UNIVERSITAT POLITÈCNICA
DE CATALUNYA
BARCELONATECH

Gender bias in natural language processing

Christine Raouf Saad Basta

Universitat Politècnica de Catalunya

PhD Thesis

Gender Bias in Natural Language Processing

Christine Raouf Saad Basta

Advisor:

Marta Ruiz Costa-Jussà

Tutor:

Josè Adrián Rodríguez Fonollosa

2022

"The journey is the treasure."

LLOYD ALEXANDER, The Golden Dream of Carlo Chuchio

"Success is not final, failure is not fatal: it is the courage to continue that counts."

WINSTON CHURCHILL

"Anyone who stops learning is old, whether at twenty or eighty. Anyone who keeps learning stays young. The greatest thing in life is to keep your mind young."

HENRY FORD

Abstract

Gender bias is a dangerous form of social bias that impacts an essential group of people. The effect of gender bias is propagated to our data, causing the accuracy of the predictions in models to be different depending on gender. In the deep learning era, our models are highly impacted by the training data transferring the negative biases in the data to the models. Natural Language Processing models encounter this amplification of bias in the data.

To understand and manage the effect of bias amplification, we are exploring the evaluation and mitigation approaches. The scientific society has exerted significant efforts in these two directions to enable proposing solutions to the problem. Our thesis is devoted to these two main directions; proposing evaluation schemes, whether as datasets or mechanisms, besides suggesting mitigation techniques. For evaluation, we proposed techniques for evaluating bias in contextualized embeddings and multilingual translation models. Besides, we presented benchmarks for evaluating bias for speech translation and multilingual machine translation models. For mitigation direction, we proposed different approaches in machine translation models by adding contextual text, contextual embeddings, or relaxing the architecture's constraints.

Our evaluation studies concluded that gender bias is encoded strongly in contextual embeddings representing professions and stereotypical nouns. We also unveiled that algorithms amplify the bias and that the system's architecture impacts the behavior. For the evaluation purposes, we contributed to creating several benchmarks. Firstly, we introduced a benchmark that evaluates gender bias in speech translation systems. This research suggests that the current state of speech translation systems does not enable us to evaluate gender bias accurately because of the low quality of speech translation systems. Additionally, we proposed a toolkit for building multilingual balanced datasets for training and evaluating NMT models. These datasets are balanced within the gender occupation-wise. We found out that high-resource languages usually tend to predict more precise male translations.

Our mitigation studies in NMT suggest that the nature of datasets and languages needs to be considered to apply the right approach. Mitigating bias can rely on adding contextual information. However, in other cases, we need to reconsider the model and relax some influencing conditions to the bias that do not affect the general performance but reduce the effect of bias amplification.

Acknowledgement

I am so grateful I came to the point of writing this in my PhD journey. The journey was worth living; it shaped my thinking and character, and I would not do it in any other way. There were so many valuable people along the journey who will always have a special place in my heart, and I will always keep them in my mind and heart.

First, I will start with my mentors. Marta, you were not only a PhD supervisor but also a companion, supporter, and friend. I can not thank you enough for all you taught me. You taught me lessons in research, life, career, and everything. I will always be grateful to you for all the conversations, the encouragement, the stories you shared, and the mentorship. Adrian, thanks for being such a role model to whom I am looking up. Thank you for enduring my basic questions, whether in teaching or research. I am grateful to have you as a mentor and have learned much from you. I also want to thank Kellie Webster in Google for being a great, helpful, knowledgeable mentor. I appreciate you a lot.

Now, coming to my family, two special people left our world while being in Barcelona, doing the PhD. Khalo Onsy, my dear uncle, I lost you in my PhD journey, but I know you would be proud of me now. You will always be in my heart, and I will never forget you. Tante Mervat, thanks for teaching me the most important lessons while leaving our world; thanks for your big heart. Pray for me while being with Khalo.

Now I would acknowledge the most important person in life, my mum, who always believed in me when I least did. Mum, you are my real hero in life. Throughout my life, you encouraged me and supported me in every way. Mum, you are amazing and inspiring. I love you; thank you for everything. Dad, thanks for always waiting for me late at night at the airport. Thank you for always ensuring that I do not need anything more and do not lack anything. Thanks for all the prayers you do. Lola, my dear sister, thanks for taking out so many responsibilities. You made sure to take care of me in every way. I love you and am blessed to have a fantastic sister like you. I always look up to you and appreciate your opinions and advice. My awesome cousins, who were the first to welcome me at home, Feby, and Engy. Thank you for always making me feel like I am there and not missing anything. Thanks for telling me the things Lola forget to tell :D. I always missed you here. Merna and Marlo, I know how special and strong you are. You make us proud and you are always in my thoughts. Mira and Mady, you are so special and close to my heart. To you all and all my other cousins, also to my dear aunts, I love you so much.

Coming to my awesome friends, Monica, thank you for always making sure I am ok and asking and caring about me, and thanks for always keeping me in your mind and life. I am so blessed for this. Thanks to my friends who visited me in Barcelona and left me with memories that will stay with me all over the years. First, I would thank Mariam, my close friend and sister, who always gives me warmth and happiness. Marie, Shimaa, Mai, Michael, and Rana, my very dear friends, thank you all for coming and making memories with me. To my friends in my work in Egypt, especially Reem, thank you for being such a good friend who always cared for me and took much responsibility for the paperwork.

