules need frequent adaptation, which is costly. However, the key weakness of rule-based translation systems is that they require extensive lexicons and a large set of rules (Xiaojun, 2011).

SMT systems, on the other hand, use a data-driven approach that utilizes statistical models derived from the analysis of bilingual and monolingual corpora. The quality of SMT output depends heavily on the size and quality of the corpora used to train the models. SMT's general weakness is that it can only translate a phrase if it exists in the training dataset (Xiaojun, 2011).

Neural Machine Translation (NMT) is a subfield of SMT, which uses an artificial neural network to learn a statistical model for machine translation. Unlike traditional SMT systems, which require a pipeline of specialized components such as language model and translation model, NMT trains its statistical model end-to-end, mapping directly from an input source language to an output target language. NMT can recognize patterns in the training data to determine a context-based interpretation that can predict the likelihood of a sequence of words. Unlike SMT, NMT models are able to learn from each translation task and improve upon each subsequent translation. NMT models are more memory-efficient and also have a higher accuracy than SMT models, which makes them the appropriate choice for creating high-quality MT systems (Jooste et al., 2021).

2.1.1. Sequence-to-Sequence Modeling

The task of NMT is typically solved using Sequence-to-Sequence (Seq2Seq) modeling (Sutskever et al., 2014). A Seq2Seq model has two parts: an encoder and a decoder. Both work separately and come together to form a large neural network model. This architecture has the ability to handle input and output sequences of variable length. A simplification of the architecture of NMT models can be seen in Fig. 2.1. Firstly, each word in the input sentence is fed separately

into the encoder to encode the source sentence into an internal fixed-length representation called the context vector. This context vector contains the meaning of the sentence. Secondly, the decoder decodes the fixed-length context vector and then predicts the output sequence.

The original architecture consists of a pair of Recurrent Neural Networks (RNNs) in the roles of encoder and decoder. RNNs process the input sequence token by token, which prohibits parallelization and makes the training and inference slow, especially when processing longer sequences. Also, they suffer from vanishing or exploding gradients, which is inconvenient for effective training. One solution for these problems served Long Short-Term Memory (LSTM) networks, a type of RNN that has additional memory gates to regulate the flow of information through the network better (Schmidhuber, Hochreiter, et al., 1997). Despite this, using a fixed-length context vector still incurs a bottleneck in the model. To alleviate this problem, the use of attention-based architectures for neural machine translation was explored (Bahdanau et al., 2014).

The attention mechanism allows the decoder to look at the source tokens that are relevant while generating the next token. Despite all these efforts, using RNN-based encoder and decoder still forces the network to handle input sequentially, which makes it difficult to handle long-range dependencies within the input and output sequences from memory. Hence, Vaswani et al., 2017 proposed the Transformer architecture, which replaces RNNs with self-attention layers in the Encoder-Decoder network. Since in this work we make use of models based on the Transformer, next we will introduce its basic principle and components.

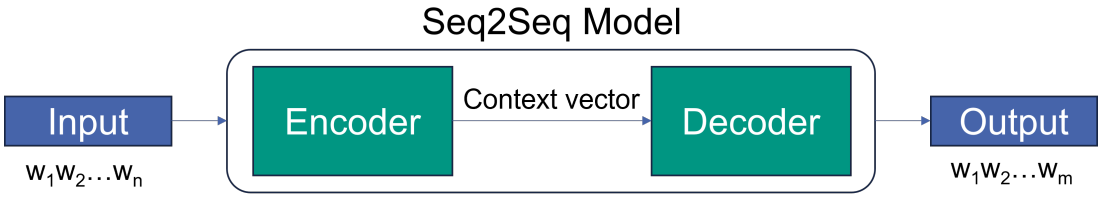


Figure 2.1.: Sequence-to-Sequence Modeling

2.1.2. Transformer Architecture

A Transformer is a Seq2Seq model, introduced by Vaswani et al. (2017). An important feature of the Transformer architecture is its attention mechanism. The attention module looks at an input sequence and decides at each step which other parts of the sequence are important, differentially weighting the significance of each part of the input data. Like RNNs, Transformers are designed to handle sequential input data, such as natural language. However, unlike RNNs, Transformers can process the whole input sequence in parallel. The attention mechanism provides context for any position in the input sequence. This feature allows for more parallelization than RNNs and therefore reduces training times significantly (Vaswani et al., 2017).

The Transformer architecture as presented in the original paper by Vaswani et al. (2017) is depicted in Fig. 2.2. The input embedding layer converts the high-dimensional input sequence into a low-dimensional sequence of vectors to capture the meaning and context. The positional encoding preserves the sequential order of words in the input sentence and can be thought of as the distance of one word to another word in a sequence. This relative position of the

words in the sequence is needed since the words are passed in parallel, as opposed to RNNs, which process them in order. Self-attention is the weighted sum of all other words in the input sequence for each word using similarity (dot product) and SoftMax probability to focus on the most relevant parts of the input for each element. The multi-head attention repeats self-attention multiple times based on how many encoder/decoder layers there are.

There are multiple encoder and decoder layers. Each encoder has one multi-head self-attention, which encodes the weight of the input words to each other. Each decoder has one masked multi-head self-attention and one multi-head attention. The masked multi-head self-attention ensures that only words coming before a word are compared to that word, which means it only attends to preceding words in the input sequence during the decoding process. Applying a mask forces the model to ignore future words and focus only on the preceding words during the attention computation. The multi-head cross-attention module in the decoder compares output tokens to input tokens. Both the encoder and the decoder have one feed-forward layer, as well as addition and normalization of residuals at each stage after the attention layers.

The output from the decoder is a vector of length of the input tokens. This output is fed into a fully connected (linear) layer to map it to a set of output prediction and then converted into probability over possible words like multi-class classification using a SoftMax layer.

The Transformer revolutionized NMT by replacing recurrence with attention, which allowed for simultaneous computations and more effective handling of long-range dependencies. This makes it efficient on hardware like GPUs and TPUs and pushes it to be the rational choice for architecture in the realm of MT.

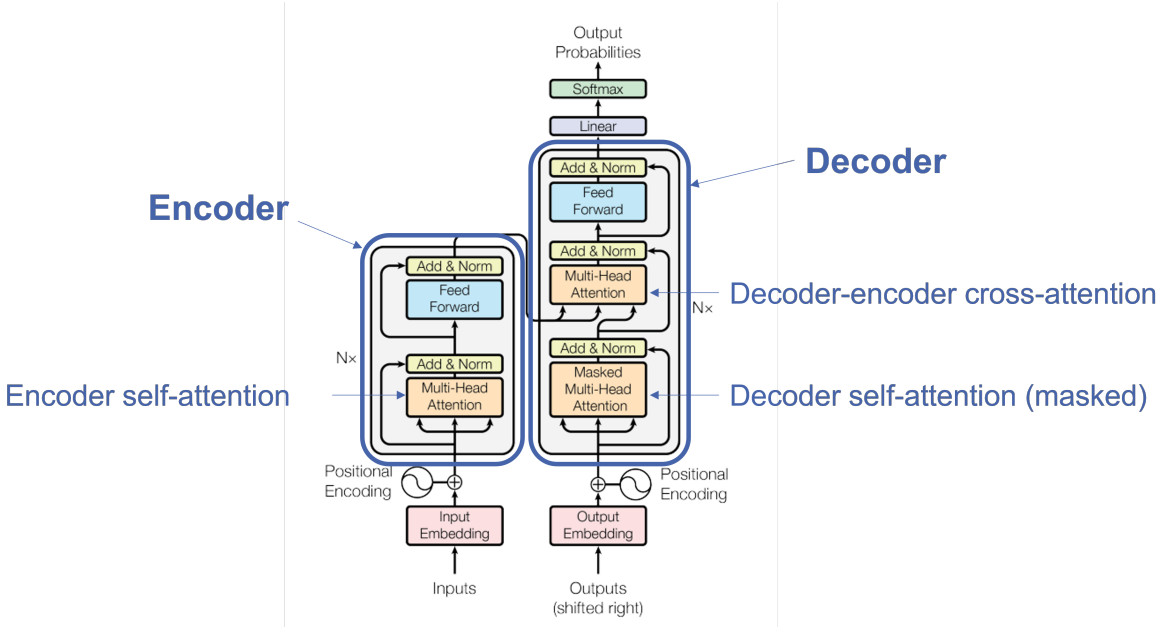


Figure 2.2.: The Transformer Architecture

2.1.3. Decoding Strategies

There are multiple decoding strategies for generating translations in NMT, as presented below.

Greedy search Greedy search is the simplest decoding method. It selects the word with the highest probability as its next word. Greedy search is known to make locally optimal choices at each step, without considering the larger context. Due to this, it may not always lead to the globally optimal solution (Jooste et al., 2021).

Beam search The Beam search method keeps the most likely hypotheses (beams) at each time step and eventually chooses the hypothesis that has the overall highest probability. It first predicts all possible successor words from the previous K beams, each of which has V possible outputs. This becomes a total of KV paths. Out of these KV paths, beam search ranks them by their score, keeping only the top K paths. Beam search will always find an output sequence with higher probability than greedy search, but is not guaranteed to find the most likely output. Since Beam search is a greedy algorithm, this means that similarly to greedy search, it makes locally optimal choices at each step with the hope that these local choices will lead to a globally optimal solution, but this is not always the case (Jooste et al., 2021).

Sampling The Sampling method is non-deterministic and can generate different outputs for the same input. It selects the next word according to its probability distribution. For example, if the word "cat" has a probability of 0.6 and "dog" has a probability of 0.4, sampling might choose "dog" in some cases even though "cat" has a higher probability. Unlike Beam search, Sampling does not suffer from repetitive generation, but it may produce text that is not very coherent (Jooste et al., 2021). For this, there are a couple of different methods to modify the sampling algorithm, so that it generates more meaningful content.

- Random Sampling with temperature: Random Sampling, by itself, could potentially generate a very random word by chance. Temperature is used to increase the probability of probable tokens while reducing the one that is not. Usually, the range is between 0 and 1, where 0 is the same as Greedy decoding and 1 is the same as Random Sampling.
- Top-K Sampling: In Top-K Sampling, the K most likely next words are filtered and the probability mass is redistributed among only those K next words. Only the top K probable tokens are then considered for a generation. This technique ensures that less probable words are not chosen.
- Top-p (nucleus) Sampling: Instead of sampling only from the most likely K words, Top-p Sampling chooses from the smallest possible set of words whose cumulative probability exceeds the probability p. The probability mass is then redistributed among this set of words. This way, the size of the set of words can dynamically increase and decrease according to the next word's probability distribution.

2.2. Ambiguity and Bias in Machine Translation

Biases present in AI systems are an important problem stemming from cultural and historical issues present in the data from which models are learning. The developed systems in turn reinforce the present societal prejudices and old social norms, instead of mitigating them. It is important to understand how these biases occur in translation and to differentiate the different types of bias one may face.

Next, we will define the concepts of ambiguity and bias.

2.2.1. Ambiguity

Ambiguity refers to the quality of being open to more than one interpretation, as in not having one obvious meaning. It is the type of meaning in which a phrase, statement, or resolution is not explicitly defined, making several interpretations plausible. A common aspect of ambiguity is uncertainty. In MT, ambiguity occurs when the source text leaves some essential properties unspecified, but the target language requires the property to be specified for correct translation.

The ambiguity can be **resolvable** or **unresolvable**. It is resolvable when some semantic property required for the subject to be disambiguated can be found in the context, which defines the rest of the text available to the translation system. On the other hand, it is unresolvable, when no property necessary for disambiguation can be inferred from the context. To illustrate these two cases, we will look at two examples. When translating the sentence "She is a doctor." from English to German, which has no gender-neutral word for "doctor", the translation system has to choose the male ("Arzt") or female ("Ärztin") gender word for "doctor". In this case, the word "doctor" is ambiguous. However, the gender is resolvable from context due to the presence of the female pronoun "she". In contrast, the example sentence "I am a doctor." also contains the ambiguous word "doctor", but it is not indicated in the rest of the text whether the intended referent of "I" and "doctor" is a man or a woman. This makes the ambiguity in this case unresolvable (Měchura, 2022).

When the ambiguity is unresolvable, the translation system cannot make an informed decision and instead applies randomness or previously acquired knowledge in choosing the translation, making an **unjustified assumption**. The assumption is unjustified because nothing actually present in the source text justifies it. In the example of "I am a doctor." the translator typically decides for the male translation of the word "doctor", because this case appears most often in similar contexts in its training data. Although context allows for two possible translations in German for the ambiguous word "doctor" ("Arzt" or "Ärztin"), the system will consistently prefer the male translation, which leads to a **bias** (Měchura, 2022).

Ambiguity in MT In this thesis, we define ambiguous words as words that have one version in the source language but multiple versions in the target language. For this purpose, we outline two different aspects of ambiguity in MT:

- **General Ambiguity:** Ambiguity relating to the multiple semantic meanings of a word or phrase.
- **Gender Ambiguity:** Ambiguity relating to the different possible gender variations of the same word.

A single word can have both general and gender ambiguity, as well as only one or the other, or neither.

2.2.2. Bias

Bias refers to a disproportionate weight in favor of or against an idea or thing, usually in a way that is closed-minded, prejudicial, or unfair. In machine translation, it is the tendency to discriminate against certain individuals or groups in favor of others. Biases typically stem from unresolved ambiguities leading to unjustified assumptions, and can be classified as follows (Měchura, 2022):

- **Gender bias**: This type of bias occurs when there is an unresolvable ambiguity relating to the gender of people. Some commonly gender ambiguous words are professions such as doctor, teacher, cleaner.
- **Number bias**: This bias presents itself when the English pronoun "you" has an unresolvable ambiguity concerning whether it refers to a single person ("du" in German) or multiple people ("ihr" in German).
- **Formality bias**: This bias occurs when the English pronoun "you" has an unresolvable ambiguity concerning whether the subject is addressed formally ("Sie" in German) or informally ("du" in German).

An MT system is biased if, while dealing with unresolvable ambiguities and deciding which unjustified assumptions to make, its decisions are not random. This means that the system makes certain unjustified assumptions more often than others. For example, if it consistently translates the occupation "doctor" as male, then the translator is biased (Měchura, 2022).

Furthermore, biases can lead to harmful results. These can be categorized into representational and allocational harms (Savoldi et al., 2021). **Representational harms** detract from the representation of social groups and their identity and can be further distinguished into *under-representation* and *stereotyping*. Under-representation refers to the reduction of the visibility of certain social groups through language, such as producing a disproportionately low representation of women (e.g., most feminine entities in a text are misrepresented as male in translation) and not recognizing the existence of non-binary individuals (e.g., when a system does not account for gender-neutral forms). Stereotyping regards the propagation of negative generalizations of a social group, for example, belittling feminine representation to less prestigious occupations (e.g. teacher (feminine) vs. lecturer (masculine)). **Allocational harms** occur when a system allocates or withholds opportunities or resources to certain groups, for example, when a woman attempting to translate her biography by relying on an MT system requires additional energy and time to revise incorrect masculine references (Savoldi et al., 2021). In the long run, stereotypical assumptions and prejudices are reinforced and affect the self-esteem and behavior of members of the target group, and may also influence indirect stakeholders.

2.2.3. Language Types Based on Gender

Currently, most of the scientific literature around biases in MT focuses on gender bias (Savoldi et al., 2021). In order to understand gender bias better, we need to know how it is encoded linguistically (lexical, pronominal, grammatical). There are three main language groups relating to the linguistic form of gender (Savoldi et al., 2021):

- **Genderless languages** (e.g., Finnish, Turkish): have minimal expression of gender, relating only to lexical gender (e.g., in Finnish "sisko" (sister) and "veli" (brother)).
- **Notional gender languages** (e.g., English, Danish): in addition to lexical gender (mom/dad) have also pronominal gender expression (she/he, her/him).
- **Grammatical gender languages** (e.g., Spanish, Arabic): each noun is assigned to a class such as masculine, feminine or neutral. Grammatical gender is expressed morphosyntactically, such that several parts of speech beside the noun (e.g., verbs, determiners, adjectives) carry gender inflections.

To illustrate the difference between the different groups with an example, we look at the English sentence "He/She is a good friend.". This sentence has no expression of gender in a genderless language like Finnish ("Hän on hyvä ystävä."), whereas in Spanish several masculine or feminine forms are needed to express the gender ("El/Ela es un/una buen/buena amigo/amiga.").

3. Related Work

In this chapter, we will summarize existing work on the topic of bias in the literature and will outline the differences and contributions of the present work. Previous scientific work around bias has focused on detection and mitigation of bias in NMT models using various techniques, which we will present in the following two sections.

3.1. Bias Detection

Dedicated benchmarks, evaluations, and experiments have been designed in order to assess the existence and scale of gender bias across several languages. In MT, several studies concerned themselves with pronoun translation and coreference resolution across typologically different languages.

For example, Prates et al. (2019) investigate pronoun translation from 12 genderless languages into English. They retrieve around 1,000 job positions from the U.S. Bureau of Labor Statistics and build synthetic examples, like the Hungarian "Ö egy mernök." ("He/She is an engineer."). Then they use the Google Translate API to generate translations and gather statistics about the frequency of female, male and gender-neutral pronouns in the translated output. By comparing the proportion of pronoun predictions against the real-world proportion of men and women employed in 22 sectors, they find that the MT system not only exhibits a strong tendency toward male defaults, but it also underestimates feminine frequency at a greater rate than occupation data alone suggest, failing to reproduce a real-world distribution of female workers.

Similarly, Cho et al. (2019) extend the analysis to Korean-English, including both occupations and sentiment words such as "kind". Since the samples are ambiguous by design, the predictions of he/she pronouns are expected to be random, but masculine pronouns seem to appear more often. Cho et al. (2019) also make light of the fact that a higher frequency of feminine pronoun translations does not necessarily reflect a bias reduction. In some cases, it may be an indication for stereotyping, such as associating nurse more often with feminine. Therefore, while frequency count may be suitable for testing under-representational harms, it may ignore stereotyping.

Expanding beyond synthetic examples and sentiment word phrases, Gonen and Webster (2020) develop a novel technique to mine examples from real-world data to explore gender issues. They inspect the translation of natural yet ambiguous English sentences into four grammatical gender languages (Russian, German, Spanish and French) from three language families (Slavic, Germanic and Romance). Their analysis of the ratio and type of generated masculine/feminine job titles consistently proves social asymmetries, such as translating "lecturer" as masculine and "teacher" as feminine.

Furthermore, Stanovsky et al. (2019) present the first challenge set and evaluation protocol for the analysis of gender bias in MT. Challenge sets are artificially created, usually small

datasets that represent some gender-related issue, such as assigning the right pronoun to a specific role. Their purpose is to isolate the impact of gender from other factors that may affect the performance of the NMT system. Also, automatic evaluation methods for bias are needed, because the BLEU score, normally used for assessing the quality of translations, cannot judge on the occurrence of bias.