My dear PhD mates, Bardia, Casimiro, Magdalena, Noe and Carlos. You were the first ones there on the team before I came. I remember your support and your help all over the way. I was very blessed to have you through this journey. Special thanks to Carlos, who helped me a lot through the paperwork and was always ready to answer any question even before that. To my new colleagues who added so much fun and exceptional taste to my PhD's last period: Gerard, Javier, Ioannis, Andre, Sant, and Belen. The last outings shaped super lovely memories for me. Thanks for being such nice company. Oriol, thanks for the pleasant cooperation. I enjoyed working with you. Everyone taught me something, and I am grateful to know each of you.

My special Barcelona Friends who became family and now special friends. Fatma, you added so much to my life here. I will always appreciate our time together. You made our time together adventurous, fun, and full of life. I am happy to gain such a friend for life. Marina, thanks for being such a faithful and understanding friend. Having you here was one of the best things. I would not give up on our lack of communication and will always keep you as a friend. Denise, I appreciate our talks a lot, you made me understand new concepts about life, and I appreciate our memories together. For many more :) To my other beautiful friends, Margo, Koki, Lili, Martina and Michael, Ireny and Peter, Mina Sameh, Besho, Pepo, Mina and Diana, you made Barcelona home. Special thanks to Marina Abadir, who helped me a lot from the beginning and always cared.

There are a lot of other beautiful friends that shared with me some parts of the journey. I appreciate everyone who shared with me any time during this journey. I am truly blessed to have you all and to have you as a part of the growth.

Contents

List of Figures

List of Tables

1 Introduction

Recently, fairness and ethical Artificial Intelligence (AI) have been a concern for AI researchers and scientists. Efforts from the industry and research target are flourishing to adopt AI principles toward more Ethical AI[1]. Bias can critically impact marginalization and suppression of under-represented societal categories and amplify discrimination towards vulnerable groups. Therefore, AI systems should be attainable and reliable to all classes, regardless of gender, race, or disability. Stakeholders should start giving strict consideration to fairness principles[2]. A famous form of bias is gender bias; it is a form that affects people's lives, especially the marginalized categories [Nadeem et al., 2020, Stanczak and Augenstein, 2021, Kiritchenko et al., 2021]. Gender bias is mainly the preference of one gender over the other in our systems.

Regarding applications influenced by gender bias, Natural Language Processing (NLP) is one of the most affected. NLP is a branch of AI that automatically teaches the machine to analyze and understand natural language machines. It significantly impacts our lives as we use its tools in our daily tasks, e.g., automatic translation, Google search auto-complete, and speech recognition systems. These applications improve our daily lives. However, we have seen that these applications can amplify social biases, e.g., gender bias [Sun et al., 2019, Mehrabi et al., 2021]. This bias can be perpetuated to models and downstream tasks, causing other harm to the end-users. This thesis focuses on the relation of gender bias with NLP for various reasons. Most importantly, at the time of starting the thesis, this topic was new, under-studied, and required multiple efforts to address the problem. Another important reason is the wide range of people affected by such bias.

Moreover, NLP and gender bias interaction is two-fold; NLP can be a tool for detecting bias in society and amplifying the gender bias in society by producing

[1] https://ai.google/principles

[2] https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

gender-biased systems [Costa-jussà, 2019]. NLP systems should encourage less discriminatory language within communities. If NLP systems are designed more fairly, we can trust the automated decisions, as they will be less contaminated by our biases and preferences [Stanczak and Augenstein, 2021].

Harms of gender bias in NLP are various and can be representation or allocation harms. Representation harm can be shown when certain concepts are associated with gender. This harm perpetuates inappropriate stereotypes about which activities men and women can do. It gives a false impression about what women are allowed or expected to perform, e.g., leading to less professional females in STEM (Science, technology, engineering, and mathematics) [McGuire et al., 2020]. It is also reflected when associations between gender with certain concepts are captured in word embeddings and model parameters [Sun et al., 2019, Stanczak and Augenstein, 2021]. When such word representations are used in downstream NLP applications, there is an additional risk of unequal performance across genders [Gonen and Webster, 2020].

On the other hand, allocation harm is reflected when a model performs with better accuracy on data associated with the majority gender (machine translation [Font and Costa-jussà, 2019, Stanovsky et al., 2019, Savoldi et al., 2021] and coreference resolution [Webster et al., 2018]).

Many questions need to be addressed. How should we design our NLP systems? How should we treat the data for fairer systems? How should we remove the human biases from our models? These questions lead us to understand the problem in our systems and help us mitigate their bias. We believe that building a fair and free-bias language system could have a beneficial social impact [Tomalin et al., 2021]; better systems that are not replicating our behaviors are way better and can positively impact society.

Bias Conceptualization Motivated by the recent work [Blodgett et al., 2020], researchers have recommended conceptualizing bias in future work. One relevant suggestion is to understand the harm bias can cause to the NLP systems and to whom. This understanding is crucial for moving forward in the right direction with the gender bias definition.

Gender bias can be defined as systematic, unequal treatment based on gender. It is one relevant factor that prevents our systems from being equitable. In machine translation, an example of occupation bias is demonstrated when translating from English to Spanish in our translation systems, generating more translations to male doctors than to female doctors. An example is translating *my friend is a doctor* to *mi amigo es un doctor* (male version). Sometimes, the system tends to predict biased translations even when there is context like *My friend and her daughter are doctors*, the system favors neglecting the context, translating *doctor* to the male version.

1.1 Objectives

Our long term research goal is to build fair NLP systems, that are representative and useful to all people. During this PhD, we focused on gender bias problem in NLP models. Our research can pave the road to reduce gender stereotypes in our systems and can benefit the NLP theory and practice to create more accountable systems. Bias can take different forms and affect our systems negatively; therefore, we need to understand how our models and data respond to such effects. Besides mitigating the effect of bias amplification in our models should be a main goal in designing our models. We are considering one type of bias; binary gender bias. We consider that this type of bias is one of the most harmful types affecting an important social sector. Therefore our objectives lie in two main research lines:

Bias Evaluation. We believe this is a critical step for any progress in this field of research. We must evaluate the systems to discover how the biases are encoded in the models and deal with such biases. This thesis proposed different methods to assess gender bias in contextualized embeddings and NMT architectures. We also analyzed multilingual scenarios in our evaluation studies. This objective is mainly covered in chapter 4. Besides, we would not accomplish an accurate evaluation without challenge datasets suitable for different tasks. For this purpose, we aim to propose a method to create such datasets used in training or evaluation. This aim is covered in chapter 6.

Bias Mitigation. This objective aspires to ensure the adoption of different strategies helps mitigate the bias. We focus here on the NMT task, as it is a multi-facet problem and the most challenging among all NMT tasks. Gender bias affects this kind of application in a way harming many users. The proposed methodologies vary from modifying the training procedure to aggregating different contextual information. The experimental results show that our methods can reduce bias without significantly affecting the model's performance. This objective is discussed in chapter 5.

1.2 Outcomes of the Thesis

Here we describe the outcome of the thesis in terms of publications and scientific contribution. We have publications directly related to the thesis (accepted and under submission) and some publications that are not. Besides, we mention here other related activities.

Thesis publications These publications are directly related to the thesis: Chapter 4 covers the following publications:

- [Basta et al., 2019] **Basta, C.**, Costa-jussà, M.R. and Casas, N. Evaluating the Underlying Gender Bias in Contextualized Word Embeddings, CORR, arXiv:2019, Proceedings of the 1st ACL Worskhop on Gender Bias for Natural Language Processing, 2019.
- [Basta et al., 2021] **Basta, C.**, Costa-jussà, M.R. and Casas, N. Extensive study on the underlying gender bias in contextualized word embeddings. Neural Computing and Applications 33, 3371–3384 (2021). https://doi.org/10.1007/s00521-020-05211-z
- [Costa-jussà et al., 2022] Costa-jussà, M.R., Escolano, C., **Basta C.**, Ferrando, J., Batlle, R. and Kharitonova, K., Interpreting Gender Bias in Neural Machine Translation: The Multilingual Architecture Matters, Accepted in 36th AAAI Conference, 2022.

Chapter 5 covers the following publications:

- [Basta et al., 2020] **Basta, C.**, Costa-jussà, M.R, Fonollosa, J.A.R, Towards Mitigating Gender Bias in a decoder-based Neural Machine Translation model by Adding Contextual Information, ACL Widening NLP Workshop (WiNLP), 2020.

- [Basta et al., 2022] **Basta, C.**, Escolano, C.,Costa-jussà, M.R. To Add or Relax? Examining Approaches for Mitigating Gender Bias in Machine Translation. (Under Submission Process).

Chapter 6 covers the following publications:

- [Costa-jussà et al., 2022b] Costa-jussà, M.R., **Basta, C.**, Gállego, G. I, Evaluating Gender Bias in Speech Translation, LREC 2022.
- [Costa-jussà et al., 2022a] Costa-jussà, M.R., **Basta, C.**, Domingo, O., Niyongabo, A., OccGen: Selection of Real-world Multilingual Parallel Data Balanced in Gender within Occupations (Under Submission Process).

Other publications We worked on some publications related to studying gender bias but not directly related to the work presented in this thesis:

- [Basta and Costa-jussà, 2021b] **Basta, C.** and Costa-jussà, M.R. Impact of Gender Debiased Word Embeddings in Language Modeling, CORR, arXiv:2021, Lecture Notes in Computer Science series LNCS Springer, CICLING, 2019. (Top 30% papers)
- [Basta and Costa-jussà, 2021a] **Basta, C.** and Costa-jussà, M.R. Impact of COVID-19 in Natural Language Processing Publications: a Disaggregated Study in Gender, Contribution and Experience, EACL First Workshop on Language Technology for Equality, Diversity, Inclusion, online.
- [Escolano et al., 2021b] Escolano, C., Ojeda, G., **Basta, C.** and Costa-jussà, M. R. Multi-Task Learning for Improving Gender Accuracy in Neural Machine Translation, 18th International Conference on Natural Language Processing (ICON), 2021.
- [Kharitonova et al., 2021] Kharitonova, K., Costa-jussà, M.R., Escolano, C., **Basta, C.** and Armengol-Estapé, J., Neutralizing Gender Bias in Neural Machine Translation by Introducing Linguistic Knowledge, WiNLP EMNLP 2021.