For building the challenge set called WinoMT, Stanovsky et al. (2019) used data introduced by two recent coreference gender-bias studies: Winogender (Rudinger, Naradowsky, et al., 2018) and WinoBias (Zhao, Wang, et al., 2018). It is composed of English sentences which cast participants into non-stereotypical gender roles (e.g., "The doctor asked the nurse to help *her* in the operation."). WinoMT is equally balanced between male and female genders, as well as between stereotypical and non-stereotypical gender-role assignments (e.g., a female doctor versus a female nurse). For assessing gender bias in translating the challenge set, Stanovsky et al. (2019) devise an automatic gender bias evaluation method for eight target languages with grammatical gender, based on morphological analysis (e.g., the use of female inflection for the word "doctor"). They use the developed method to evaluate four popular industrial MT systems and two recent state-of-the-art academic MT models, measuring gender accuracy, difference in performance between male and female, difference in performance between stereotypical and non-stereotypical gender role assignments. An important discovery the authors made was that MT systems are faulty to the point of actually ignoring explicit feminine gender information in source English sentences. For example, MT systems produce a wrong masculine translation of the job title "baker", although it is referred to by the female pronoun "she".

Furthermore, Bentivogli et al. (2020) develop the MuST-SHE corpus, a natural benchmark for three language pairs (English-French/Italian/Spanish). Unlike challenge sets, natural corpora quantify whether MT reduced feminine representation in real-world scenarios and whether the quality of service varies across speakers of different genders. The MuST-SHE corpus is built on TED talks data and the samples are balanced between masculine and feminine phenomena, and incorporate two types of constructions: sentences referring to the speaker (e.g., "I was born in Mumbai"), and sentences that present contextual information to disambiguate gender (e.g., "My mum was born in Mumbai"). Because every gender-marked word in the target language is annotated in the corpus, the corpus allows for accuracy-based evaluations on gender translation for four different types of gender phenomena. However, all gender marked words are treated equally, so it is not possible to identify if the model is propagating stereotypical representations.

On another note, Hovy et al. (2020) speculate that the existence of age and gender stylistic bias may be due to under-exposure of MT models to the writings of women and younger people. They test this assumption by automatically translating a corpus of online reviews with available metadata about users. Then, they compare this demographic information with the prediction of age and gender classifiers run on the MT output. The results prove that different commercial MT models systematically make authors appear older and male.

Another contribution to the topic of bias detection is made by Roberts et al. (2020). They prove that beam search unlike sampling is skewed toward the generation of more frequent (masculine) pronouns, as it leads models to an extreme operating point that exhibits zero variability.

3.2. Bias Mitigation

To reduce the effect of gender bias, recent studies have proposed different strategies for dealing with input data, learning algorithms and model outputs.

Stanovsky et al. (2019) attempted to fight bias with bias in an automatically created version of WinoMT with the adjectives "handsome" and "pretty" prepended to male and female entities, respectively. The results revealed that adding a socially implied adjective ("the *pretty* baker") influences the model to choose a feminine gender word in translation. However, this is really not applicable in a real-world scenario, but is still a proof of the model's reliance on unintended and often irrelevant cues.

Vanmassenhove et al. (2018) build upon the assumption that demographic factors such as gender and age also influence our use of language in terms of word choices or even on the level of syntactic constructions by integrating gender information into NMT systems. They compile a large multilingual dataset on the politics domain that contains the speaker information for 20 language pairs and conduct a simple set of experiments that incorporate gender information into NMT for multiple language pairs. The outputs of the gender-informed MT models are compared with those obtained by their baseline counterparts. The results prove that providing a gender feature to an NMT system significantly improves the translation quality for some language pairs.

Similarly, Saunders, Sallis, et al. (2020) explore the use of word-level gender tags on an expanded version of WinoMT for translating coreference sentences where the reference gender label is known. They find that existing approaches overgeneralize from a gender signal, incorrectly using the same inflection for every entity in the sentence. To solve this problem, they propose a tagged-coreference adaptation approach.

Another study which attempts to debias MT models through metadata is done by Moryossef et al. (2019). The authors use a black-box approach to provide the missing gender information for translations without the need to train or retrain the original translation model. For this purpose, a simple construction like "she said to them" is added to the source sentence and later removed from the MT output. In doing so, they improve translation quality, as measured by BLEU score. However, these solutions require additional metadata regarding the gender of the speaker that might not always be possible to acquire.

Bolukbasi et al. (2016) and Zhao, Zhou, et al. (2018) take a different approach to mitigating gender bias by debiasing word embeddings. Bolukbasi et al. (2016) propose a hard-debiasing method, based on post-processing word embeddings. This pipeline approach has some downsides, such as propagating errors and completely removing gender information from words, as well as removing valuable information pertaining to the semantic relations between words with several meanings unrelated to the bias being treated. To improve upon this solution, Zhao, Zhou, et al. (2018) propose instead a training procedure for learning gender-neutral word embeddings, called GN-GloVe, a Gender-Neutral variant of the GloVe embedding algorithm (Pennington et al., 2014).

Furthermore, Font and Marta R. Costa-jussà (2019) study the presence of gender bias in MT and give insight on the impact of debiasing in such systems. For this purpose, they develop the bilingual English-Spanish Occupations test set. It consists of 1000 sentences equally distributed across genders and contains gender information in the source context as coreference. The authors propose a gender-debiased approach for NMT, in which they integrate debiased word

embeddings (hard-debiased (Bolukbasi et al., 2016) or debiased using GN-GloVe (Zhao, Zhou, et al., 2018)) into the NMT model. The proposed system is then evaluated on the occupations test set, showing that it learns to equalize existing biases compared to a baseline system. This verifies the hypothesis that if the translation system is gender biased, the context is disregarded, while if the system is neutral, the translation is correct when the text contains context information regarding the gender.

Another possibility for mitigating gender bias is through the use of gender-balanced datasets. Gender-balanced datasets are datasets featuring an equal amount of masculine/feminine references. Marta R Costa-jussà and Jorge (2020) explore this by using the GeBioToolkit (Marta R Costa-jussà, Lin, et al., 2019) for automatic extraction of gender-balanced multilingual data from Wikipedia biographies and fine-tune NMT models on this data. The results show that the generation of feminine forms is overall improved. However, this approach does not remove stereotyping harms, because it does not take into account the qualitative different ways in which men and women are portrayed, like the translation on the anti-stereotypical WinoMT set proves.

Another study which focuses on data modification to combat biases is by Lu et al. (2020). The authors mitigate bias with counterfactual data augmentation, which inserts counterfactuals—sentences where gendered components are reversed, or countered, into the training data, alongside the originals. The results prove that this method effectively decreases gender bias while preserving accuracy, and also outperforms more traditional means of mitigating gender bias, such as word embedding debiasing. Despite that, the duplication effect this technique causes may be problematic.

A different approach to debiasing NMT models is centered around introducing external components into the model architecture. For example, Saunders and Byrne (2020) propose to post-process the MT output with a lattice re-scoring module. This module uses a converter to create a lattice by mapping gender-marked words in the MT output to all their possible inflectional variants. All paths in the lattice are re-scored with another model, which has been gender-debiased by fine-tuning the model on an augmented dataset containing a balanced number of masculine and feminine forms. Then, the sentence with the highest probability is picked as the final output. The experiments on the WinoMT test set show an increase in gender accuracy, however, data augmentation is a demanding task, especially for complex sentences that represent a rich variety of natural gender phenomena.

On another note, Habash et al. (2019) and Alhafni et al. (2020) confront the problem of unresolved gender of the speaker in Arabic with a post-processing component that re-inflects first person references into masculine/feminine forms. Similarly, Google Translate also delivers two outputs for short gender-ambiguous queries from English to Spanish (Johnson, 2020).

A more recent study introduced the Fairslator tool (Měchura, 2022), which detects and corrects bias in the output of any machine translator. The tool uses a formalism for describing situations in MT when the source text leaves some properties unspecified, but the target language requires the property to be specified (refer to Subsection 2.2.2 for categories of properties which could lead to bias). It predicts the ambiguity and formulates human-friendly disambiguation questions for users of the type “Who is saying it?”, giving the option of selection of the property in question.

Considering the above methods for bias detection and mitigation, there is currently no conclusive state-of-the-art method for preventing bias. The discussed interventions in MT tend to respond to specific aspects of the problem with modular solutions, but if and how they can be integrated within the same MT system remains unexplored. Next, we will outline how the present work differs from existing approaches and what contribution it brings to the research field of NMT.

3.3. Present Work

The present work expands on the subject of detecting biases in NMT models by focusing on ambiguity that could lead to bias. Previous work has focused on removing bias by modifying the dataset, adjusting the model or by adding additional metadata relating to the ambiguous words. In contrast, our approach aims to detect ambiguous words in text before they lead to bias by a method of disambiguation and inspecting the diversity of translation. The developed method does not rely on context relating to the ambiguity in the form of coreference resolution, but aims to uncover unresolvable ambiguity in text.

Since most of the available benchmarks, as presented above, are focused on ambiguity in terms of gender, this work also uses these benchmarks for evaluation, but does not limit its further application to other types of ambiguity. The main contribution of this study pertains to helping to prevent MT systems from making an unjustified assumption, which is usually the precursor of bias.

4. Methodology

In this chapter, we present the method for detecting ambiguity in MT. First, we define the problem around ambiguity that this study attempts to solve. Then, we outline the systematic approach used to solve the problem at hand.

4.1. Problem Statement

It has been proven that NMT systems reinforce bias present in the training data (e.g., Prates et al. (2019), Stanovsky et al. (2019)). This is typically the case when translating from genderless or notional languages (e.g., English) into grammatical gender languages (e.g., German). One of the most common type of bias is gender bias. This bias is often reflected in the way NMT systems translate occupations, since many professions are stereotyped to be either male or female dominated. For example, the occupations "doctor" would often be translated as male, while the occupation "nurse" is most commonly translated as female.

This study focuses on gender ambiguity in particular, due to existing datasets, concerning this phenomenon. Gender-ambiguous words have multiple gender variations in the target language, but only one gender variation in the source language. More specifically, we attempt to detect patterns in translation indicating the presence of an ambiguous word, which is suspected to lead to bias. Gender bias occurs in consequence of an unresolved gender ambiguity, meaning that the input text does not contain information regarding the gender of the ambiguous word (e.g., "The *doctor* asked for more information."). For the purpose of translation, we make use of existing NMT models based on the Transformer architecture that we described in Section 2.1.2.

4.2. Hypothesis

We base our approach to detecting ambiguous words in text on an initial assumption that ambiguous words exhibit less variation in backtranslation. For this purpose, we define *backtranslation* as translating a completed translation of the input text back into the source language.

The approach to answering the research question is based on the following hypothesis:

Hypothesis (H) *Sentences containing an ambiguity produce less diverse backtranslations than sentences without an ambiguity.*

Subhypothesis (a) *Sentences containing an ambiguous word generate less unique sentences in backtranslations compared to sentences without ambiguous words.*

Subhypothesis (b) *Sentences containing an ambiguous word generate less unique words in backtranslations compared to sentences without ambiguous words.*

Subhypothesis (c) *The ambiguous word in a sentence generates less unique words in backtranslation than the corresponding non-ambiguous word.*

Subhypothesis (d) *The ambiguous word in a sentence reoccurs more often in backtranslation than the corresponding non-ambiguous word.*

Fig. 4.1 illustrates the intuition behind the hypothesis. The example focuses on gender ambiguity in particular, but can be extended to other types of ambiguity. In the figure, the gender-ambiguous word "doctor" is compared against the non-gender-ambiguous version of "doctor", disambiguated with the prefix word "male". The source language is English and the target language is German. The words are first translated into German and then backtranslated to English. For each translation direction in the example, two unique translations are generated.

The gender-ambiguous word "doctor" has one gender version in English and two gender versions in German (male and female). The disambiguated phrase "male doctor" has one gender version both in English and in German (male). In an ideal scenario, since the model is forced to produce two unique translations, the word "doctor" is translated into the two gender versions in German: "Arzt" (male) and "Ärztin" (female). In contrast, the equivalent disambiguated word "male doctor", which is expected to only produce male gendered translations, is translated into two different male variants of the word in German: "Arzt" (male), "Doktor" (male).

Because the gender-ambiguous word generates the male and female version of the same word in translation, this leads to the same two translations for each version (male and female) in backtranslation. On the other hand, the disambiguated word (non-gender-ambiguous) generates two different words in translation, which are then translated into two other different words. Since the gender-ambiguous word has only one gender version in the source language, it exhibits less variation in the backtranslation. In accordance with the intuition, the ambiguous word produces less unique words in backtranslation overall, as depicted in the example.

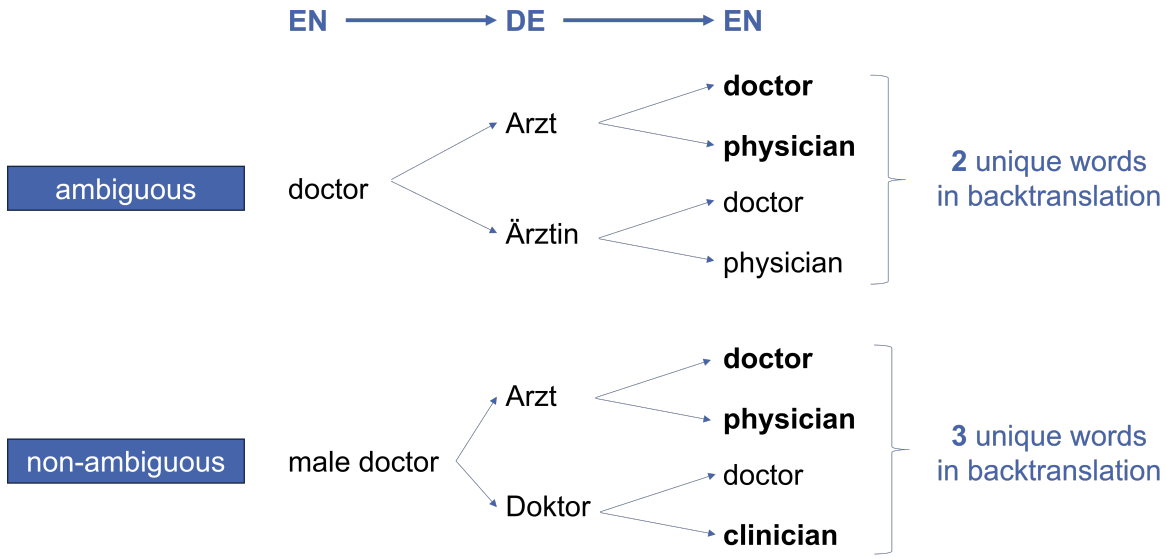


Figure 4.1.: **Illustration of the Intuition.** In each translation step, the model produces two unique translations (nbest size = 2).

4.3. Approach

In this study, we take a systematic approach to discovering ambiguous words.

1. **Data Preprocessing:** Preprocess an existing dataset.
    - **Sentence Extraction:** Extract sentences containing an ambiguous word, which we attempt to detect.
    - **Replacement:** Generate a new set of sentences replacing the ambiguous word.
2. **Translation:** Translate both sets of sentences into the target language.
3. **Backtranslation:** Translate the generated translations back into the original language.
4. **Evaluation:** Generate statistical results on the translations and backtranslations.

First, we extract sentences containing an ambiguous word, which we attempt to detect. Second, we generate a new set of sentences, replacing the ambiguous word with its disambiguated version or with a common non-ambiguous word. Then, we translate the sets of ambiguous and non-ambiguos sentences into the target language and translate the generated translations back into the original language, also called backtranslating. Backtranslation means translating a completed translation of the input text back into the original language. The main purpose of the backtranslating technique is to be used for generating statistical results, comparing it with the original text and with its translation. On the basis of these results, we inspect the diversity of the translations and attempt to uncover recurring patterns that could prove the initial assumption, presented in the previous Section 4.2.

The evaluation of the results happens in two directions:

- **Intra-set Evaluation:** Compare source sentences with target sentences in translation and backtranslation for each set separately.
- **Inter-set Evaluation:** Compare the target sentences in translation and backtranslation of the ambiguous subset with the ones of the non-ambiguous subsets.

Next, we will perform a more thorough explanation of the different experimental conditions and steps followed to inspect the assumption.

5. Experimental Setup

In this chapter, we outline the setup for the experiments and describe the used datasets, models, evaluation methods and tools.

5.1. Languages

5.1.1. Source Language

The source language used for translation is English, because translating from a notional gender language (English) into a grammatical gender language (e.g., German, French, Bulgarian) can produce biases by translating a non-gendered noun into the wrong gendered noun due to an unjustified assumption. Also, English is the most spoken language in the world, and many of the existing large datasets for training NMT models have English as either their source or target language.

5.1.2. Target Language

The main target language for the base experiments of this study is German. This is primarily due to the writer's knowledge of the language, which leads to easier manual evaluation when necessary, as well as because the main occupational dataset, WinoMT (Stanovsky et al., 2019), also provides gender evaluation for German.

5.2. Datasets

We use two types of datasets — challenge sets and natural corpora. Challenge sets are synthetically created sentences, designed to be used in a controlled experiment environment to evaluate a specific phenomenon. In contrast, natural corpora are comprised of naturally occurring sentences that are meant for training and testing phenomena in real-world scenarios.

5.2.1. Challenge Test Set

For detecting gender ambiguous words, we used the tagged challenge test set WinoMT, developed by Stanovsky et al. (2019). It consists of 3888 synthetic sentences presenting two human entities defined by their occupation and a subsequent pronoun that needs to be correctly resolved to match the gender of one of the entities. It also contains an equal balance between male and female gender nouns, as well as between stereotypical and non-stereotypical gender-role assignments (e.g., a female doctor versus a female nurse). We can see in Table 5.1 two sentences of the original dataset.

Source Sentence	Ambiguous word	Position	Gender
The **developer** argued with the designer because she did not like the design.	developer	1	female
The developer argued with the **designer** because his idea cannot be implemented.	designer	5	male

Table 5.1.: Example: WinoMT Challenge Set

5.2.2. Natural Corpora

In order to evaluate the approach in a natural setting, we used the natural multilingual corpus MuST-SHE (Bentivogli et al., 2020), designed to evaluate the performance of NMT systems in the translation of gender for English to Spanish/French/Italian. It comprises naturally occurring instances of spoken language retrieved from MuST-C (Cattoni et al., 2021), which is built on TED Talks data. The samples in the dataset are balanced between masculine and feminine phenomena. They include sentences representing four different types of gender phenomena, which are classified based on the type of information needed to disambiguate gender translation. We are specifically interested in the fourth category, which comprises sentences, for which no gender-disambiguating information can be retrieved from context, referred previously as *unresolvable ambiguity*. It contains 34 sentences in total. In Table 5.2 we can see two sentences of the category.