We co-operated in other publications for WMT Translation tasks:

- [Casas et al., 2019] Casas, N., Fonollosa, J.A.R., Escolano, **C., Basta**, C. and

Costa-jussà The TALP-UPC Machine Translation Systems for WMT19 News Translation Task: Pivoting Techniques for Low Resource MT. Proceedings of the ACL 4th Conference on Machine Translation, 1st-2nd August, Florence.

- [Escolano et al., 2021c] Escolano, C., Tsiamas, I., **Basta, C.** , Ferrando, J., Costa-Jussà, M.R., Fonollosa, J.A.R., The TALP-UPC Participation in WMT21 News Translation Task:an mBART-based NMT Approach, Proceedings of the 6th Conference on Machine Translation.

Scientific Contribution **Industrial Internship in Google AI Research (12/2020-03/2021)**. I worked on *Latent Bridge Augmentation for Machine Translation*, designing an NMT model to bridge the source language as a bridging representation that is scalable for any pair of languages and obtains the main features in the source language.

Co-organizing GEBNLP Workshop. Besides, I am co-organizing a workshop on *Gender Bias in Natural Language Processing*, which is totally dedicated to the research concerning this problem and raises the awareness of it [Hardmeier et al., 2022].

1.3 Thesis Outline

Our thesis is divided into seven chapters, following **Chapter 1** which describes the problem and our objectives; the other chapters cover the rest of the thesis. **Chapter 2 and 3** mainly provide the suitable background and literature essential for the following chapters. **Chapter 4** demonstrates the methodologies of evaluation of the gender bias in contextual embeddings and multilingual machine translation architectures. **Chapter 5** illustrates different mitigation techniques in NMT; increasing context and adding gender tag, adding documental information, and relaxing positional information in NMT task. **Chapter 6** presents WinoST, the multilingual speech evaluation dataset, and OccGen, the toolkit for selecting the real-world multilingual parallel data balanced in gender within occupations. **Chapter 7** concludes the contributions of the thesis, besides the insights about the current and future situation of the problem in NLP.

2 Background

This chapter covers all the theoretical parts and the essential concepts and techniques for the NLP tasks discussed throughout the thesis.

2.1 Natural Language Processing

Natural Language Processing (NLP) is a broad area that creates methods for dealing with unstructured natural data. Usually, human language is ambiguous. Consider, for example, *Before passing by the bank to get my salary, I sat on the bank of the nile*. Humans can distinguish between the two cases of *bank* because they naturally perceive and illustrate the languages. When humans learn languages, they do not know the rules of their native languages; they start to understand, perceive, interpret, and reproduce naturally. This concept is quite a challenge for defining the regulations of the languages. To teach the computer to understand these languages, we need these rules that make language processing quite challenging. What makes the task more language is that language is compositional: words from letters and sentences from words. Such facts lead to data sparseness; the terms can be combined indefinitely and infinitely to form sentences [Goldberg, 2017]. This is challenging when the machine learns from examples; there will always be examples that were never seen in the training set.

As described, the task is quite challenging, but the NLP community has worked tremendously to develop different methods and architectures to tackle different problems; semantic and syntactic parsing, parts of speech understanding, co-reference resolution, sentiment analysis, etc. In the last decade, enormous progress has been made toward better accuracy of the NLP systems, thanks to the new technologies and advances.

2.2 Word Embeddings

Recently, new approaches have been revealed to compute the embeddings differently, employing different neural language model architectures. These approaches include such as ULMfit [Howard and Ruder, 2018], ELMo [Peters et al., 2018], OpenAI GPT [Radford et al., 2018, Radford et al., 2019] and BERT [Devlin et al., 2019], which are mainly pre-trained language models (LMs). They provide new LM architectures, and the pre-trained weights are available for usage in downstream tasks. These techniques prove that they enhanced the performance of several state-of-the-art benchmarks, including question answering on SQuAD, (cross-lingual) natural language inference, and named identity recognition.

ELMo [Peters et al., 2018] was one of the first techniques that depended on training recurrent neural networks as language models and then reusing the context vectors for each token as pre-trained word (token) vectors [Smith, 2020]. The neural architecture employed in ELMo consists of a character-level convolutional layer that creates a word representation after processing the characters of each word. Consequently, a language model task training is done with this representation into a 2-layer bi-directional long-short term memory [Hochreiter and Schmidhuber, 1997]. Since it uses a bi-directional architecture, the embedding relies on both the next and previous words in the sentence. ELMo provides word-level representations. [Peters et al., 2019] and [Liu et al., 2019a] assured the viability of using ELMo representations as features for downstream tasks without retraining the entire model on the target task.

Bert [Devlin et al., 2019] is a multi-layer bi-directional transformer-encoder model for learning contextualized embeddings, adopting the transformer architecture with self-attention layers [Vaswani et al., 2017]. BERT proposes a masked language modeling (MLM) objective, where some tokens are masked, and the aim is to predict them, given the masked sequence input. Special tokens are used in Bert to obtain a single contiguous sequence for each input sequence. Sentences are separated given a special separator token [SEP], and the first token is a special classification token [CLS]. BERT utilizes a pre-training technique followed by a fine-tuning scheme. Sentence-level tasks employ the final hidden state of [CLS], while the token-level tasks use the last hidden state of each token [Liu et al., 2020].