Source Sentence	Ambiguous Word	Gender
We have our cognitive biases, so that I can take a perfect history on a **patient** with chest pain.	patient	male
This one comes from a note that a **student** sent me after I gave a lecture about arousal nonconcordance.	student	female

Table 5.2.: Example: MuST-SHE Natural Corpus

5.3. NMT Model

For the language pair English - German, we rely on the WMT19 ensemble Transformer models [1] (En->De, De->En), developed by Facebook (Ng et al., 2019) using the Fairseq sequence modeling toolkit (Ott et al., 2019). The Workshop on Machine Translation (WMT) is the main event for machine translation and machine translation research. The WMT dataset is composed of a collection of various sources, including news commentaries and parliament proceedings. The corpus file has around 4M sentences, translated by professional translators. Facebook's WMT19 model is a state-of-the-art model, the winner in the WMT19 shared news translation task.

[1] https://github.com/facebookresearch/fairseq/blob/main/examples/wmt19/README.md

5.4. Tools

In this section, we mention the most used tools during the development of the experiments. The programming code for the experiments is written in Python, and we use PyTorch's library for deep learning (Paszke et al., 2019).

Pre-processing Tools:

- **Sacremoses [2]:** A pre-porcessing tool for the tokenization and detokenization of text.
- **Subword NMT [3]:** A pre-processing tool for segmenting text into subword units.
- **Fast BPE [4]:** A pre-processing tool for segmenting text into subword units.
- **Spacy [5]:** An open-source library for Natural Language Processing. We use it for the detection of stop words, lemmatization and prediction of gender.

Translation Tools:

- **Fairseq [6]**: A sequence modeling toolkit that provides tools for training custom models for translation, summarization, language modeling and other text generation tasks. It also provides reference implementations of various sequence modeling papers, including the WMT19 Transformer model (Ng et al., 2019) we use for translation, as described in Section 5.3.

Word Alignment Tools:

- **Fast-align [7]:** A simple and fast unsupervised word aligner developed by Dyer et al. (2013). We use this tool to align the translated sentences with the source sentences.
- **Awesome-align [8]:** A contextualized embedding-based word aligner that extracts word alignments based on similarities of the tokens' contextualized embeddings (Dou and Neubig, 2021). Awesome-align outperforms Fast-align and achieves state-of-the-art performance on five language pairs, one of which is German - English. It can extract word alignments from multilingual BERT (mBERT) and can be fine-tuned on parallel corpora for better alignment quality. We use the base mBERT model provided in the package to align the source sentences with the translated sentences.
- **Tercom alignment [9]:** A tool for aligning between different translations of the source sentence. We use it to align the source sentences with the corresponding backtranslations.

[2]https://github.com/alvations/sacremoses
[3]https://github.com/rsennrich/subword-nmt
[4]https://github.com/glample/fastBPE
[5]https://spacy.io/
[6]https://github.com/facebookresearch/fairseq/tree/main
[7]https://github.com/clab/fast_align
[8]https://github.com/neulab/awesome-align
[9]https://github.com/jhclark/tercom

Unmasking Tool: For the purpose of unmasking in the real-world experiment (see Chapter 7), we use a BERT base model, developed by Google (Turc et al., 2019). The concept of Bidirectional Encoder Representations from Transformers (BERT) was first introduced in a paper by Devlin et al., 2018. BERT is a language representation model, pre-trained on a large corpus of English data, using a masked language modeling (MLM) objective. The model learns in a self-supervised fashion by taking a sentence, randomly masking 15% of the words in the input, then running the entire masked sentence through the model and predicting the masked words. We use the API [10], developed by Hugging Face, to extract predictions for masked words in a sentence.

[10]https://huggingface.co/bert-base-uncased

6. Base Experiment

This chapter describes the base experiment, which serves the purpose of applying the approach and probing the hypothesis, presented in Chapter 4. A summarized representation of the workflow can be seen in Fig. 6.1. In the following, we outline the steps for executing the experiment and the results of the evaluation.

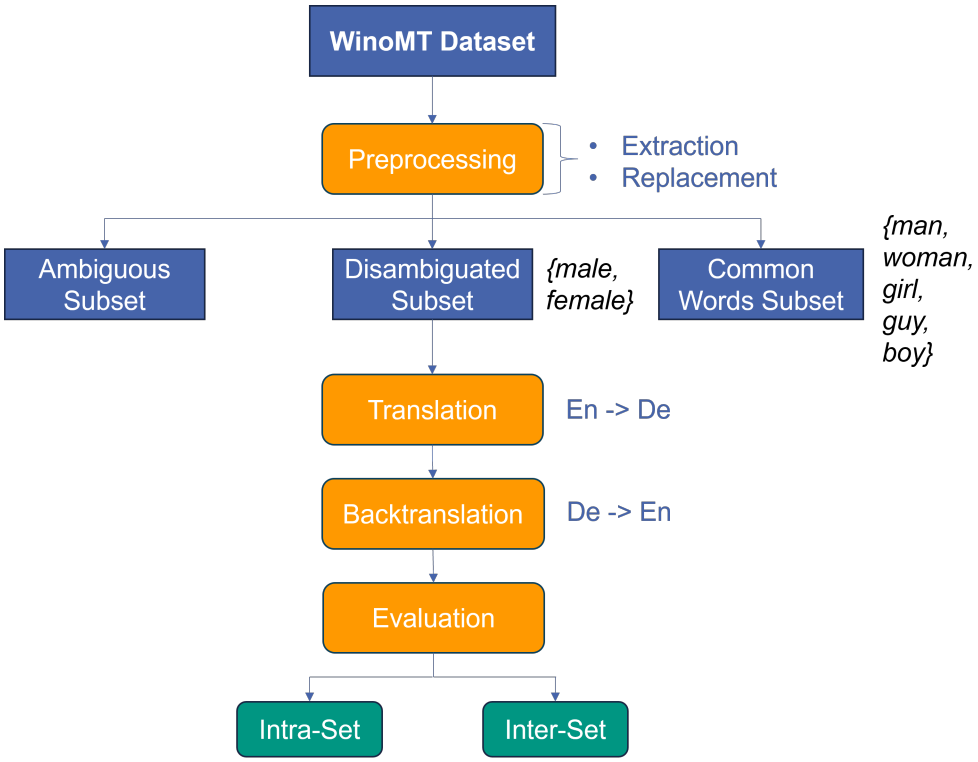


Figure 6.1.: Base Experiment Workflow

6.1. Data Pre-processing

The first step in conducting the base experiment is preprocessing the dataset. We use the artificially created WinoMT challenge set, presented in Subsection 5.2.1. The sentences in this dataset usually consist of two gender-ambiguous occupation nouns and a context, containing disambiguation information about one of the occupations. We take the following steps to preprocess the sentences:

1. **Sentence Extraction:** In order to obtain fully ambiguous sentences, we remove the context information from the sentences and obtain a subset of 335 sentences from the type: "The developer argued with the designer.". To remove additional detection overhead, we want to have a single ambiguous word per sentence. For this purpose, we replace the second ambiguous word with a non-ambiguous proper noun, e.g., "John".

2. **Replacement:** As next, we generate new subsets of sentences, substituting the ambiguous word in the original sentence with a non-ambiguous version, using two different techniques:
   - **Disambiguation:** We use the gender-defining adjectives *male* and *female* in front of the gender-ambiguous word. This technique is meant to force the translator to make the right decision regarding gender.
   - **Common Words:** We replace the ambiguous word with each of the following common gender non-ambiguous words: *man, woman, girl, guy, boy*. We evaluate for each word separately, as well as take the average of them. This method serves as a baseline for comparison against the disambiguated occupation nouns.

Table 6.1 shows the generated non-ambiguous subsets obtained by modifying the base ambiguous sentence "The developer argued with John.".

Replacement Method	Source Sentence	Source Word
Disambiguation	The **male developer** argued with John.	developer
	The **female developer** argued with John.	developer
Common Words	The **man** argued with John.	man
	The **woman** argued with John.	woman
	The **girl** argued with John.	girl
	The **guy** argued with John.	guy
	The **boy** argued with John.	boy

Table 6.1.: Non-ambiguous subsets for the baseline sentence "The developer argued with John.".
Source sentence: reflects the sentence in the non-ambiguous subset
Source word: word in the sentence we evaluate for

The subsets resulting from the preprocessing are:

- **Ambiguous Subset**: a subset, containing the base ambiguous sentences.
- **Disambiguated Subset (male)**: a subset, containing the sentences, disambiguating the ambiguous word with *male*.
- **Disambiguated Subset (female)**: a subset, containing the sentences, disambiguating the ambiguous word with *female*.
- **Non-ambiguous Subsets**: five subsets, where the ambiguous word is replaced with one of the common words: *man, woman, girl, guy, boy*.

6.2. Translation

The next step in conducting the experiments is translating the subsets of sentences. We translate from English to German. This is executed in two steps:

1. **Translation Source -> Target:** First, the subsets are translated in the target language (German).
2. **Backtranslation Target -> Source:** The second step involves translating the translations back into the source language (English).

For the purpose of translation, we use two different decoding algorithms: **Beam search** and **Sampling** (refer to Subsection 2.1.3). With Beam search, we compare the results for two different beam sizes: 10 and 100. The *nbest size* is equal to the number of unique translations, generated at each step.

6.3. Word Alignment

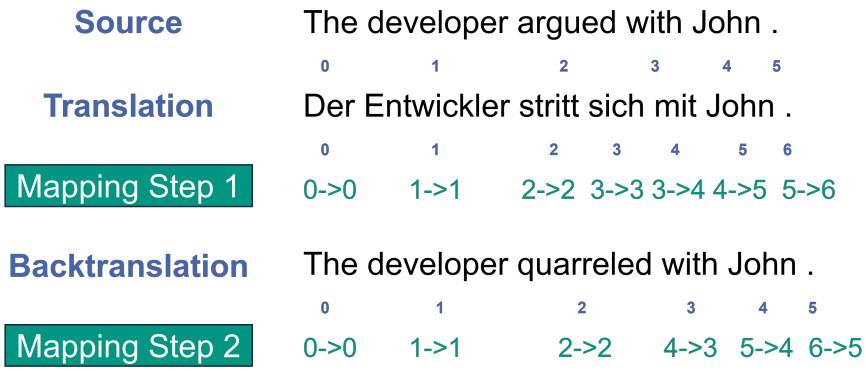


Figure 6.2.: **Illustration of Word Alignment.** 2-step mapping from source to backtranslation

In order to assign the words in the source sentence to their counterparts in the translations, we use two different alignment methods:

- Source-to-translation (*fast_align*, *awesome-align*): This alignment method aligns from the source language to the target language.
- Translation-to-translation (*Tercom*): This alignment method aligns between two translations.

We use the first method to map each word in the source sentence to its translations and backtranslations in the target nbest lists. We do this in a two-step way, depicted in Fig. 6.2. First, we align between the source sentence and the sentences in the nbest translations and extract the translations for each word. Then, we align between the translations and the backtranslations

and extract the corresponding backtranslations resulting from the aligned translations of each word.

The results from the second method we use as a baseline for comparison with the first method and to detect possible errors, which may occur in the information transfer between the two steps in the first method.

6.4. Evaluation

The last step in the experiments involves evaluating the translations and backtranslations to detect patterns using different statistical methods. These methods aim to probe the initial Hypothesis H, discussed in Section 4.3. We apply all methods to all subsets, as presented in Section 6.1, and extract diversity information regarding the subsets themselves, as well as compare the results of the subsets against each other.

6.4.1. Reoccurrence Evaluation

We evaluate how many of the source sentences and source words reoccur in the backtranslations:

1. Gather the backtranslations for each source sentences.

2a. Count the number of source sentences that reappear in their list of backtranslations.

2b. Count the number of source sentences which contain the source word in their backtranslations.

The purpose of this evaluation is to determine which of the subsets are able to reconstruct more of the source sentences/words.

We also evaluate how often the source sentences and source words reoccur in the backtranslations:

1. Gather the backtranslations for each source sentences.

2a. Count in how many of the backtranslations the source sentences reappear.

2b. Count in how many of the backtranslations the source word reappears.

We use both evaluations to verify whether sentences in the ambiguous subset would be restored more often than the disambiguated and the non-ambiguous subsets, with which we probe the Hypothesis d).

6.4.2. Uniqueness Evaluation

We evaluate the number of unique words and sentences in the translations and backtranslations for each sentence of the subsets.

For the sake of the evaluation, we follow this routine (*[translations | backtranslations]* denotes that we follow the same step for both the translations and backtranslations):

1. Collect the *[translations | backtranslations]* for each source sentence.

2a. Count how many of the *[translations | backtranslations]* are unique.

2b. Count how many unique words there are in the *[translations | backtranslations]* and normalize the number by the total amount of words.

3. Average the result for all sentences.

We use this method to probe the Hypotheses a) and b).

6.4.3. Gender Evaluation

We perform the evaluation of gender on the translations of the subsets. First, we align each word in the source sentence with its corresponding word in the translations and backtranslations (see Subsection 6.3 for more detail). Then, we perform the following steps:

1. Gather the translations of the source word for each source sentence.

2. Detect the gender of the translations for each source sentence.

3a. Determine the proportion of source sentences producing *male* versus *female*.

3b. Calculate how many of the source sentences produce *both genders*.

The purpose of this method is to assess if the translations produce the right gender (in the non-ambiguous subsets) or both genders (in the ambiguous subset) and how often they produce both genders for each subset.

6.4.4. Alignment Evaluation

Another form of evaluation is based on the alignment of the words between the source sentence, the translations and backtranslations (see Subsection 6.3 for more detail).

In order to assess the translations and backtranslations of the **source word**, we do:

1. Collect all *[translations | backtranslations]* of the source word.

2. Count how many of the *[translations | backtranslations]* are unique.

3. Average the result for all sentences.

Similarly, to assess the translations and backtranslations of the **rest of the sentence** excluding the source word, we do:

1. Collect all *[translations | backtranslations]* of the sentence rest.

2. Count how many of the *[translations | backtranslations]* are unique.

3. Average the result for all sentences.

The idea behind this evaluation method is to assess the Hypothesis c).

Next, we will present the results of these statistical evaluations of the subsets.

6.5. Results

We evaluate the translations and backtranslations of the different subsets based on the presented evaluation methods and present the results in the following subsections.

6.5.1. Reoccurrence Evaluation Results

The results from the evaluation of reoccurrence are listed in Table 6.2. As we can observe, the average from the subsets of common words dominates the highest score in both the recurring sentences and words. This is to be expected, because the words in these subsets are most generic and have the highest probability of being predicted, compared to the occupational words from the WinoMT sentences in the other three subsets.

Most interestingly, the female-disambiaguated subset has the lowest score for reoccurring sentences. When investigating the results, we found some discrepancy between the way the words "female" and "male" are translated. The *female* prefix is very often lost in the backtranslation, which results in the backtranslated sentence being regarded as differing from the source sentence. In contrast, the *male* prefix is most often preserved, resulting in the same sentence in backtranslation. We illustrate this with the following examples:

- **Source (EN):** The *female* developer argued with John.
  Translation (DE): Die Entwicklerin argumentierte mit John.
  Backtranslation (EN): The developer argued with John.

- **Source (EN):** The *male* developer argued with John.
  Translation (DE): Der *männliche* Entwickler argumentierte mit John.
  Backtranslation (EN): The *male* developer argued with John.

As we can see, the *male* prefix is translated to its corresponding word in German "männlich", while the *female* prefix is lost in the translation, but its meaning is reflected in the female gender of the German word for developer "Entwicklerin". This is the expected outcome of the disambiguation method using gender forcing. According to these results, we can conclude that disambiguation with "female" is more successful in achieving this.

Furthermore, both disambiguation subsets sometimes generate the opposite gender with the correct prefix, for example, "der *weibliche* Entwickler" (the *female* male developer) and "die *männliche* Entwicklerin" (the *male* female developer). This presents a contradiction that influences the translation quality in the next step as well. We tried mitigating this by removing the German words for "female" and "male" ("weiblich" and "männlich") from the translations with Beam search with beam size 10 (see Table 6.2a). As we can see, this has a significant effect on the reoccurrence in backtranslation for the disambiguated subsets. For the male-disambiguated subset, only 6 out of 335 sentences reoccur, while for the female-disambiguated subset, 25 out of 335 sentences reoccur. This means that when we remove the gender prefix words in the German translations, they are a lot less likely to be backtranslated to their English counterparts in backtranslation.

We note that the findings from these results are important and will have an effect on the further experiments.

	Ambiguous	Disambiguated (male)	Disambiguated (female)	Non-ambiguous average
#Sentences	295/335	293/335	118/335	308/335
#Sentences (R)	295/335	6/335	25/335	308/335
Sentences	6.5/100	4.72/100	1.13/100	6.69/100
#Words	329/335	330/335	314/335	335/335
#Words (R)	329/335	330/335	314/335	335/335
Words	56.65/100	58.8/100	50.47/100	71.26/100

(a) **Beam search with beam size 10**. Backtranslation. Nbest size 10. Highest scores are underlined.
R: removed German words for *male* and *female*.
First and second row: number of source sentences that reappear in the backtranslations.
Third row: averaged number of times the source sentences reoccur in 100 backtranslations.
Fourth and fifth row: number of source sentences which contain the source word in the backtranslations.
Sixth row: averaged number of times the source words reoccur in 100 backtranslations.

	Ambiguous	Disambiguated (male)	Disambiguated (female)	Non-ambiguous average
#Sentences	329/335	330/335	281/335	329/335
Sentences	72.51/10000	48.59/10000	21.85/10000	70.94/10000
#Words	335/335	335/335	335/335	335/335
Words	4714.74/10000	4788.99/10000	4192.54/10000	5925.98/10000

(b) **Beam search with beam size 100**. Backtranslation. Nbest size 100. Highest scores are underlined.
First row: number of source sentences that reappear in the backtranslations.
Second row: averaged number of times the source sentences reoccur in 10000 backtranslations.
Third row: number of source sentences which contain the source word in the backtranslations.
Fourth row: averaged number of times the source words reoccur in 10000 backtranslations.

	Ambiguous	Disambiguated (male)	Disambiguated (female)	Non-ambiguous average
#Sentences	270/335	249/335	87/335	295/335
Sentences	3.99/100	2.52/100	0.58/100	4.54/100
#Words	332/335	333/335	326/335	335/335
Words	50.83/100	49.02/100	43.86/100	66.83/100

(c) **Sampling**. Backtranslation. Nbest size 10. Highest scores are underlined.
First row: number of source sentences that reappear in the backtranslations.
Second row: averaged number of times the source sentences reoccur in 100 backtranslations.
Third row: number of source sentences which contain the source word in the backtranslations.
Fourth row: averaged number of times the source words reoccur in 100 backtranslations.

Table 6.2.: Reoccurrence Evaluation Results

Table 6.2a shows the results from the evaluation of reoccurrence for beam search with beam size 10. Here, contrary to the expectation derived from Hyp. d), the ambiguous word reoccurs more often for the male-disambiguated subset than for the ambiguous subset. To inspect this further, we increase the beam size to 100 and change the decoding strategy to Sampling.