GPT [Radford et al., 2018] uses a two-stage learning paradigm: unsupervised pre-training employing a language modeling objective and supervised fine-tuning. The main goal is learning transferable embeddings to be used in multiple downstream tasks. These approaches pre-process each sentence as a single contiguous sequence of tokens through special tokens including [START] (the start of a sequence), [DELIM] (delimiting two sequences from the text input) and [EXTRACT] (the end of a sequence).

2.3 Neural Machine Translation (NMT)

2.3.1 Sequence-to-Sequence Models

Sequence-to-sequence models are the models that takes sequence of input $x = \{x_1, x_2, ..., x_n\}$ and generates sequence of output $y = \{y_1, y_2, ..., y_n\}$. The inputs can be different modalities, such as text, image, and speech. In this thesis, we discuss the text models only, as we are concerned with them in these studies.

Two main components are used for building different task architecture; the encoder and the decoder.

Encoder. The encoder is mainly responsible for constructing contextual output from the input tokens. The input sequence tokens are first fed to the encoder input. These input tokens are then embedded and provided through units or layers to create contextual representation as the output. The encoder represents the inputs and feeds them to the decoder in the case of encoder-decoder architecture.

Decoder. The decoder receives the encoder output as a context vector and starts generating the tokens. Each new sentence is marked with the input BOS (beginning of the sentence) token. The output is a softmax representing the categorical probability distribution over the output token space.

2.3.2 Neural Machine Translation Architectures

NMT, one of the important milestones in MT, has led to enormous improvements in accuracy. NMT has gained a lot of success in the last era, giving competitive results compared to human translations. Below we describe the main NMT architectures.

NMT models define a probability distribution over the target tokens P(y|x) by decomposing it into conditional probabilities:

$$p(y|x) = \prod_{j=1}^{J} p(y_j|y_1^{j-1}, x) \tag{2.1}$$

Recurrent Neural Networks (RNN). The first NMT architectures were encoder-decoder sequence-to-sequence RNN models [Sutskever et al., 2014], where both encoder and decoder, either vanilla RNN, LSTM [Hochreiter and Schmidhuber, 1997] or GRU [Cho et al., 2014].

RNNs [Elman, 1990] is a family of neural networks dealing with sequential data. RNN is a neural network specialized for processing a sequence of values $(x_i, ..., x_T)$ and can deal with variable length input. Parameter sharing allows it to extend and apply the model to examples of different forms (e.g., different lengths) and generalize across them, attending to structured properties. The main problem with RNNs is the vanishing gradient when the sequences tend to be extended. Therefore variances came to participate in this problem; Gated Recurrent Unit (GRU) [Cho et al., 2014], and Long Short-Term Memory (LSTM) [Hochreiter and Schmidhuber, 1997] which are capable of capturing the statistical regularities in sequential inputs.

LSTM architecture has proven a success in many applications. It was mainly designed to solve the vanishing gradients problem and was the first to introduce the gating mechanism. The structure of LSTM has two splits for the state vectors; memory cells and working cells. At each input, the gate determines a certain amount of the new input to be written to the memory cell, which can be forgotten. The memory cell is responsible for maintaining the memory for all inputs and error gradients.

A notable problem of RNN models is that the encoder has to fit all the information from the source sentence into a fixed-length vector representation (i.e., the context vector passed from encoder to decoder). [Bahdanau et al., 2015] introduced the concept of attention to avoid having a fixed-length source sentence representation. The model does not need the whole context vector; instead, the decoder attends to certain parts of the source sentence that are useful for producing the next token. The attention-based NMT models [Bahdanau et al., 2015] also allowed the decoder to use the weighted sum of the encoder's context vectors. These attention-based models outperformed the vanilla RNN in translation accuracy and quality.

At inference time, the encoder receives the sequence of tokens as the input, generating a representation given to the decoder. The decoder then generates the probability distribution over the target token space, conditioning on the previous token generated. A decoding algorithm must select the predicted token. While a greedy algorithm seems to be the solution, it does not guarantee that the best sequence of tokens will be generated. Therefore, the **beam search** decoding algorithm [Graves, 2012] is used. It is s a decoding algorithm that depends on the hypothesis that sequences with high probability have high probability conditionals. It follows a greedy search, but instead of keeping only the highest probable token, it keeps n most probable ones, known as the beam. Every step, the search generates predictions based on the beam of the previous step.

For evaluation, **BLEU** score (BiLingual Evaluation Understudy) [Papineni et al., 2002] is the most standard evaluation metric. It is based on comparing the candidate translation (hypothesis) with one or multiple reference translations. In most cases, due to limited resources, only one reference is considered.

Transformer. Transformer [Vaswani et al., 2017] is an encoder-decoder architecture, each with multiple layers of multi-head attention, normalization, feed-forward layers with residual connections as depicted in Figure 2.1. The multi-head attention concept was first introduced in transformers, where each hidden state has multiple keys (K), values (V), and queries (Q) vectors which generate different attention distributions. Each head computes as in eq. 2.2. This gives the feasibility of paying attention to different heads simultaneously.

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d}})V \tag{2.2}$$

The input is not provided sequentially through layers. In the beginning, positional embeddings are added to the input and output tokens to keep track of their positions. The decoder self-attention blocks are masked to ensure the causality of predictions. In training time, the model is trained in a complete parallel mode. In inference time, it is auto-regressive.

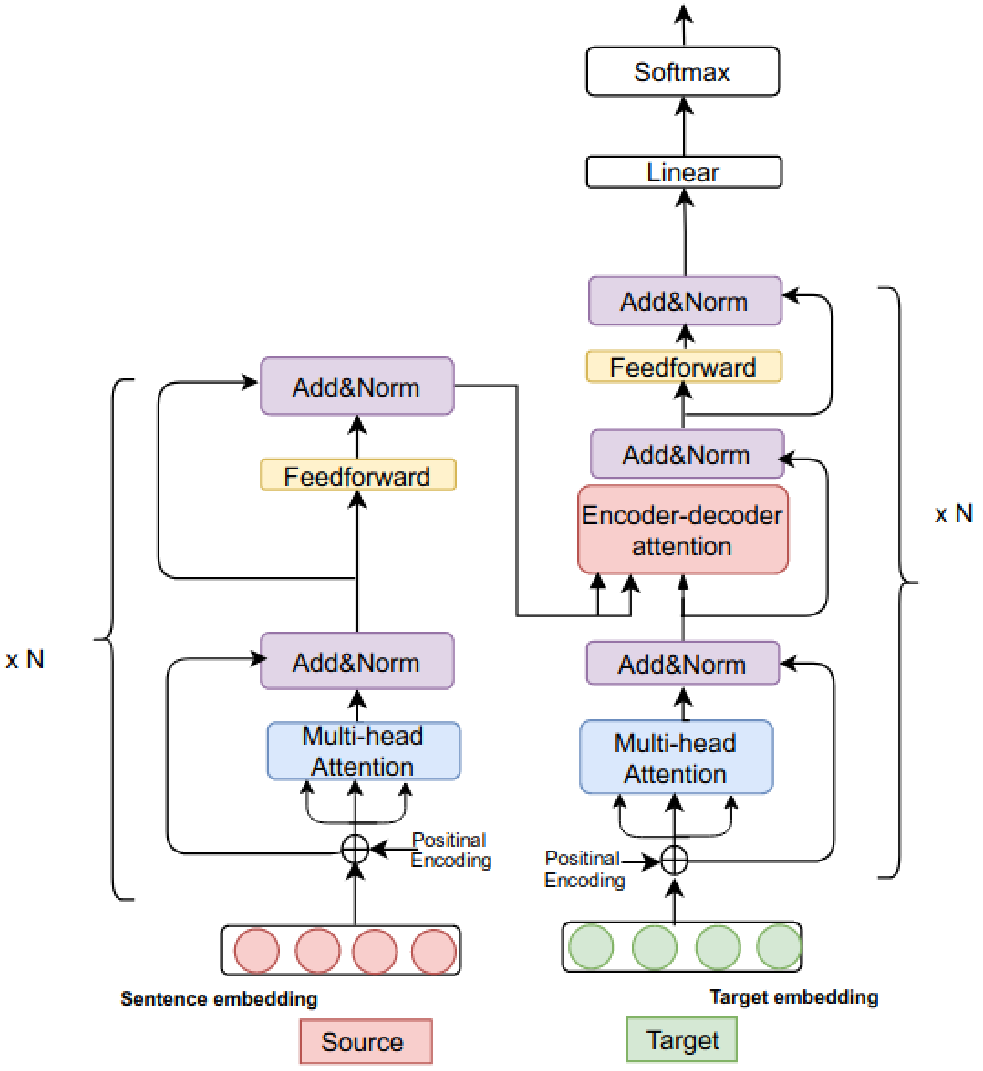


Figure 2.1: Transformer architecture.

Multilingual Architectures. As described above, most NMT architectures are based on the Transformer [Vaswani et al., 2017]. Different from the bilingual NMT Transformer, which devotes the entire representation capacity of the model to a single task, capturing specific features and correlations of the language pair, several alternatives exist to extend it to a multilingual system. Here, we briefly describe the multilingual architectures explored in Chapter 4.

- **Shared Encoder-Decoder** [Johnson et al., 2017] train a single encoder and decoder with multiple input and output languages. A shared architecture has a universal encoder and decoder fed with all initial language pairs at once. The model shares vocabulary and parameters among languages to ensure that no additional ambiguity is introduced in the representation. By sharing a single model across all languages, the system can represent all languages in a single space. The model then allows translation between language pairs never seen during the training process, known as zero-shot translation.
- **Language-Specific Encoders-Decoders** Architectures of this category may vary from sharing some layers [Firat et al., 2017, Lu et al., 2018] to no sharing at all [Escolano et al., 2021a, Escolano, 2022]. The latter approach is the most contrastive to the shared encoder-decoder. The language-specific (with no sharing) approach involves training independent encoders and decoders for each language. Different from standard pairwise training, in this case, there is only one encoder and one decoder for each language. Since parameters are not shared, this joint training enables new languages without retraining the existing modules, which is a clear advantage relative to the previously shared encoder-decoder.