Table 6.2b shows the results from the evaluation of reoccurrence for beam search with beam size 100. Here we can observe that the ambiguous word still reoccurs more often for the male-disambiguated subset than for the ambiguous subset, in contradiction with Hyp. d). Also, more female sentences are reoccurring in backtranslation compared to beam 100, which means that increasing the beam increases the possibility for preservation of the *female* prefix in the backtranslations. This indicates more variability in the translation, but also a worsening of the translation quality with increasing the beam size.

The results from Sampling, as seen in Table 6.2c, show that the ambiguous word reoccurs more often in backtranslations for the ambiguous subset than for the disambiguated subsets (male and female), which was not the case with Beam search. This is a partially positive result for Hyp. d), because the ambiguous word is expected to reoccur more often in backtranslation than the corresponding non-ambiguous word. However, the non-ambiguous word in the non-ambiguous subset still reoccurs most often.

6.5.2. Uniqueness Evaluation Results

The results from the evaluation of uniqueness are listed in Table 6.3 for translation and Table 6.4 for backtranslation. Since we apply the Beam search algorithm for decoding with beam size 10, using fairseq's built-in option for generating the nbest list of 10, we expect it to generate 10 unique sentences per translation. This is almost always the case, as we see from the score for the number of unique sentences in translations (e.g., 9.94 unique sentences for the ambiguous subset). We suspect this may happen due to subword tokenization, in which two different subwords, coming from separate beams in the Beam search algorithm, may create the same word, which in the end would lead to the same sentence. However, the error is no more than 1% and can be disregarded. With Sampling, we sample a higher number of translations (e.g., 50), and extract 10 unique translations from the generated samples after the fact. This means that we always have complete 10 unique translations per sentence on average.

Most notable are the results for the number of unique backtranslations. As we can see in the first row of Table 6.4a, the ambiguous subset produces the least amount of unique sentences in backtranslation, which proves Hyp. a). To inspect this further, we increase the beam size to 100, and as we see from Table 6.4b this experiment confirms the results.

However, when removing the German prefix gender words *male* and *female*, we get mixed results in relation to the disambiguated subsets. The value for the male-disambiguated subset is the lowest, however the value for the ambiguous subset is still lower than the female-disambiguated subset. Then, we can speculate that modifying the translation in the suggested manner would not be further useful to the inspection of the approach.

To investigate this further, we regard the results for the different common words separately instead of averaged, as we can see in Fig. 6.3. The minimum value in the range, 43.9 (for the word "boy") for beam size 10 and 3074.01 (for the word "guy") for beam size 100, is still lower for the common words compared to the ambiguous subset, 45.98 for beam size 10 and 3181.51 for beam size 100.

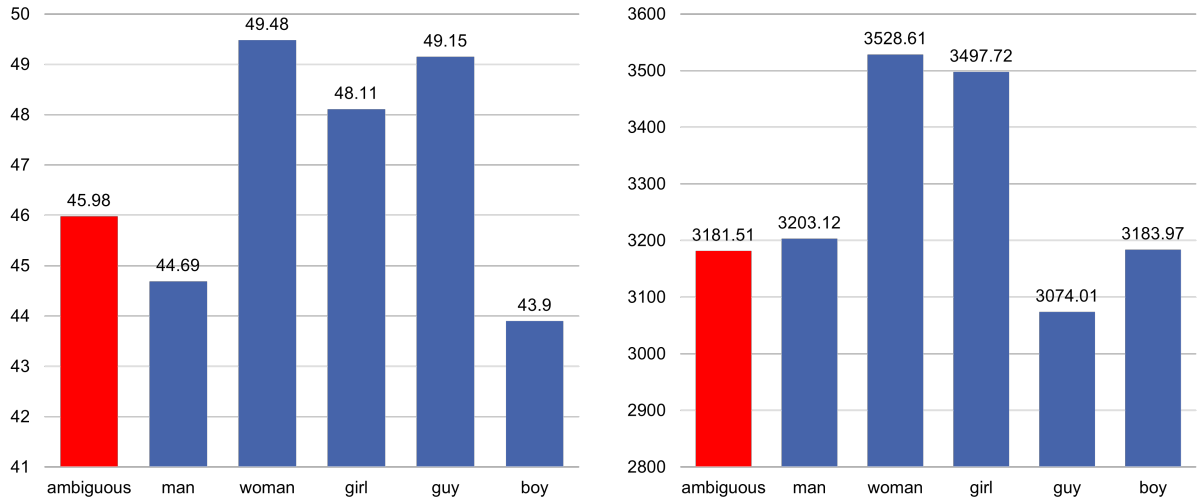
	Ambiguous	Disambiguated (male)	Disambiguated (female)	Non-ambiguous average
Sentences	9.94/10	9.95/10	9.87/10	9.97/10
Words	0.205	0.19	0.201	0.202

(a) **Beam search with beam size 10**. Nbest size 10. Highest scores are underlined.
First row: Averaged number of unique sentences per source sentence out of 10 translations.
Second row: Averaged number of unique words per source sentence, normalized by the average total number of words in 10 translations.

	Ambiguous	Disambiguated (male)	Disambiguated (female)	Non-ambiguous average
Sentences	10/10	10/10	10/10	10/10
Words	0.278	0.277	0.29	0.263

(b) **Sampling**. Nbest size 10. Highest scores are underlined.
First row: Averaged number of unique sentences per source sentence out of 10 translations.
Second row: Averaged number of unique words per source sentence, normalized by the average total number of words in 10 translations.

Table 6.3.: Uniqueness Evaluation Results for Translation



(a) Beam search with beam size 10

(b) Beam search with beam size 100

Figure 6.3.: Comparison Between the Number of Unique Backtranslations for Common Words and Ambiguous Subsets

The results for the number of unique words in translation and backtranslation are inconclusive. Considering Hyp. b), we expected the ambiguous subset to generate the least unique words in backtranslation, but this is not the case.

We also investigated the generated backtranslations from the Sampling method, as seen in 6.4c. Interestingly, Sampling generates almost twice as many unique backtranslations,

	Ambiguous	Disambiguated (male)	Disambiguated (female)	Non-ambiguous average
Sentences	45.98/100	50.73/100	50.82/100	47.06/100
Sentences (R)	45.98/100	41.36/100	47.03/100	47.06/100
Words	0.044	0.039	0.043	0.045
Words (R)	0.044	0.042	0.043	0.045

(a) **Beam search with beam size 10**. Nbest size 10. Highest scores are underlined.
R: removed German words for *male* and *female*.
First and second row: Averaged number of unique sentences per source sentence out of 10 translations.
Third and fourth row: Averaged number of unique words per source sentence, normalized by the average total number of words in 100 backtranslations.

	Ambiguous	Disambiguated (male)	Disambiguated (female)	Non-ambiguous average
Sentences	3181.51/10000	3391.55/10000	3424.98/10000	3297.48/10000

(b) **Beam search with beam size 100**. Nbest size 100. Highest scores are underlined.
Averaged number of unique sentences per source sentence out of 100 translations.

	Ambiguous	Disambiguated (male)	Disambiguated (female)	Non-ambiguous average
Sentences	81.81/100	86.87/100	85.75/100	80.6/100
Words	0.104	0.103	0.108	0.098

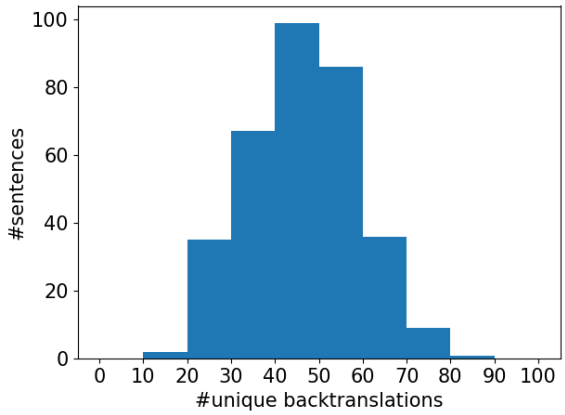
(c) **Sampling**. Nbest size 10. Highest scores are underlined.
First row: Averaged number of unique sentences per source sentence out of 10 translations.
Second row: Averaged number of unique words per source sentence, normalized by the average total number of words in 100 backtranslations.

Table 6.4.: Uniqueness Evaluation Results for Backtranslation

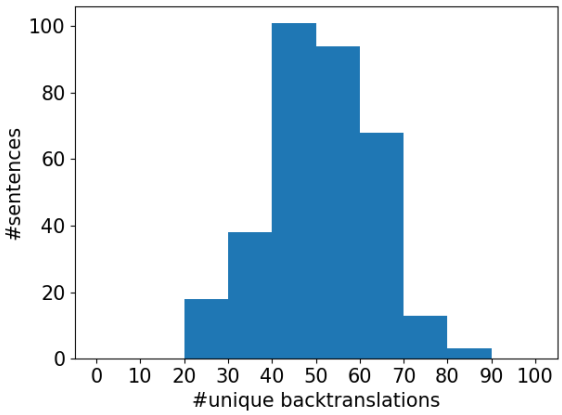
regarding both words and sentences. Here, the lowest score belongs to the non-ambiguous subset, contrary to the expectation from Hyp. a). However, the ambiguous subset still has less unique backtranslations than the corresponding disambiguated subsets, which is what we expected.

Furthermore, we plot the distribution of unique backtranslations, which can be seen in Fig. 6.4 for Beam search with beam size 10, containing the histograms for each subset. Most sentences have between 40 and 50 unique backtranslations, except for the female-disambiguated subset, which has between 50 and 60 unique backtranslations, confirming the highest score from Table 6.4. Another observation we can make pertains to the difference between the ambiguous subset 6.4a and the non-ambiguous subset (6.4c). There, we can see that the range is smaller ([20, 80]) and more of the non-ambiguous sentences have the average amount of unique backtranslations (around 140 sentences), while the ambiguous sentences produce more variable amount of backtranslations in terms of range ([10, 90]) but less amount of sentences producing

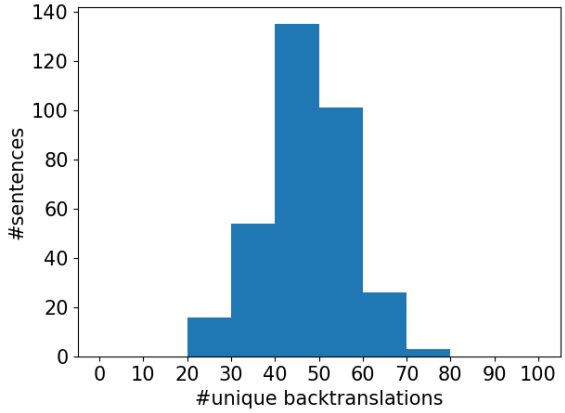
the average amount of unique backtranslations (around 100 sentences). The distribution of unique backtranslations for Beam search with beam size 100 and Sampling show the same trend and can be seen in Appendix A.



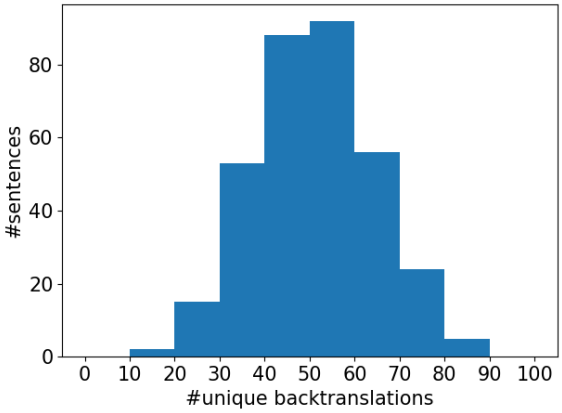
(a) Ambiguous Subset



(b) Disambiguated Subset (male)



(c) Non-ambiguous Subset Average



(d) Disambiguated Subset (female)

Figure 6.4.: Distribution of Unique Backtranslations: Beam search with beam size 10

As next, we want to inspect the translations for the source word separately.

6.5.3. Gender Evaluation Results

The results from the evaluation of gender are listed in Table 6.5. We can observe that, as expected, the male-disambiguated sentences have predominantly male translations, and similarly the female-disambiguated sentences have mostly female translations. The same applies to the male words “man”, “guy” and “boy”, as well as the female word “woman”. The female word “girl” presents an exception, because in German it is a neutral noun.

Interestingly, when comparing the disambiguation subsets, the disambiguation with “female” seems to be more successful overall, with more sentences producing the right gender and less of both genders appearing in the translations.

Also, as expected, the ambiguous source sentences produce the most translations of both genders, while the common non-ambiguous words produce the least. Despite this, the disambiguation subsets still have a rather high amount of sentences producing both genders. With Sampling, a higher amount of sentences produces both genders of the source word in translations than with Beam search, which points to Sampling producing more variable output in translations. The conclusion we can make from these results is that the gender prefix words *male* and *female* do not always fulfill the expected role of enforcing the correct gender.

Furthermore, we can make the observation that a significantly higher percentage of generated translation are male (86.27% for Beam search, 80.54% for Sampling) versus female (12.81% for Beam search, 14.33% for Sampling). Sampling seems to produce more balanced translations in terms of gender, but there are still predominantly male translations of the source word. This can only point to the already biased pre-trained NMT model.

	Ambiguous	Disambiguated (male)	Disambiguated (female)	Non-ambiguous
Male	86.27%	89.46%	6.81%	*man*: 95.01% *woman*: 0.51% *girl*: 0.39% *guy*: 93.07% *boy*: 96.15%
Female	12.81%	11.19%	92.33%	*man*: 0.18% *woman*: 96.69% *girl*: 0.81% *guy*: 0.18% *boy*: 0.27%
Both genders	38.21%	35.22%	28.06%	average: 0.72%

(a) **Beam search with beam size 10**. Translation. Nbest size 10. Highest scores are underlined. First and second row: Percentage of the source sentences producing male versus female translations. Third row: Percentage of the source sentences producing both genders in translation.

	Ambiguous	Disambiguated (male)	Disambiguated (female)	Non-ambiguous
Male	78.66%	81.96%	8.27%	*man*: 83.80% *woman*: 0.59% *girl*: 0.75% *guy*: 78.59% *boy*: 86.48%
Female	14.88%	14.24%	86.90%	*man*: 0.46% *woman*: 88.57% *girl*: 5.66% *guy*: 0.98% *boy*: 0.75%
Both genders	91.64%	91.34%	82.98%	average: 23.04%

(b) **Beam search with beam size 100**. Translation. Nbest size 100. Highest scores are underlined. First and second row: Percentage of the source sentences producing male versus female translations. Third row: Percentage of the source sentences producing both genders in translation.

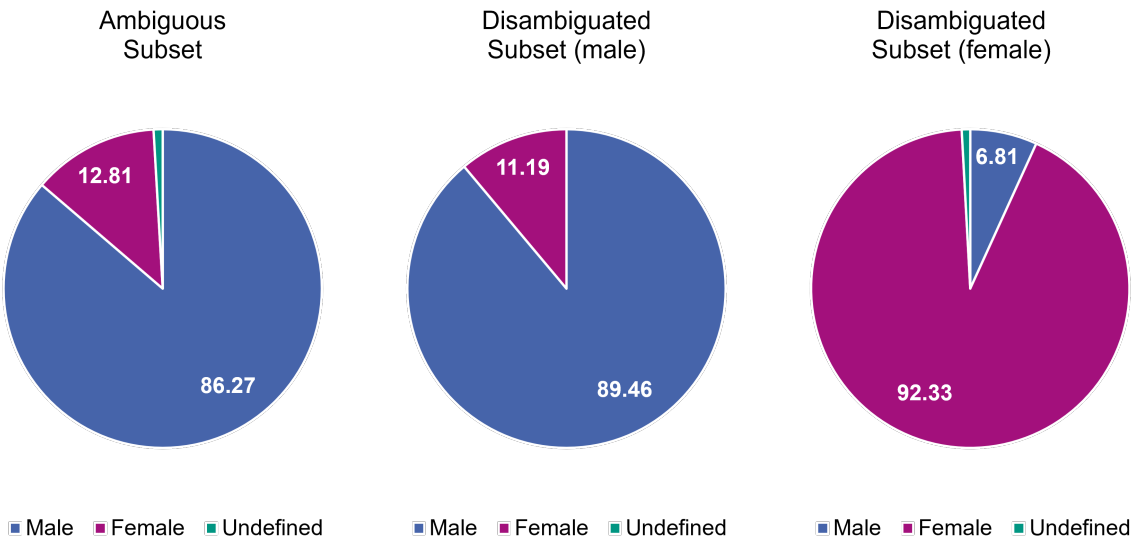
Table 6.5.: Gender Evaluation Results

	Ambiguous	**Disambiguated (male)**	**Disambiguated (female)**	**Non-ambiguous**
Male	80.54%	82.60%	7.64%	*man*: 89.91% *woman*: 0.45% *girl*: 0.63% *guy*: 85.04% *boy*: 91.31%
Female	14.33%	13.76%	87.64%	*man*: 0.42% *woman*: 92.96% *girl*: 2.72% *guy*: 0.6% *boy*: 0.54%
Both genders	48.06%	55.52%	34.92%	average: 2.98%

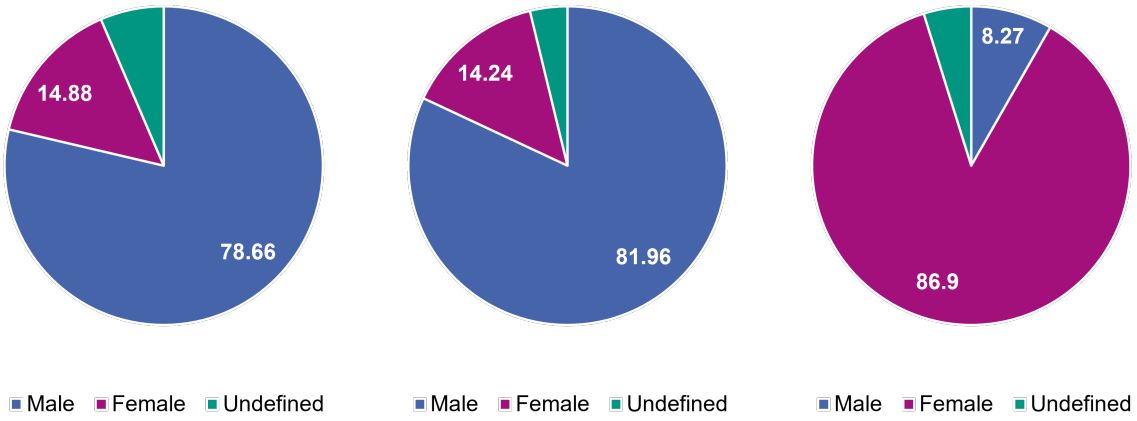
(c) **Sampling**. Translation. Nbest size 10. Highest scores are underlined.
First and second row: Percentage of the source sentences producing male versus female translations.
Third row: Percentage of the source sentences producing both genders in translation.

Table 6.5.: Gender Evaluation Results

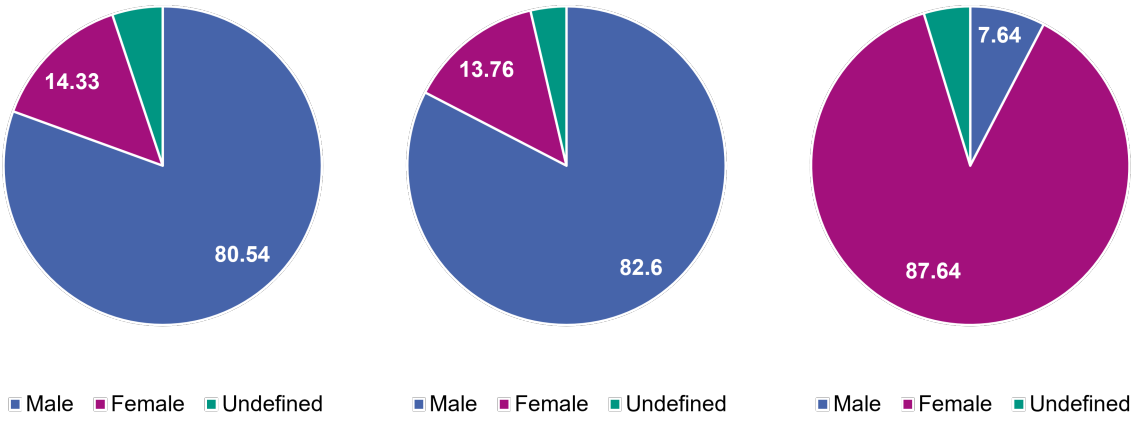
Fig. 6.5 further shows the influence of increasing the beam size tenfold and the difference between Beam search and Sampling. We can observe that more of both genders occur in translation of the ambiguous subset with beam size 100 compared to beam size 10. But this is also the case for the disambiguated subsets, where we do not expect this. Also, there is more balance between female and male words in the translations of the ambiguous subset, with the difference between 78.66% and 14.88% being smaller than the difference between 86.27% and 12.81%. But on the other hand, more male gender translations occur in the female-disambiguated subset, which is a downside. Sampling, on the other hand, achieves more balance between male and female translations for the ambiguous subset without increasing the occurrence of both genders in the disambiguated subsets, unlike Beam search with beam size 100.



(a) Beam search with beam size 10



(b) Beam search with beam size 100



(c) Sampling

Figure 6.5.: Gender Representation in Translation

6.5.4. Alignment Evaluation Results

The results from the evaluation of alignment are listed in Table 6.7 for translation and Table 6.8 for backtranslation. An example of aligning the source words with their target translations and extracting the translations and backtranslations for each word from the nbest list can be seen in Table 6.6.

Sentence	**Translations**	**Backtranslations**
The	{Der}	{The}
developer	{Entwickler, Bauunternehmer, Bauträger}	{property, building, contractor, estate, Developer, real, builder, developer, developers}
argued	{streitete, sich, stritt, argumentierte}	{fought, disagreed, out, argued, argument, argues, clashed, quarreled, disputed, was, reasoned, quarrelled, arguing, had}
with	{mit}	{with}
John	{'Johannes', 'John'}	{John, Johannes, him}
.	{.}	{.}

Table 6.6.: Nbest translations and backtranslations for each word in the source sentence "The developer argued with John.". Marked ambiguous word.

The alignment results for translation feature unique words with and without gender information. For example, the male and female German word for developer: "Entwickler" and "Entwicklerin", are considered one unique word when disregarding gender information. The removal of gender information is performed using a rule-based approach. The assumption is that this would reduce the number of unique words for the ambiguous subset compared to the other subsets. However, the effect is not significant enough to influence the results. We can see that the non-ambiguous subset has the least amount of unique translations to the source word compared to the other subsets for all algorithms. The only difference is between the ambiguous subset and the disambiguated subsets. With Sampling and Beam search with beam size 100, the ambiguous subset produces the least unique translations, while with Beam search with beam size 10, the male-disambiguated subset has the least translations.

While the results from the translation are not conclusive, the results from the backtranslation exhibit a noticeable pattern. We can see that for the source word, the non-ambiguous subset has the least amount of unique backtranslations, contrary to the expectation from Hyp. c), which postulated that the ambiguous subset should produce the least unique backtranslations. However, the ambiguous subset has less unique translations than the disambiguated subsets when applying Sampling and Beam search with beam size 100, which partially confirms the hypothesis. For Beam search with beam size 10, however, the male-disambiguated subset produces the least amount of backtranslations.

	Ambiguous	Disambiguated (male)	Disambiguated (female)	Non-ambiguous average
Source word (FA, WIG)	2.87/10	2.83/10	<u>3.02/10</u>	1.83/10
Source word (FA, WOG)	2.66/10	2.64/10	<u>2.81/10</u>	1.83/10
Source word (AA, WIG)	2.60/10	<u>2.63/10</u>	2.55/10	1.70/10
Source word (AA, WOG)	2.38/10	<u>2.43/10</u>	2.33/10	1.70/10
Sentence rest (AA)	1.87/10	1.67/10	1.64/10	<u>1.94/10</u>

(a) **Beam search with beam size 10**. Nbest size 10. Highest scores are underlined.
FA: *fast_align*, **AA**: *awesome-align*. **WIG** (with gender): gender information preserved, **WOG** (without gender): gender information removed.
First to fourth row: Averaged number of unique translations of the source word per source sentence in the 10 translations.
Fifth row: Averaged number of unique translations of the sentence rest per source sentence in the 10 translations.

	Ambiguous	Disambiguated (male)	Disambiguated (female)	Non-ambiguous average
Source word (FA, WIG)	12.37/100	13.70/100	<u>14.76/100</u>	7.18/100
Source word (FA, WOG)	11.04/100	12.47/100	<u>13.38/100</u>	7.18/100
Source word (AA, WIG)	10.81/100	12.12/100	<u>13.12/100</u>	5.26/100
Source word (AA, WOG)	9.46/100	10.88/100	<u>11.75/100</u>	5.26/100
Sentence rest (AA)	6.52/100	5.39/100	5.85/100	<u>7.23/100</u>

(b) **Beam search with beam size 100**. Nbest size 100. Highest scores are underlined.
FA: *fast_align*, **AA**: *awesome-align*. **WIG** (with gender): gender information preserved, **WOG** (without gender): gender information removed.
First to fourth row: Averaged number of unique translations of the source word per source sentence in the 100 translations.
Fifth row: Averaged number of unique translations of the sentence rest per source sentence in the 100 translations.

Table 6.7.: Alignment Evaluation Results for Translation

	Ambiguous	Disambiguated (male)	Disambiguated (female)	Non-ambiguous average
Source word (FA, WIG)	3.93/10	4.31/10	4.72/10	2.33/10
Source word (FA, WOG)	3.71/10	4.09/10	4.48/10	2.33/10
Source word (AA, WIG)	3.44/10	4.01/10	4.06/10	1.95/10
Source word (AA, WOG)	3.21/10	3.79/10	3.80/10	1.95/10
Sentence rest (AA)	2.33/10	2.23/10	2.18/10	2.37/10

(c) **Sampling**. Nbest size 10. Highest scores are underlined.
FA: *fast_align*, **AA**: *awesome-align*. **WIG** (with gender): gender information preserved, **WOG** (without gender): gender information removed.
First to fourth row: Averaged number of unique translations of the source word per source sentence in the 10 translations.
Fifth row: Averaged number of unique translations of the sentence rest per source sentence in the 10 translations.

Table 6.7.: Alignment Evaluation Results for Translation

	Ambiguous	Disambiguated (male)	Disambiguated (female)	Non-ambiguous average
Source word (FA)	8.84/100	8.32/100	9.79/100	5.25/100
Source word (AA)	7.48/100	7.04/100	7.85/100	4.80/100
Source word (Tercom)	8.02/100	7.90/100	9.82/100	6.52/100
Sentence rest (AA)	4.13/100	3.72/100	3.73/100	4.66/100

(a) **Beam search with beam size 10**. Nbest size 10. Highest scores are underlined.
FA: *fast_align*, **AA**: *awesome-align*.
First-third row: Averaged number of unique backtranslations of the source word per source sentence in the 100 backtranslations.
Fourth row: Averaged number of unique backtranslations of the sentence rest per source sentence in the 100 backtranslations.

	Ambiguous	Disambiguated (male)	Disambiguated (female)	Non-ambiguous average
Source word (FA)	147.93/10000	148.34/10000	163.53/10000	69.76/10000
Source word (AA)	120.08/10000	121.96/10000	136.14/10000	48.82/10000
Sentence rest (AA)	68.00/10000	66.39/10000	63.41/10000	66.06/10000

(b) **Beam search with beam size 100**. Nbest size 100. Highest scores are underlined.
FA: *fast_align*, **AA**: *awesome-align*.
First-third row: Averaged number of unique backtranslations of the source word per source sentence in the 10000 backtranslations.
Fourth row: Averaged number of unique backtranslations of the sentence rest per source sentence in the 10000 backtranslations.

Table 6.8.: Alignment Evaluation Results for Backtranslation

	Ambiguous	Disambiguated (male)	Disambiguated (female)	Non-ambiguous average
Source word (FA)	19.09/100	20.60/100	23.86/100	10.48/100
Source word (AA)	15.28/100	17.35/100	19.77/100	8.35/100
Source word (Tercom)	16.63/100	20.03/100	24.58/100	12.72/100
Sentence rest (AA)	8.28/100	8.12/100	8.10/100	8.41/100

(c) **Sampling**. Nbest size 10. Highest scores are underlined.
FA: *fast_align*, **AA**: *awesome-align*.
First to third row: Averaged number of unique backtranslations of the source word per source sentence in the 100 backtranslations.
Fourth row: Averaged number of unique backtranslations of the sentence rest per source sentence in the 100 backtranslations.

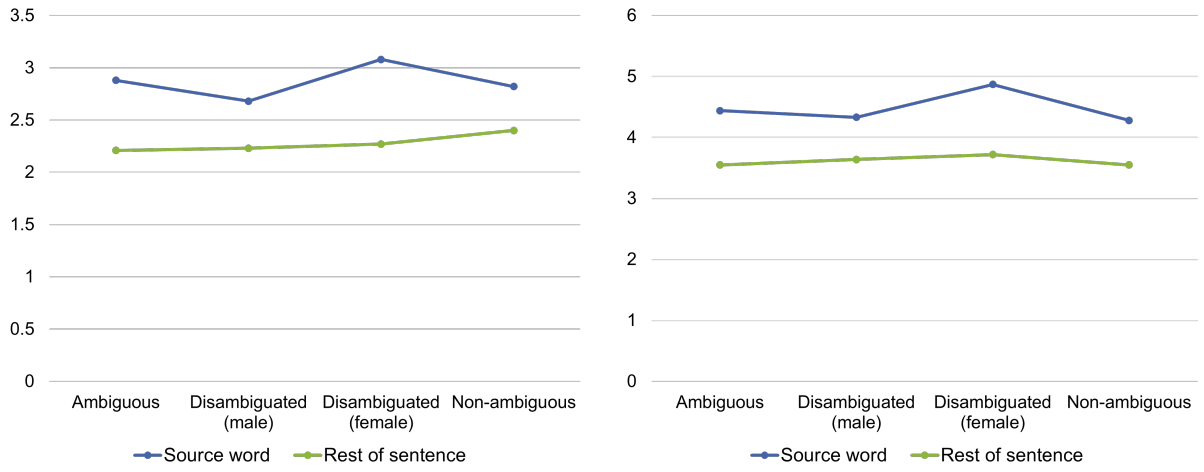
Table 6.8.: Alignment Evaluation Results for Backtranslation

We also investigate the relationship between translation and backtranslation for the source word and the rest of the sentence. Fig. 6.6 shows that there is not a significant discrepancy between the amount of translations compared to the amount of backtranslations for the different subsets. This may indicate a correlation between the number of translations and backtranslations.

A notable trend in the results is that the non-ambiguous subset has the significantly lowest score in uniqueness for the translations of the source word, but it also has the highest score for the rest of the sentence. Also, the difference in the scores for the source word and the rest of the sentence ($2.38 - 1.87$) is smaller for the non-ambiguous subset than the difference for the ambiguous subset ($1.70 - 1.94$). A similar tendency is observed also in backtranslation, as well as for Beam search with beam size 100 and Sampling.

To inspect this further, we evaluate the relationship between the source word and the rest of the sentence, as seen in Fig. 6.7. We observe that for the non-ambiguous subset the source word and the rest of the sentence have a similar amount of translations and backtranslations, while the ambiguous and the disambiguated subsets generate almost twice as many translations and backtranslation for the source word compared to the rest of the sentence. This is an indication of the stronger tendency of the decoding algorithm to put more emphasis on the ambiguous word rather than the rest of the sentence, when such an ambiguous word is present.

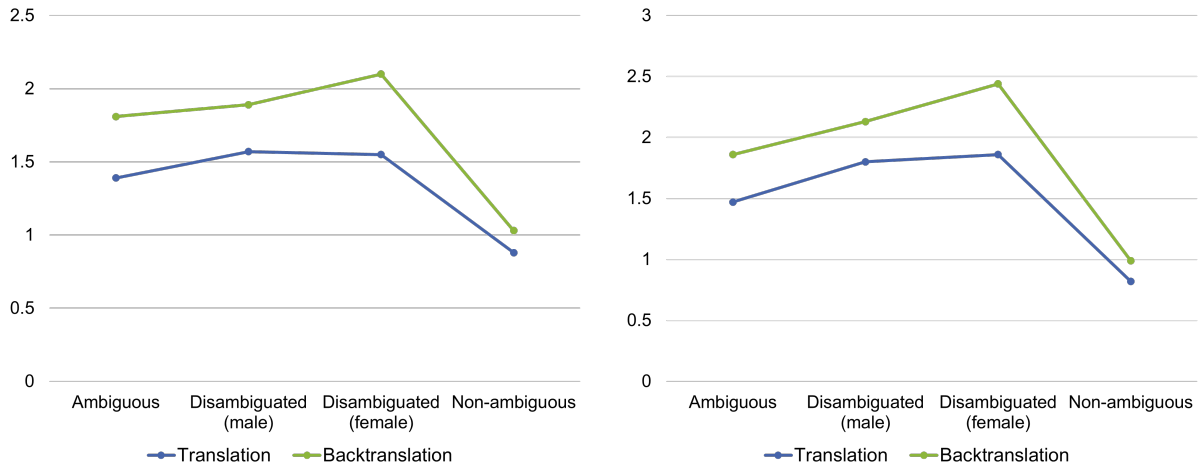
Interestingly, for Beam search with beam size 10 as well as for Sampling, both in translation and backtranslation the rest of the sentence for the non-ambiguous subset produces the most diverse translations compared to the ambiguous subset. This may indicate that more emphasis is given on diversity of the rest of the sentence, when the source word is unambiguous itself.



(a) Beam search with beam size 10

(b) Sampling

Figure 6.6.: **Relationship Between Translation and Backtranslation.** The results represent the amount of backtranslations divided by the amount of translations.



(a) Beam search with beam size 10

(b) Sampling

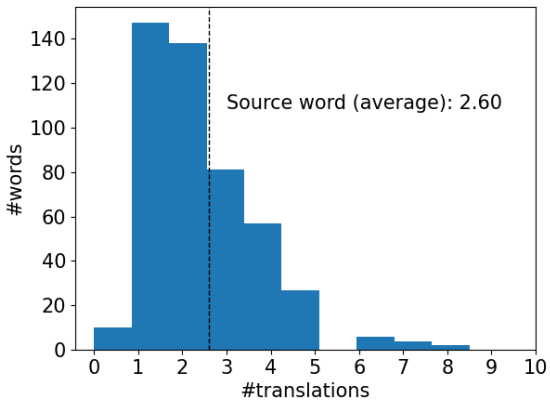
Figure 6.7.: **Relationship Between Source Word and Rest of Sentence.** The results represent the amount of translations or backtranslations of the source word divided by the amount of translations or backtranslations of the rest of the sentence.

Furthermore, we plot the distribution of unique translations and backtranslations for every word in the subsets, which can be seen in Fig. 6.8 and 6.9, containing the histograms of each subset for Beam search with beam size 10. We can observe in the plots in Fig. 6.8 that for the ambiguous subset as well as for the disambiguated subsets most words have between one and two unique translations, while for the non-ambiguous subset they have closer to two unique backtranslations. However, less of the words in the non-ambiguous subset have the average amount of unique translations (around 120 sentences), compared to the ambiguous subset (around 140 words) and the disambiguated subsets (around 150 words). On the other hand, the

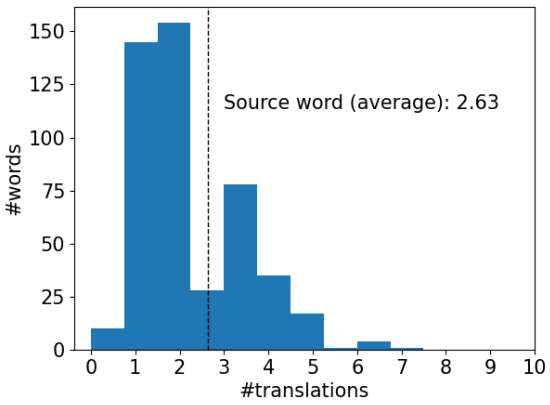
plots for backtranslations in Fig. 6.9 show the opposite tendency. Here, more of the words in the non-ambiguous subset have the average amount of unique translations (more than 150 sentences), compared to the ambiguous subset (less than 150 words) and the disambiguated subsets (less than 150 words). This observation could be positive for the hypothesis (Hyp. H), indicating that ambiguous sentences produce less diverse backtranslations than non-ambiguous sentences.

Another discovery we can make refers to the position of the average value for the source word in the histograms for both translations (Fig. 6.8) and backtranslations (Fig. 6.9). For the ambiguous subset, as well as for the disambiguated subsets, the value is higher than the maximum, while for the non-ambiguous subset it matches the maximum. This could be one indication for the non-ambiguity of the word in the non-ambiguous subset, since the number of unique translations matches the average number of translations for most words in the subset.

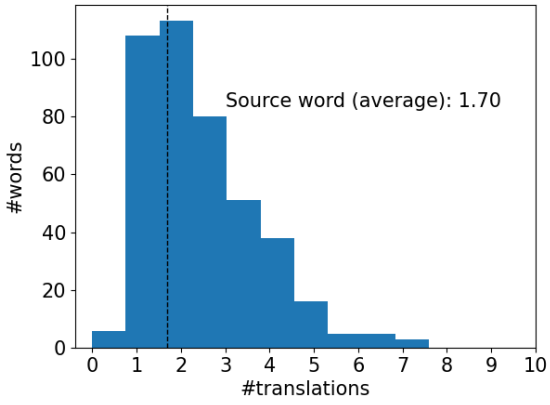
The results for Beam search with beam size 100 and Sampling show a similar trend in the distribution and can be seen in Appendix A.