3 Literature Survey and Related Work

In this chapter, we describe the research approaches the scientific community has explored to work towards gender bias problem resolution in tasks related to our thesis. We also describe some concepts related to gender bias.

3.1 Gender Bias

Type	**Bias Definition**
Group bias	A system's decisions are skewed toward a particular group of people [Mehrabi et al., 2021].
Individual bias	A system is biased if it gives different predictions, which are less favorable to individuals within a particular group, where there is no relevant difference between these groups that justifies such harms [Dwork et al., 2012].
Predictive Bias	The mismatch of ideal and actual distributions of labels and user attributes in training and application of a system [Hovy and Prabhumoye, 2021, Shah et al., 2020].

Table 3.1: Examples of bias definitions.

As argued by [Blodgett et al., 2020], there are several under-specified concepts of bias addressing the datasets in NLP tasks, fostering the imprecision of the terminology *Gender Bias*. *Gender bias* is manifested in several distinct ways in NLP tasks; consequently, it is not a homogenous phenomenon. The focus of gender bias is different according to the task; accordingly, it is easier to detect in some tasks than in others. For example, in the context of MT, the focus has been mainly on the representational harms that arose from stereotyping specific linguistic structures and items. This type of bias is identifiable within a dataset. Nevertheless, identifying the effect of power, society, and politics on gender bias in the dataset is highly complicated.

Bias can be defined from different perspectives, as mentioned in the table 3.1; however, we can wonder why NLP is impacted by such definitions immensely. Understanding the connection between the power of language and stereotypical beliefs about women's invisibility explains the automatic transfer of this connection to NLP tasks. Many individuals can be impacted by such bias, as we see women misgendered in the context of MT and losing their linguistic style.

3.2 Sources of Gender Bias

Some sources of bias are common across all NLP applications, and those sources are the main root of causing different types of biases [Suresh and Guttag, 2021]. Nevertheless, some sources of bias can differ from a task to another.We are stating the common sources as follows:

Data is a Main Source. Scientists believe that the primary source of bias is data. Data can be gathered containing constructed biases and inaccuracies. Consequently, feeding such data into an AI system may change its behavior and cause inaccurate results. The question here is whether the integrity of the data should be guaranteed before training, or this is a highly complex task to do and can not be guaranteed. The community is raising awareness that such inaccuracies and biases should be removed from data. However, there is an argument saying that eliminating gender bias from data leads to another sort of bias, as the data would be skewed in a 'positive' way rather than a 'negative' way [Tomalin et al., 2021]. Another argument is that if data capture distinctive skewings in the sample population, the data is not biased [Prates et al., 2020]. We can agree that processes and data sets must be tested and documented at each step, such as planning, training, testing, and deployment.

Bias in Data Annotation and Selection. A possible significant source of bias can be the underspecified annotations guidelines and the positionality of annotators. Annotators can get distracted and uninterested in the annotation task, choosing 'wrong' labels for annotating the data, thus introducing bias. The data selection is the other face of the problem, which should be tackled together with the annotation. When choosing a text dataset to work with, there are many decisions to be taken about the demographic groups represented in the data. Such decisions are humanly made and thus can have built-in prejudices, adding bias to the system. These steps lead to ageist, racist, or sexist models biased against the respective user groups [Cao and Daumé, 2021, Hovy and Prabhumoye, 2021].

Bias Encoded in Learnt Embeddings. Word embeddings have been shown to raise the racial and gender biases in the training data [Bolukbasi et al., 2016]. These biases are resistant to corrections and hard to remove from the embeddings [Gonen and Goldberg, 2019]. Moreover, these biases are even transferred to the contextualized embeddings [Zhao et al., 2018b]. Thus, the bias is encoded in the embeddings and learned during the model training, adding a profound reason for perpetuating gender bias to the system [Hovy and Prabhumoye, 2021].

Bias in Model. Models themselves may over-amplify bias. One source can be the loss objective used in training the models. The main target of these objectives is increasing the predictions so the models might utilize spurious examples (e.g., all positive examples coming from female nurses, then the gender is considered a discriminative feature) or statistical irregularities in the dataset. Such behavior is rooted in the models, hard to track, and hard to address [Hovy and Prabhumoye, 2021]. Another source is how the model is defined, what features it uses, what decisions are made, and how predictions are ranked. For instance, a model can make positive sentiment guesses depending on a map of certain words.

Bias in System Testing. Evaluation metrics may be a source of bias. The metrics may weight errors differently, reflecting the false cost of weight. For example, if a coreference resolution refers to a name with the wrong gendered pronoun, maybe the high importance of the crucial social error is not given to such error [Hovy and Prabhumoye, 2021]. Another source is improper and unbalanced benchmarks for evaluating applications, which can exaggerate such biases [Mehrabi et al., 2021].