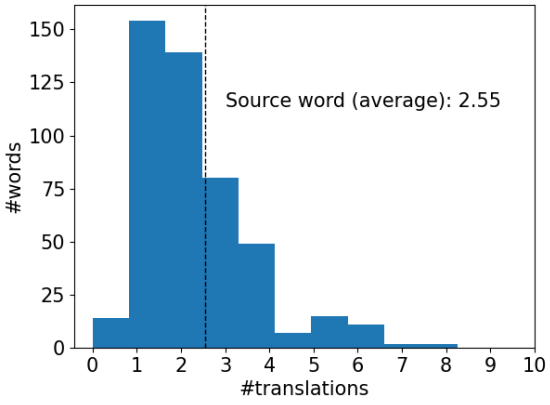
(a) Ambiguous Subset



(b) Disambiguated Subset (male)

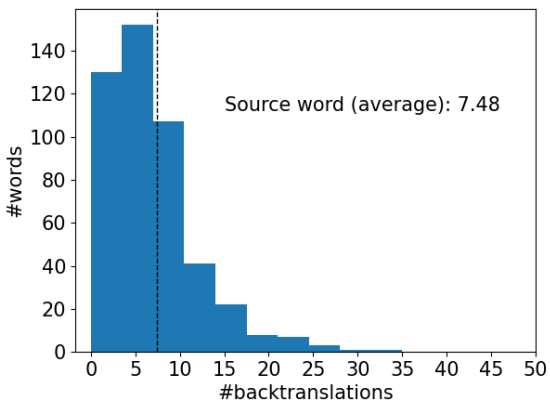


(c) Non-ambiguous Subset Average

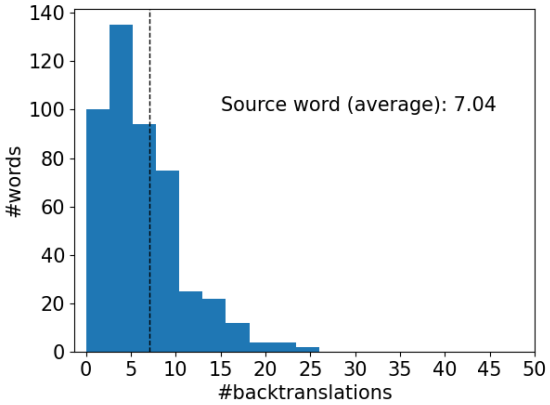


(d) Disambiguated Subset (female)

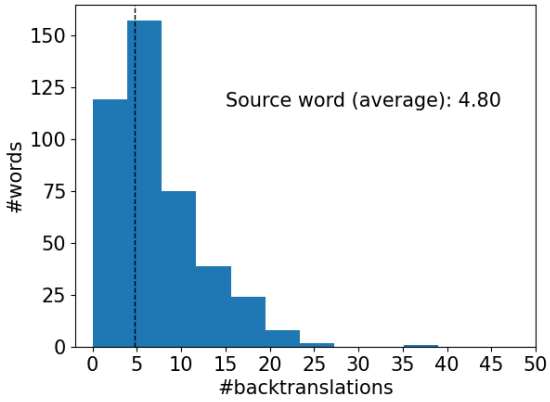
Figure 6.8.: **Distribution of Unique Translations for Words**. Beam search with beam size 10. Nbest size 10. Alignment with *awesome-align*. The dashed line marks the average number of unique translations for the source word, the value displayed to the right.



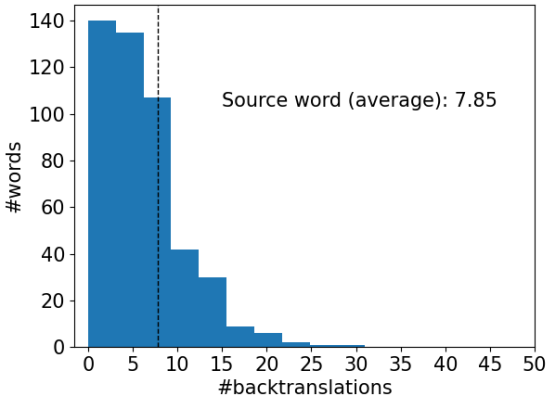
(a) Ambiguous Subset



(b) Disambiguated Subset (male)



(c) Non-ambiguous Subset Average



(d) Disambiguated Subset (female)

Figure 6.9.: **Distribution of Unique Backtranslations for Words**. Beam search with beam size 10. Nbest size 10. Alignment with *awesome-align*. The dashed line marks the average number of unique translations for the source word, the value displayed to the right.

7. Real-world Experiment

This chapter describes an experiment in a real-world setting. The purpose of this experiment is to test the hypothesis in natural conditions. In the following, we outline the steps for executing the experiment and the results from the evaluation.

7.1. Data Extraction

First, we extract the natural sentences from the MuST-SHE corpus, presented in Subsection 5.2.2. We choose 10 sentences for the experiment, that contain no context information regarding the gender of the ambiguous words. The sentence set is balanced, containing 5 sentences of each two genders - male and female. All sentences are listed in Table 7.1.

7.2. Data Preprocessing

As next, we preprocess the data by unmasking for each word in each original sentence. For unmasking we use the BERT base model, introduced in Section 5.4. The model generates five most probable words for each masked word. We replace the unmasked words in the sentences with three of the generated words. For each original sentence, we have three times the number of words in the sentence unmasked sentences.

Sets of Sentences We have the following multiple sets of sentences:

- Original set: the extracted sentences, as shown in Table 7.1.
- Unmasked word sets: $3 * |words|$ unmasked sentences for each sentence ($|words|$ denotes the number of words in the sentence).

Manual Replacement Since very often the ambiguous words in the original sentences are unmasked with ambiguous words as well, we try to mitigate this by manually replacing the gender ambiguous noun words with the following non-ambiguous words: *man, woman, girl, guy, boy*. We compare this approach with the originally replaced words by the BERT model.

Source Sentence	Ambiguous Word(s)	Gender
So now Thomson becomes the more likely **suspect**.	suspect	male
There was one black **professor** and one black assistant **dean**.	professor, dean	male
We have our cognitive biases, so that I can take a perfect history on a **patient** with chest pain.	patient	male
That's the **officer** who emailed me back, saying I think you can have a few classes with us.	officer	male
Steve, a physician, told me about a **doctor** that he worked with who was never very respectful, especially to junior staff and nurses.	doctor	male
What do you think a batting average for a **cardiac surgeon** or a **nurse practitioner** or an **orthopedic surgeon**, an **OBGYN**, a **paramedic** is supposed to be?	surgeon, nurse practitioner, OBGYN, paramedic	female
Fortunately for Mama Jane and her **friend**, a **donor** had provided treatment so that we could take them to the nearest hospital three hours away.	friend, donor	female
The three words are: Do you remember? "Do you remember that patient you sent home?" the other **nurse** asked matter-of-factly. "Well she's back," in just that tone of voice.	nurse	female
This one comes from a note that a **student** sent me after I gave a lecture about arousal nonconcordance.	student	female
At the end of a conference in a hotel lobby once, I'm literally on my way out the door and a **colleague** chases me down. "Emily, I just have a really quick question."	colleague	female

Table 7.1.: **Extracted Natural Sentences.** MuST-SHE 4th Category: No gender-disambiguating information can be retrieved. 10 sentences in total.

7.3. Translation

We translate the sets of sentences from English to German in the two steps:

1. **Translation Source -> Target:** Translate the sets in the target language (German).
2. **Backtranslation Target -> Source:** Translate the translations back into the source language (English).

For translation, we use the Beam search decoding strategy (see Subsection 2.1.3) with beam size 10. We choose this strategy based on the uniqueness evaluation results from the base experiment, where the number of unique sentences for the ambiguous subset is smaller than the number of unique sentences for the unambiguous subset, as expected.

7.4. Evaluation

For each original sentence, we execute the following algorithm:

1. Count the number of unique sentences in the backtranslations.
   $N \leftarrow |unique\ backtranslations|$
2. Count the number of unique sentences in the backtranslations for each unmasked word.
   $w_{ij} \leftarrow |unique\ backtranslations\ for\ sentence\ with\ masked\ word\ w_i\ and\ mask\ j|$
   $j \leftarrow \{1, 2, 3\}$
   $i \leftarrow \{1, 2, ..., |words|\}$
   $[w_{1j}, w_{2j}, ..., w_{|words|j}]$
3. Average the result for the three masks of each word.
   $w_i^{'} = \sum_{j=1}^{3} \frac{w_{ij}}{3}$
   $[w_1^{'}, w_2^{'}, ..., w_{|words|}^{'}]$
4. Subtract the number of unique backtranslations of the original sentence from the average.
   $[|w_1^{'} - N|, |w_2^{'} - N|, ..., |w_{|words|}^{'} - N|]$
3. Extract the words, which generate the 5 biggest differences.

Based on the Hypothesis a), the original sentence containing an ambiguous word is expected to generate considerably less unique sentences than the sentence with the corresponding unmasked ambiguous word, which is non-ambiguous. Therefore, this sentence is also expected to produce the biggest difference, when subtracting the number of unique sentences, pointing to the ambiguous word in the sentence. Since the natural sentences often exhibit more than one ambiguous word, we extract the three biggest differences, indicating the three most ambiguous words in the original sentence.

7.5. Results

The results from the evaluation can be seen in Table 7.2. We also compare the results of the replacement of the gender ambiguous nouns with BERT against the manual replacement. As we can see from the results, the manual replacement method achieves more detected ambiguous words in the list with the five most probable ambiguous words. This proves that the theory of replacing ambiguous words with non-ambiguous words may help in detecting ambiguous words in a sentence.

Source Sentence	Ambiguous Word(s)	BERT Replacement	Manual Replacement
So now Thomson becomes the more likely **suspect**.	suspect	[now, the, Thomson, **suspect**, more]	[now, the, Thomson, **suspect**, more]
There was one black **professor** and one black assistant **dean**.	professor, dean	[**professor**, There, and, . , was]	[**dean**, There, and, . , was]
We have our cognitive biases, so that I can take a perfect history on a **patient** with chest pain.	patient	[cognitive, chest, a, history, I]	[cognitive, chest, a, history, I]
That's the **officer** who emailed me back, saying I think you can have a few classes with us.	officer	[classes, saying, . , you, can]	[classes, saying, . , can, **officer**]
Steve, a physician, told me about a **doctor** that he worked with who was never very respectful, especially to junior staff and nurses.	doctor	[nurses, respectful, Steve, staff, especially]	[nurses, **doctor**, respectful, Steve, staff]
What do you think a batting average for a **cardiac surgeon** or a **nurse practitioner** or an **orthopedic surgeon**, an **OBGYN**, a **paramedic** is supposed to be?	surgeon, nurse practitioner, OBGYN, paramedic	[think, an, for, or, be]	[**paramedic**, think, an, for, or]
Fortunately for Mama Jane and her **friend**, a **donor** had provided treatment so that we could take them to the nearest hospital three hours away.	friend, donor	[the, we, three, , , a]	[the, we, three, , , a]
The three words are: Do you remember? "Do you remember that patient you sent home?" the other **nurse** asked matter-of-factly. "Well she's back," in just that tone of voice.	nurse	[you , you, are , words, **patient**]	[" , **nurse**, " , remember, of]

This one comes from a note that a **student** sent me after I gave a lecture about arousal nonconcordance.	student	[one, comes, nonconcordance, arousal, note]	[one, comes, nonconcordance, arousal, **student**]
At the end of a conference in a hotel lobby once, I'm literally on my way out the door and a **colleague** chases me down. "Emily, I just have a really quick question."	colleague	[chases, way, have, end, down]	[chases, way, have, end, down]

Table 7.2.: **Natural Experiment Results.** Displays the 5 most ambiguous words (sorted in descending order) for each sentence.
The marked words are the expected ambiguous words.
BERT Replacement: Unmasking of each word with the BERT model.
Manual Replacement: Unmasking of the gender ambiguous nouns with non-ambiguous nouns manually.

8. Discussion

This chapter presents a summary of results from conducting the experiments and discusses challenges and limitations of the approach.

8.1. Base Experiment

In our base experiment, we constructed four subsets to inspect the initial assumption, that sentences containing an ambiguity produce less diverse backtranslations than sentences without an ambiguity. We extracted fully ambiguous sentences from the synthetic dataset WinoMT. In one approach, we attempted disambiguating the source ambiguous words in the sentences with gender prefix words: *male* and *female*. Another approach consisted of replacing the source ambiguous words with five common words (*man, woman, girl, guy, boy*) and averaging the results. We translated and backtranslated the subsets and evaluated the results based on gender, uniqueness of translations and backtranslations, reoccurrence of the source sentence and source word and translations of the source word and words in the rest of the sentence.

8.1.1. Replacement Method

The evaluation of the results showed that disambiguating with *male* and *female* has different effects on the translation and backtranslation. The female-disambiguated subsets often lost the *female* prefix in backtranslation due to not directly translating it to the corresponding German word, but directly producing the female gender noun in translation. On the other hand, the male-disambiguated subset most often kept the *male* prefix in translation and backtranslation. From this, we can conclude that disambiguating with *female* proved to be more successful in terms of translation quality.

The replacement of an ambiguous word using disambiguation proved successful when using Beam search with beam size 100 and Sampling for generating translations. The ambiguous subset produced less unique backtranslations, proving Hyp. a) and c). Also, the ambiguous word in the ambiguous subset reoccurred most often in backtranslation, confirming Hyp. d).

Furthermore, when we only regard the ambiguous and the female-disambiguated subsets for Beam search with beam size 10, we observe that the ambiguous subset does produce less unique sentences in the backtranslations as well as less unique backtranslations of the source word, which partially confirms the initial Hypothesis H. Since we showed that disambiguating with *female* produces better quality translations than disambiguating with *male*, this result is positive.

On the other hand, replacing the ambiguous word with a common unambiguous word proved mostly ineffective for all search strategies. From the results on uniqueness of the backtranslations, we observed that the ambiguous subset generates the least unique sentences

in the backtranslations compared to the disambiguated subsets and the average of the non-ambiguous subset, both for beam size 10 and 100. However, when comparing the result of the ambiguous subset against the result for the common words in the non-ambiguous subset individually, we noticed that there is a lower value. This could be due to the chosen words of replacement. It may mean that the occupational ambiguous words have more semantic meanings outside the ambiguity of gender. From the analysis on alignment, we saw that the occupation words have multiple versions in both the source and the target language. For example, the word *man* most often gets translated to the single word "Mann", while the gender ambiguous word *developer* has been translated to more than three different words not considering gender information: "Bauträger", "Bauunternehmer", "Entwickler". This in turn influences the backtranslation, which also produces more unique backtranslations of the sentences. This can also be observed in the results from the alignment, where the average results for the non-ambiguous subset has the lowest value in unique translations of the source word, even when disregarding gender information for the ambiguous subset. We can conclude that in this case, the multiple meanings of the word have stronger influence on the diversity of translation than its gender ambiguity. Therefore, the replacement method using common words can be deemed unsuccessful in this scenario.

8.1.2. Search Method

We decided to compare Beam search with Sampling to see what effect it would have on the diversity of translation. Also, Roberts et al. (2020) prove that Beam search unlike Sampling has a tendency to generate more frequent masculine pronouns, because it guides the model towards an extreme operating point that exhibits zero variability and consequently amplifies the biases present in the training data, resulting in a skewed output. Indeed, the Sampling method introduced more variability regarding gender. Furthermore, it led to more unique words and sentences in backtranslation overall.

While for Beam search with beam size 10 the ambiguous word in the ambiguous subset has more unique backtranslations than the male-disambiguated subset, decoding with Beam search with beam size 100 or Sampling counteracts that and shows that the ambiguous word has the least amount of backtranslations compared to the disambiguated subset, confirming Hyp. c).

8.1.3. Correlation

Furthermore, we detected a possible correlation between the number of translations and backtranslations. The number of backtranslations seem to increase proportionally to the number of translations for all subsets.

In regard to the alignment results, we observed that the rest of the sentence produces the most unique translations and backtranslations for the non-ambiguous subset. But most importantly, the difference in the scores for the source word and the rest of the sentence is a lot bigger for the ambiguous subset than the non-ambiguous subset. This proves that the most diversity in translation is given to the ambiguous word in the sentence, when such is present. Therefore, we can conclude that the number of translations of the non-ambiguous words in a sentence is correlated with the number of translations of the ambiguous word in the sentence.

8.2. Real-world Experiment

The real-world experiment aimed to probe the assumption that sentences containing ambiguous words generate less unique backtranslations, and with this to attempt detecting ambiguous words in a realistic sentence. We developed an algorithm of replacing each word in the sentence with the most probable word, unmasked with a BERT model, while also manually replacing the gender ambiguous nouns with non-ambiguous words. We concluded from the results that the method of combining BERT with manual replacement of the ambiguous words appears more effective than the simple replacement method with BERT, proving that replacing ambiguous words with non-ambiguous words in a sentence may assist in detecting ambiguous words.

8.3. Challenges and Limitations

While conducting the experiments and evaluating the results, we also met a couple of challenges.

Gender Bias Pre-trained NMT models are gender biased, which influences the balance of male and female translations. For Beam search with beam size 10, we observed that in less than half of the translations, both genders occur, which influences the way of evaluating the assumption. Furthermore, we observed from the results on gender that there is an inclined tendency of translation to produce more male nouns than female. To combat these problems, we applied Beam search with beam size 100. This successfully contributed to both genders occurring in over 90% of the translations, however it only slightly improved the balance between male and female translations of the source word. It also carries specific drawbacks, such as worsening the translation quality overall. Sampling also slightly counteracts this issue, but it is still insufficient.

Furthermore, we noticed that disambiguating with *male* versus *female* yields different results. Unexpectedly, sometimes the gender words were completely disregarded, and the wrong gender noun was produced in translation. This suggests that the method of disambiguation using gender forcing may not be very effective.

Word Alignment Another challenge we encountered relates to the word alignment methods. Often, *fast-align* and *awesome-align* produced significantly different results, although the results did not differ in terms of the end conclusion towards the initial assumption. We eventually relied mostly on the results from *awesome-align*, because it is state-of-the-art. It would also be useful to evaluate the quality of the alignment methods, which we didn't specifically investigate due to time constraints.

Translation Quality The quality of translation presented itself to be another limitation, because the WinoMT dataset does not provide reference translations currently, therefore only manual assessment was possible. An automatic assessment would have been useful for the case, when we explored translation with the Beam search algorithm with beam size 100 to inspect the difference of quality within the nbest list.

Ambiguity Definition When detecting ambiguous words, we need to take into account the type of ambiguity we are testing for. Since we defined ambiguity as having one version in the source language and multiple versions in the target language, the occupational words are gender-ambiguous in this definition, but they are not generally ambiguous. However, they have multiple semantic meanings in both chosen source and target language. Therefore, replacing them with words, which have less semantic meanings in both languages (e.g., *man, woman, girl, guy, boy*), proved not useful for assessing their gender ambiguity.

Real-world Scenario In regard to detecting ambiguous words in a real-world setting, we also met a couple of challenges. The method of replacement using BERT is not always successful in replacing the ambiguous words with non-ambiguous words, which facilitated the need for manual replacement. This manual replacement would only be applicable in the small set of sentences we use for proof-of-concept, but not in a large dataset. Therefore, we outline the need for an automatic method of replacement of ambiguous words with non-ambiguous words.

Furthermore, the marked ambiguous words in the sentences were only gender-ambiguous nouns. However, there may be other detected ambiguous words in these sentences, which were not marked. We do not limit our approach to only detecting gender ambiguous words, but we currently have only the means to evaluate for such words due to the datasets we used.

9. Conclusion and Future Work

In this work, we developed an approach to detecting ambiguous words in text by inspecting the diversity of translation. Our method is context-independent, meaning that it does not rely on context relating to the ambiguity in order to detect ambiguous words. Furthermore, despite utilizing a dataset, focused on gender ambiguity, this does not limit the application of the approach to other types of ambiguity, provided that the ambiguous words are defined as words which have one version in the source language and multiple versions in the target language. With this work, we strive to contribute to preventing MT systems from making an unjustified assumption, which may lead to bias further on.

Our approach is based on the hypothesis that sentences containing an ambiguity produce less diverse backtranslations than sentences without an ambiguity. In our base experiment, we probe this by comparing a dataset containing ambiguous sentences with a dataset which replaces the ambiguous word in the sentences with a non-ambiguous word. For this purpose, we also compare two replacement methods: disambiguation using gender forcing with gender defining adjectives and replacement with common non-ambiguous words. We translate the ambiguous and the non-ambiguous datasets into the target language and backtranslate the translations back into the source language, after which we evaluate the generated translations and extract patterns relating to the diversity of translation.

The results from the base experiment show that replacing the ambiguous word in a sentence with the corresponding disambiguated word leads to generating more diverse backtranslations when generating translations using Beam search with beam size 100 and Sampling. This is an expected result according to the hypothesis. On the other hand, replacing the ambiguous word with a common unambiguous word did not achieve the desired result. We observed that the selected unambiguous words had far fewer synonyms in both the source and the target language, compared to the occupation words in the chosen dataset, which led to less diversity in translation, but not due to ambiguity. With that, we deem this replacement method unsuccessful in this specific scenario.

Furthermore, we tested our approach in a real-world experiment by using natural occurring sentences, containing ambiguity. Our approach used an unmasking model and manual replacement to replace each word in the sentences with a non-ambiguous word one at a time and compare the results of the original sentences with each unmasked sentence. The unmasked sentence, which differs the most from the original sentence in terms of number of unique backtranslations, indicates an ambiguous word. The ambiguous word is the word in the unmasked sentence, which was replaced in the original sentence. Although not every ambiguous word was uncovered in the sentences, manually replacing the ambiguous words with non-ambiguous words in the sentence led to more ambiguous words being detected overall, which speaks for the method being partially effective.

9.1. Answers to Research Questions

In this section, we answer the research questions, posed in the Introduction (see Section 1.2).

9.1.1. Subquestion 1

The first subquestion pertained to the diversity of translations. We can conclude from the evaluation of our approach that sentences containing an ambiguous word generate less diverse backtranslations. From the analysis on the correlation between the number of unique translations and backtranslations, we estimated that the diversity of translations is proportional to the diversity of backtranslations, therefore we can state that sentences containing an ambiguous word generate less diverse translations.

9.1.2. Subquestion 2

The second subquestion related to the influence of ambiguous and non-ambiguous words on the diversity of translations. From the evaluation, we observed that sentences containing an ambiguous word place more attention in translation generation on the diversity of translations of the ambiguous word compared to the rest of the sentence, resulting in more unique translations of the ambiguous word compared to the non-ambiguous words in the same sentence.

By comparing sentences with an ambiguous word to sentences with the disambiguated version of the ambiguous word, we can conclude that disambiguation leads to more diverse backtranslations, proving the initial assumption that sentences containing an ambiguous word generate less unique backtranslations than sentences without ambiguity.

9.1.3. Main Research Question

This study aimed to answer the main research question of how we can detect ambiguous words in text. Our approach of disambiguating gender ambiguous words with gender forcing using the gender defining words *male* and *female* proved effective when decoding with Beam search with beam size 100, as well as with Sampling. However, this method is only applicable to gender ambiguous words.

On the other hand, replacing the gender-ambiguous words with common non-ambiguous words appeared ineffective in the specific scenario, because the gender ambiguous occupation words have more versions on average in both the source and the target language compared to the chosen common words. Using a similar approach in the real-world scenario seemed to work only sometimes. Therefore, more work is required towards developing a more universal method of detection of all types of ambiguity.

Next, we propose possibilities for future work, based on the findings of this thesis.

9.2. Future Work

In the future, we would like to explore the possibility of also detecting other types of ambiguity except gender. Unfortunately, our proposed disambiguation method using the gender words *male* and *female* is limited only to gender-ambiguous words. For this, we would need to develop a more general disambiguation method to be able to disambiguate generally ambiguous words in text. We would also benefit from an appropriate dataset containing ambiguous words of different types.

While the focus of this work lies mainly on the proof-of-concept, achieved with the base experiment, the real-world experiment sets the scene for a promising future for the approach. Therefore, it would be useful to execute the real-world experiment with a larger corpus and to test with different decoding strategies, such as Sampling. Also, manually replacing the ambiguous words is cost ineffective and would require automatization to be applicable for a larger dataset.

Furthermore, it may also be beneficial to use an Unsupervised Word Sense Disambiguation (WSD) approach for detecting ambiguous words. WSD is a technique in Natural Language Processing (NLP), defined as the ability to determine which meaning of a word is activated by the use of the word in a particular context. Our method can also be applied to a WSD test set by inspecting the diversity of translation of the ambiguous word in the sentence compared to the diversity of translation of the rest of the sentence.

Finally, we propose that researching methods for Quality Estimation (QE) may help to detect possible biases in translation. QE is a method for predicting the quality of a given translation rather than assessing how similar it is to a reference. For example, following multiple beams in Beam search indicates low confidence in translation, which may point to a possibility for error in translation due to ambiguity.

Bibliography

[1] Bashar Alhafni, Nizar Habash, and Houda Bouamor. "Gender-aware reinflection using linguistically enhanced neural models". In: *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*. 2020, pages 139–150.

[2] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. "Neural machine translation by jointly learning to align and translate". In: *arXiv preprint arXiv:1409.0473* (2014).

[3] Luisa Bentivogli, Beatrice Savoldi, Matteo Negri, Mattia Antonino Di Gangi, Roldano Cattoni, and Marco Turchi. "Gender in danger? evaluating speech translation technology on the MuST-SHE corpus". In: *arXiv preprint arXiv:2006.05754* (2020).

[4] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. "Man is to computer programmer as woman is to homemaker? debiasing word embeddings". In: *Advances in neural information processing systems* 29 (2016).

[5] Roldano Cattoni, Mattia Antonino Di Gangi, Luisa Bentivogli, Matteo Negri, and Marco Turchi. "MuST-C: A multilingual corpus for end-to-end speech translation". In: *Computer Speech & Language* 66 (Mar. 2021), page 101155. DOI: 10.1016/j.csl.2020.101155. URL: https://doi.org/10.1016%2Fj.csl.2020.101155.

[6] Won Ik Cho, Ji Won Kim, Seok Min Kim, and Nam Soo Kim. "On Measuring Gender Bias in Translation of Gender-neutral Pronouns". In: *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*. Association for Computational Linguistics, 2019. DOI: 10.18653/v1/w19-3824. URL: https://doi.org/10.18653%2Fv1%2Fw19-3824.

[7] Marta R Costa-jussà and Adrià de Jorge. "Fine-tuning neural machine translation on gender-balanced datasets". In: *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*. 2020, pages 26–34.

[8] Marta R Costa-jussà, Pau Li Lin, and Cristina España-Bonet. "GeBioToolkit: Automatic extraction of gender-balanced multilingual corpus of Wikipedia biographies". In: *arXiv preprint arXiv:1912.04778* (2019).

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. "Bert: Pre-training of deep bidirectional transformers for language understanding". In: *arXiv preprint arXiv:1810.04805* (2018).

[10] Zi-Yi Dou and Graham Neubig. "Word Alignment by Fine-tuning Embeddings on Parallel Corpora". In: *Conference of the European Chapter of the Association for Computational Linguistics (EACL)*. 2021.

[11] Chris Dyer, Victor Chahuneau, and Noah A Smith. "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*. 2013, pages 644–648.

[12] Joel Escudé Font and Marta R. Costa-jussà. "Equalizing Gender Bias in Neural Machine Translation with Word Embeddings Techniques". In: *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*. Association for Computational Linguistics, 2019. DOI: 10.18653/v1/w19-3821. URL: https://doi.org/10.18653%2Fv1%2Fw19-3821.

[13] Hila Gonen and Kellie Webster. "Automatically Identifying Gender Issues in Machine Translation using Perturbations". In: *Findings of the Association for Computational Linguistics: EMNLP 2020*. Association for Computational Linguistics, 2020. DOI: 10.18653/v1/2020.findings-emnlp.180. URL: https://doi.org/10.18653%2Fv1%2F2020.findings-emnlp.180.

[14] Nizar Habash, Houda Bouamor, and Christine Chung. "Automatic Gender Identification and Reinflection in Arabic". In: *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*. Association for Computational Linguistics, 2019. DOI: 10.18653/v1/w19-3822. URL: https://doi.org/10.18653%2Fv1%2Fw19-3822.

[15] Dirk Hovy, Federico Bianchi, and Tommaso Fornaciari. ""You Sound Just Like Your Father" Commercial Machine Translation Systems Include Stylistic Biases". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2020. DOI: 10.18653/v1/2020.acl-main.154. URL: https://doi.org/10.18653%2Fv1%2F2020.acl-main.154.

[16] Melvin Johnson. "A scalable approach to reducing gender bias in Google translate". In: *Google AI blog* (2020).

[17] Wandri Jooste, Rejwanul Haque, and Andy Way. "Philipp Koehn: Neural Machine Translation". In: *Machine Translation* 35.2 (June 2021), pages 289–299. DOI: 10.1007/s10590-021-09277-x. URL: https://doi.org/10.1007%2Fs10590-021-09277-x.

[18] Kaiji Lu, Piotr Mardziel, Fangjing Wu, Preetam Amancharla, and Anupam Datta. "Gender bias in neural natural language processing". In: *Logic, Language, and Security: Essays Dedicated to Andre Scedrov on the Occasion of His 65th Birthday* (2020), pages 189–202.

[19] Michal Měchura. "A Taxonomy of Bias-Causing Ambiguities in Machine Translation". In: *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*. Association for Computational Linguistics, 2022. DOI: 10.18653/v1/2022.gebnlp-1.18. URL: https://doi.org/10.18653%2Fv1%2F2022.gebnlp-1.18.

[20] Amit Moryossef, Roee Aharoni, and Yoav Goldberg. "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*. Association for Computational Linguistics, 2019. DOI: 10.18653/v1/w19-3807. URL: https://doi.org/10.18653%2Fv1%2Fw19-3807.

[21] Nathan Ng, Kyra Yee, Alexei Baevski, Myle Ott, Michael Auli, and Sergey Edunov. "Facebook FAIR's WMT19 News Translation Task Submission". In: *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*. Association for Computational Linguistics, 2019. DOI: 10.18653/v1/w19-5333. URL: https://doi.org/10.18653%2Fv1%2Fw19-5333.

[22] Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. "fairseq: A fast, extensible toolkit for sequence modeling". In: *arXiv preprint arXiv:1904.01038* (2019).

[23] Adam Paszke et al. "PyTorch: An Imperative Style, High-Performance Deep Learning Library". In: *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc., 2019, pages 8024–8035. URL: http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

[24] Jeffrey Pennington, Richard Socher, and Christopher Manning. "Glove: Global Vectors for Word Representation". In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 2014. DOI: 10.3115/v1/d14-1162. URL: https://doi.org/10.3115%2Fv1%2Fd14-1162.

[25] Marcelo O. R. Prates, Pedro H. Avelar, and Luìs C. Lamb. "Assessing gender bias in machine translation: a case study with Google Translate". In: *Neural Computing and Applications* 32.10 (Mar. 2019), pages 6363–6381. DOI: 10.1007/s00521-019-04144-6. URL: https://doi.org/10.1007%2Fs00521-019-04144-6.

[26] Nicholas Roberts, Davis Liang, Graham Neubig, and Zachary C Lipton. "Decoding and diversity in machine translation". In: *arXiv preprint arXiv:2011.13477* (2020).

[27] Rachel Rudinger, Chandler May, and Benjamin Van Durme. "Social Bias in Elicited Natural Language Inferences". In: *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*. Association for Computational Linguistics, 2017. DOI: 10.18653/v1/w17-1609. URL: https://doi.org/10.18653%2Fv1%2Fw17-1609.

[28] Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. "Gender Bias in Coreference Resolution". In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. Association for Computational Linguistics, 2018. DOI: 10.18653/v1/n18-2002. URL: https://doi.org/10.18653%2Fv1%2Fn18-2002.

[29] Danielle Saunders and Bill Byrne. "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*. Association for Computational Linguistics, 2020. DOI: 10.18653/v1/2020.acl-main.690. URL: https://doi.org/10.18653%2Fv1%2F2020.acl-main.690.

[30] Danielle Saunders, Rosie Sallis, and Bill Byrne. "Neural Machine Translation Doesn't Translate Gender Coreference Right Unless You Make It". In: *arXiv preprint arXiv:2010.05332* (2020).

[31] Beatrice Savoldi, Marco Gaido, Luisa Bentivogli, Matteo Negri, and Marco Turchi. "Gender Bias in Machine Translation". In: *Transactions of the Association for Computational Linguistics* 9 (2021), pages 845–874. DOI: 10.1162/tacl_a_00401. URL: https://doi.org/10.1162%2Ftacl_a_00401.

[32] Jürgen Schmidhuber, Sepp Hochreiter, et al. "Long short-term memory". In: *Neural Comput* 9.8 (1997), pages 1735–1780.

[33] Gabriel Stanovsky, Noah A. Smith, and Luke Zettlemoyer. "Evaluating Gender Bias in Machine Translation". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2019. DOI: 10.18653/v1/p19-1164. URL: https://doi.org/10.18653%2Fv1%2Fp19-1164.

[34] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. "Sequence to sequence learning with neural networks". In: *Advances in neural information processing systems* 27 (2014).

[35] Zhixing Tan, Shuo Wang, Zonghan Yang, Gang Chen, Xuancheng Huang, Maosong Sun, and Yang Liu. "Neural machine translation: A review of methods, resources, and tools". In: *AI Open* 1 (2020), pages 5–21. DOI: 10.1016/j.aiopen.2020.11.001. URL: https://doi.org/10.1016%2Fj.aiopen.2020.11.001.

[36] Iulia Turc, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. "Well-read students learn better: On the importance of pre-training compact models". In: *arXiv preprint arXiv:1908.08962* (2019).

[37] Eva Vanmassenhove, Christian Hardmeier, and Andy Way. "Getting Gender Right in Neural Machine Translation". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2018. DOI: 10.18653/v1/d18-1334. URL: https://doi.org/10.18653%2Fv1%2Fd18-1334.

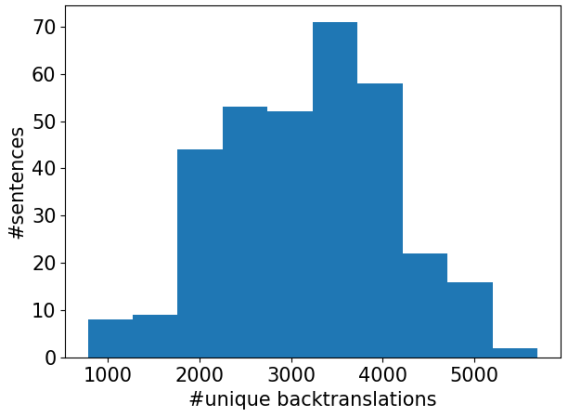
[38] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. "Attention is all you need". In: *Advances in neural information processing systems* 30 (2017).

[39] Z. Xiaojun. "Philipp Koehn: Statistical Machine Translation." In: *Applied Linguistics* 32.3 (Apr. 2011), pages 359–362. DOI: 10.1093/applin/amr017. URL: https://doi.org/10.1093%2Fapplin%2Famr017.

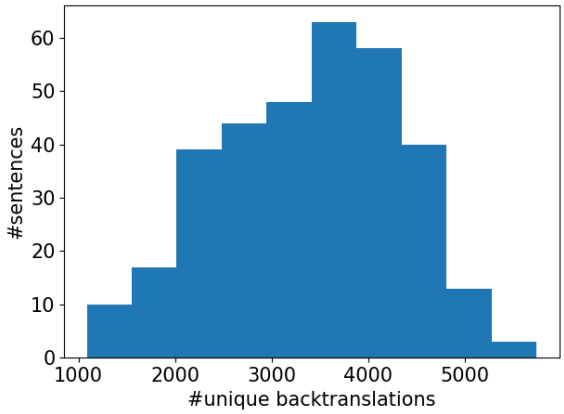
[40] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. "Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods". In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. Association for Computational Linguistics, 2018. DOI: 10.18653/v1/n18-2003. URL: https://doi.org/10.18653%2Fv1%2Fn18-2003.

[41] Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. "Learning Gender-Neutral Word Embeddings". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2018. DOI: 10.18653/v1/d18-1521. URL: https://doi.org/10.18653%2Fv1%2Fd18-1521.

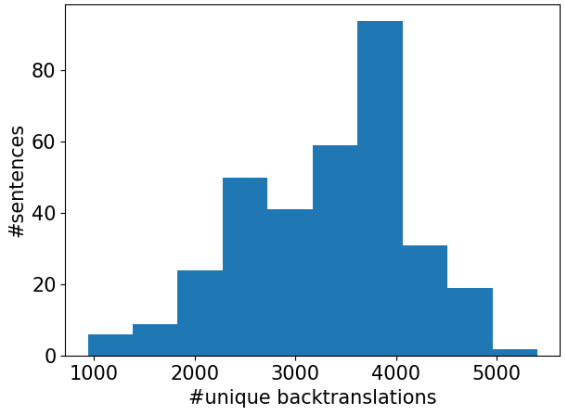
A. Appendix



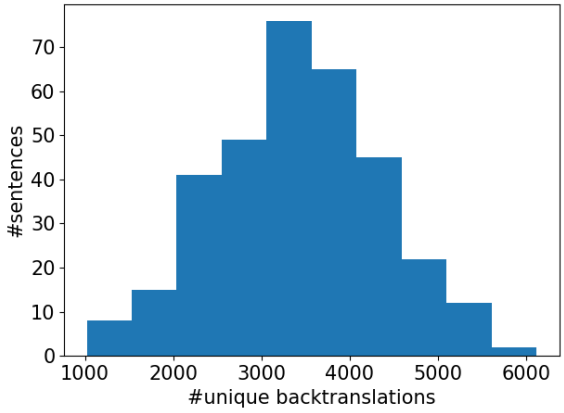
(a) Ambiguous Subset



(b) Disambiguated Subset (male)

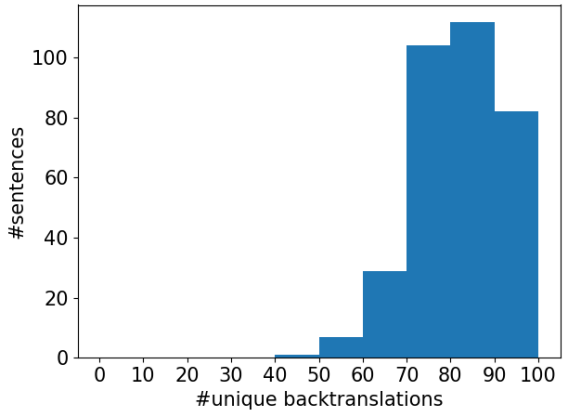


(c) Non-ambiguous Subset Average

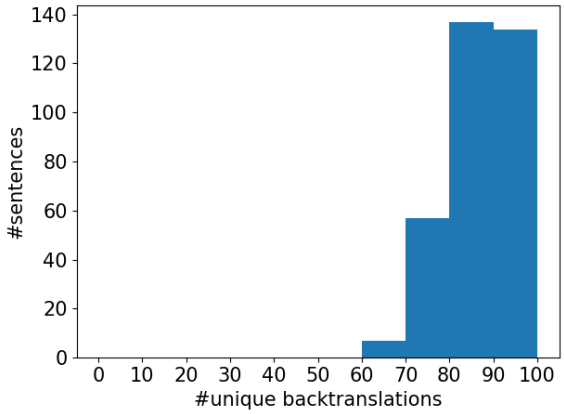


(d) Disambiguated Subset (female)

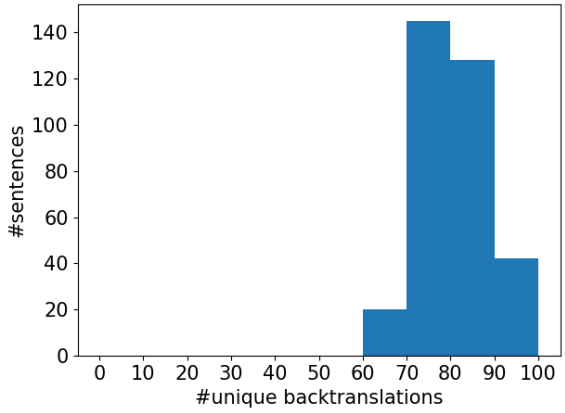
Figure A.1.: Distribution of Unique Backtranslations: Beam search with beam size 100



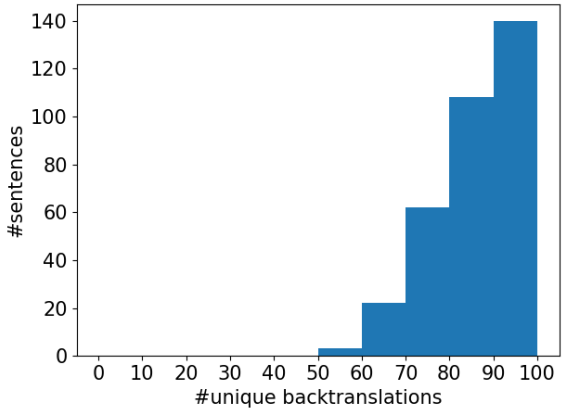
(a) Ambiguous Subset



(b) Disambiguated Subset (male)

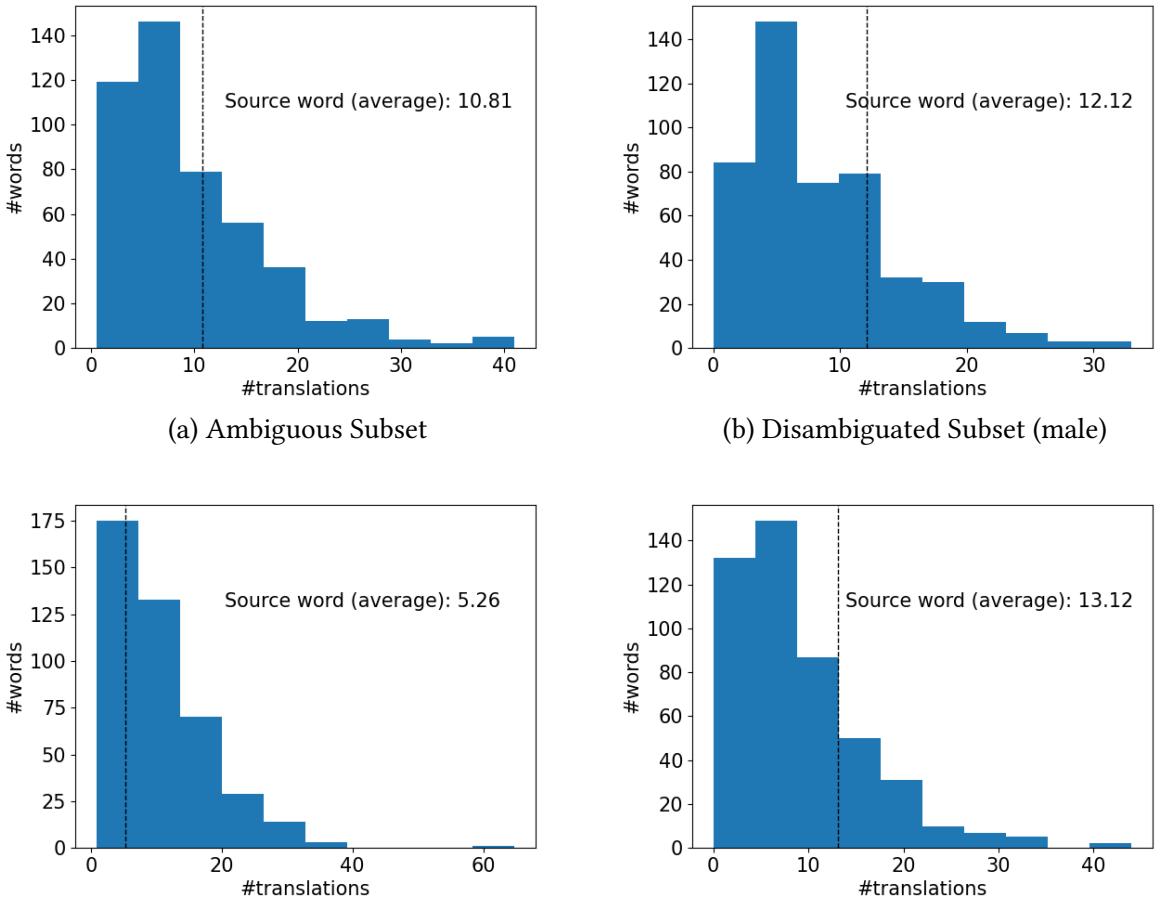


(c) Non-ambiguous Subset Average



(d) Disambiguated Subset (female)

Figure A.2.: Distribution of Unique Backtranslations: Sampling



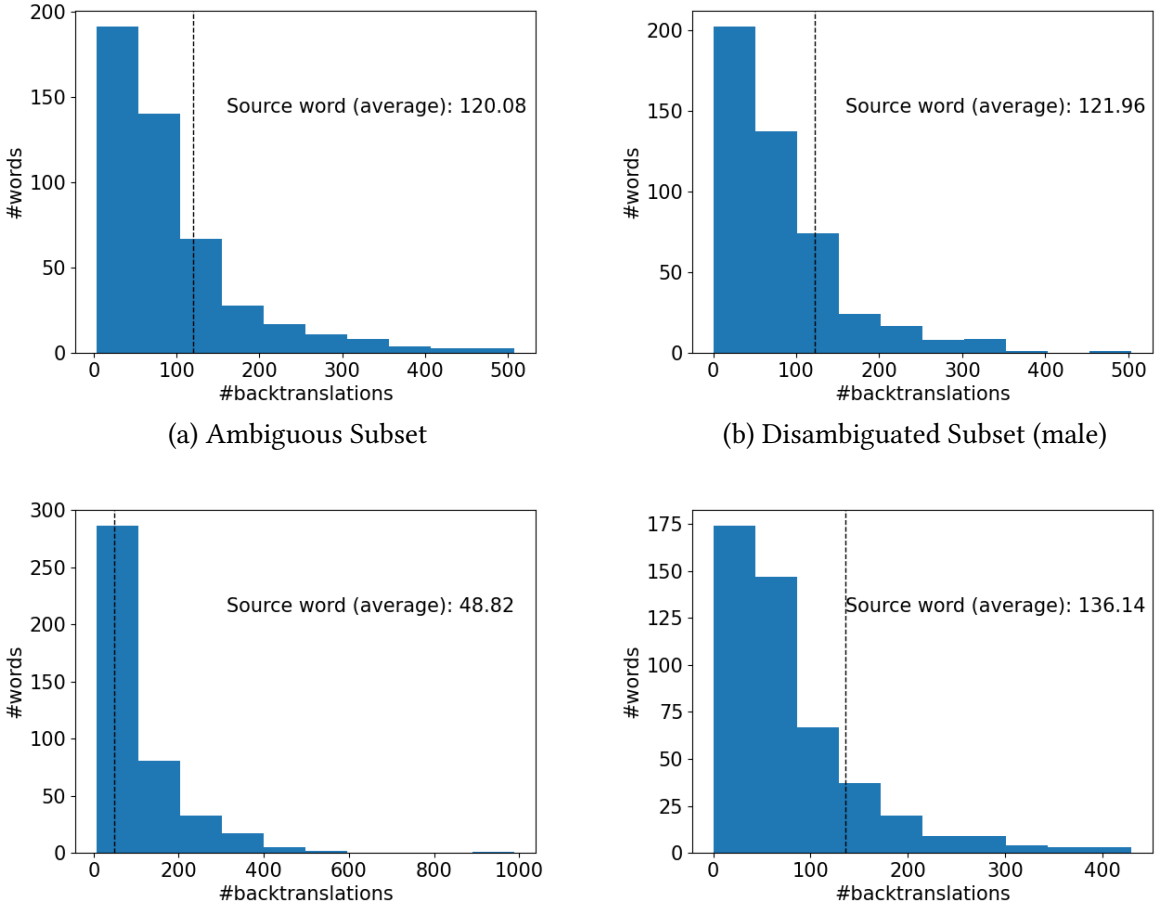
(a) Ambiguous Subset

(b) Disambiguated Subset (male)

(c) Non-ambiguous Subset Average

(d) Disambiguated Subset (female)

Figure A.3.: **Distribution of Unique Translations for Words**. Beam search with beam size 100. Nbest size 100. Alignment with *awesome-align*. The dashed line marks the average number of unique translations for the source word, the value displayed to the right.



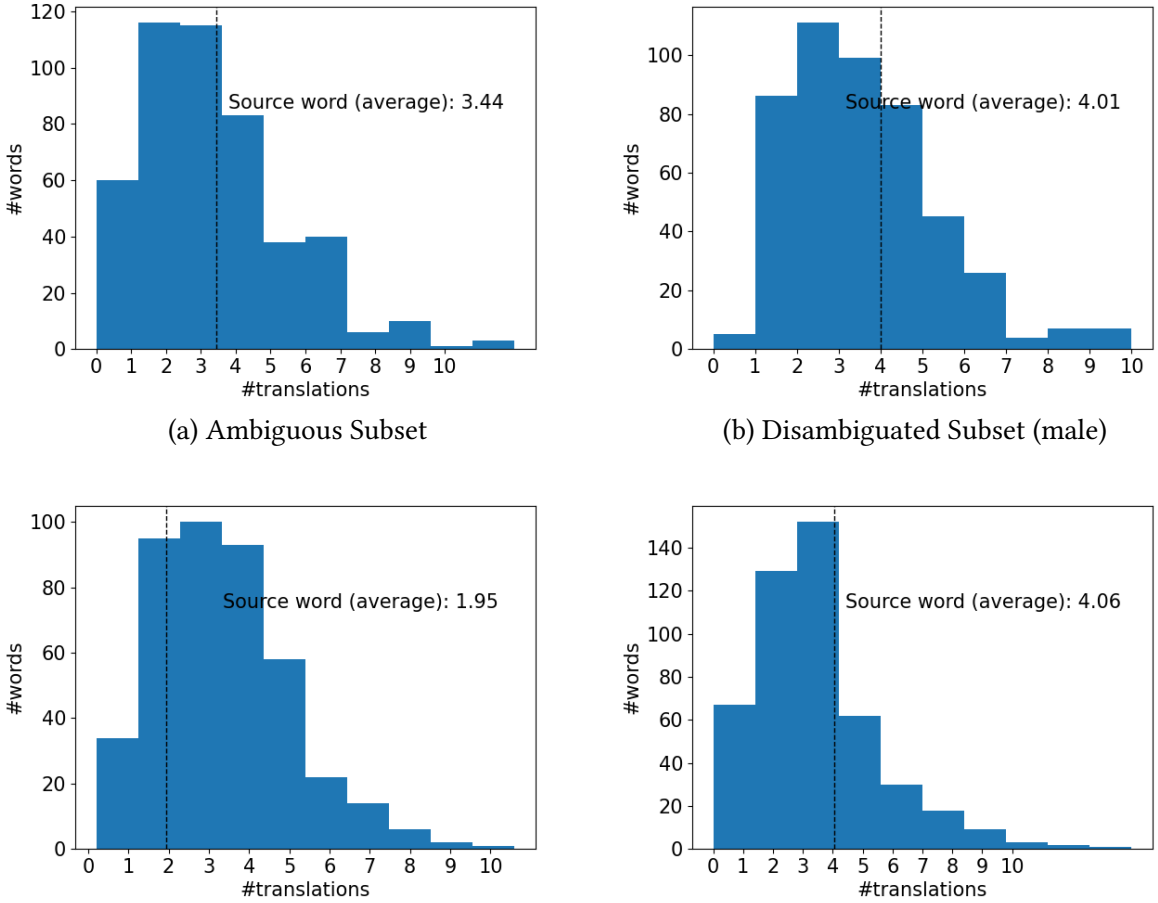
(a) Ambiguous Subset

(b) Disambiguated Subset (male)

(c) Non-ambiguous Subset Average

(d) Disambiguated Subset (female)

Figure A.4.: **Distribution of Unique Backtranslations for Words**. Beam search with beam size 100. Nbest size 100. Alignment with *awesome-align*. The dashed line marks the average number of unique translations for the source word, the value displayed to the right.



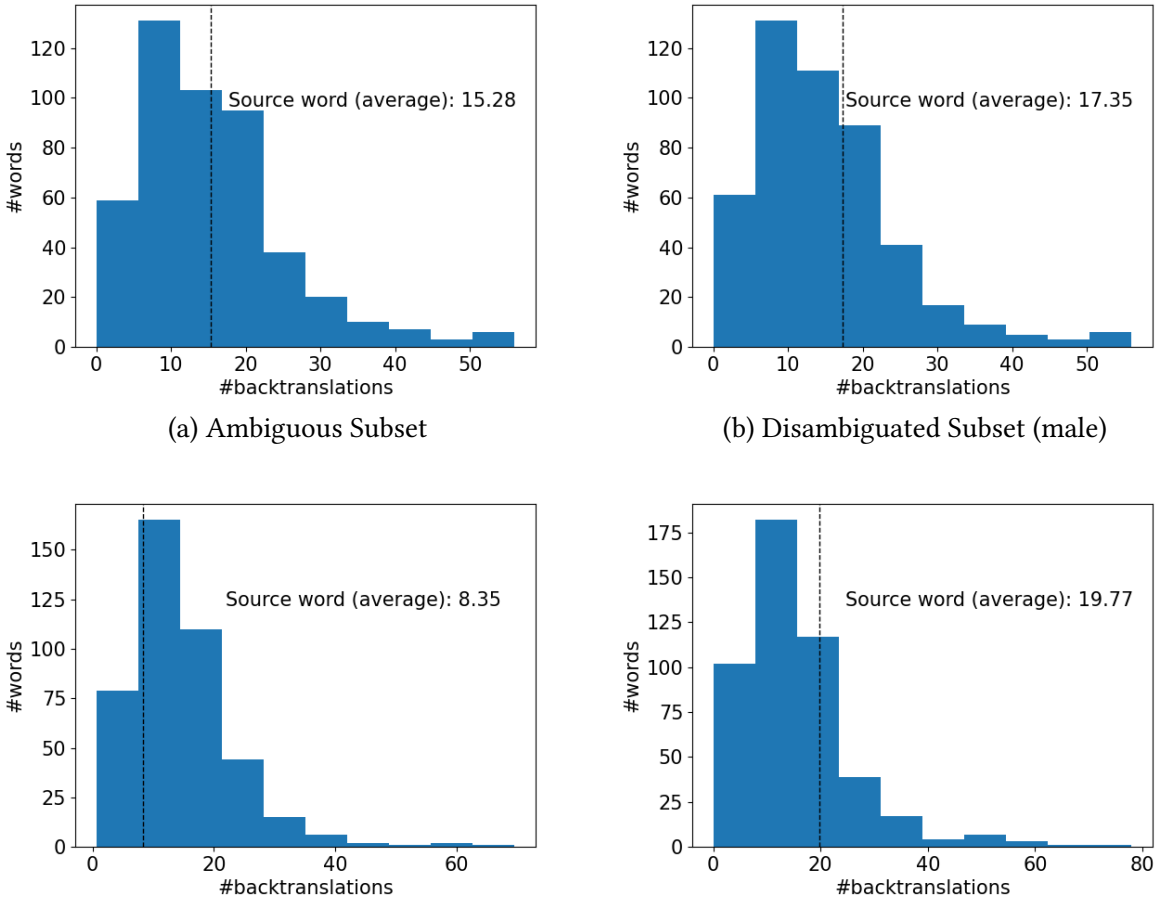
(a) Ambiguous Subset

(b) Disambiguated Subset (male)

(c) Non-ambiguous Subset Average

(d) Disambiguated Subset (female)

Figure A.5.: **Distribution of Unique Translations for Words**. Sampling. Nbest size 10. Alignment with *awesome-align*. The dashed line marks the average number of unique translations for the source word, the value displayed to the right.



(a) Ambiguous Subset

(b) Disambiguated Subset (male)

(c) Non-ambiguous Subset Average

(d) Disambiguated Subset (female)

Figure A.6.: **Distribution of Unique Backtranslations for Words**. Sampling. Nbest size 10. Alignment with *awesome-align*. The dashed line marks the average number of unique translations for the source word, the value displayed to the right.