3.3 Gender Bias and Language

Gender is expressed in several forms in languages, not only in written forms but also verbally. Some expressions are classified according to gender. Additionally, some adjectives, nouns, curses, and polite forms of words relate to one gender more than the other. Such bias is inherited historically in societies, and people tend to use social gender clues to assign gender to other considered people. For instance, it is a social gender that may cause an inference that my cousin is female in *My cousin is a librarian* or male in *My cousin is intelligent.*

The relation between language and gender needs elaboration on how languages understand and deal with gender. Not only do languages treat gender differently from the grammatical side but also from the cultural and historical aspects. Grammatically speaking, we can classify the languages in several categories, but here we will follow the classification scheme in [Savoldi et al., 2021, Gygax et al., 2019]. This classification shows the main difference between languages, illustrated in the rest of the thesis:

Gender-less Languages (e.g., Finnish, Turkish). In such languages, gender is only expressed at its minimum. They use them for the essential lexical pairs (*brother-sister*, *mum-dad*), but generally, they use a neutral form for pronouns, nouns, and adjectives.

Gender-less Languages with a Few Traces of Grammatical Gender (e.g., Oriya, Basque). Most personal pronouns are used for male or female referents without being distinguished linguistically. A few gendered forms appear in nouns with gender suffixes, gendered adjectives, or verbal forms.

Neutral/Notional Gender Languages (e.g., English). Such languages have a gender pronominal system. Inanimate and personal nouns do not have different genders. These languages can host some marked derivative nouns (*host/hostess*) and compounds (*businessman/businesswoman*).

Languages with a Combination of Grammatical and Neutral Gender (e.g., Norwegian, Dutch). Human nouns are not necessarily differentiated between males and females. We can use them for feminine and masculine referents without a linguistic difference. These languages have gender distinctions for inanimate nouns as well as for some personal nouns. Pronouns usually express the gender of the referent.

Highly Grammatical Gendered Languages. Gendered languages (e.g., Spanish, Italian, and French) have a gender assigned to all nouns; consequently, all articles, verbs, and adjectives have to agree with this noun. They are known to have a morphological agreement in the gender and number with the subject.

Type of Language	Languages	Example
Gender-less	Finish	**Hän** on lääkäri. (She/he is a doctor) **Hän** on sairaanhoitaja. (He/she is a nurse)
	Turkish	**O** bir tasarımcı. (He/she is a designer) **O** bir geliştirici. (She/he is a developer)
Gender-less + gender traces	Basque	**Gu** hiri laguntzera etorriko gatzaizki**k**. (We will come to you(male) to help) **Gu** hiri laguntzera etorriko gatzaizki**n**. (We will come to you(female) to help)
Gender-neutral	Danish	**Hun** er læge. (She is a doctor) **Han** er professor. (He is a professor)
	Swedish	**Hon** är revisor. (She is an accountant.) **Han** är sjuksköterska. (He is a nurse)
Grammatical + neutral gender	Dutch	De tafel met **haar** poten. **Zij** is mooi. Ik zie **haar**. (The table with its (her) legs. It (she) is beautiful. I see it (her)). De stoel met **zijn** poten. **Hij** is mooi. Ik zie **hem**. (The chair with its (his) legs. It (he) is beautiful. I see it (him))
Highly Grammatical-gendered	German	**Mein** **Freund** ist **Student**, **er** studiert. (My friend is a student, he is studying)) **Meine** **Freundin** ist **Studentin**, **sie** studiert. (My friend is a student, she is studying)
	Spanish	Mi **amiga** es **abogada**. (My friend is a lawyer) Mi **amigo** es **abogado**. (My friend is a lawyer)

Table 3.2: Different types of languages, **blue** for pronouns, **green** for gendered possessive pronouns and **red** for gendered nouns.

3.4 Gender Bias in Word Embeddings

In this subsection, we are studying the gender bias in word embeddings and contextualized ones. We include methods for assessing and evaluation in standard embeddings and contextualized embeddings from pretrained language models.

3.4.1 Assessing bias in Word Embeddings

Standard Word Embeddings. Researchers have been trying to understand the effect of gender bias on word embeddings. Many recent studies have proposed quantifying gender bias in standard word embeddings and contextualized embeddings. The first work in this field was introduced by [Bolukbasi et al., 2016], which showed that bias is inherited in word embeddings. The authors studied the gender bias in word embeddings from a geometrical perspective. They calculated the principal components of the difference between gendered female and male word pairs and computed the gender subspace. Two different kinds of biases within this gender space were recognized; direct and indirect bias. Regarding direct bias, the authors removed the gender information for neutral words by subtracting gender direction from these words' vectors. They equalized the distance of these words to both elements of gendered words. [Ethayarajh et al., 2019] suggested another bias score rule based on vector similarity; Relational Inner Product Association (RIPA). The main difference between the direct bias and RIPA is that RIPA performs normalization at the gender base pair level instead of at the word level.

[Caliskan et al., 2017] developed Word Embedding Association Test (WEAT) to measure bias by comparing two sets of target words with two sets of attribute words. He included sensitive, racial, occupation, and adjective terms in the attribute words. They concluded that Word2Vec and GloVe contain gender and racial biases.

Most studies focused on English, until [Zhou et al., 2019] extended the study to traditional Spanish embeddings only. The authors examined the gender bias in gendered languages. They introduced another direction that determines gender, known as grammatical direction, besides semantic direction. Grammatical direction determines the direction between feminine and masculine nouns in a gendered language.

Contextualized word embeddings. The contextualized embeddings helped begin the debate on performing more evaluations of gender bias within these embeddings. Multiple approaches are redesigned for such embeddings to assess the gender bias in them. [Kurita et al., 2019] proposed a probability-based method for assessing gender bias in BERT embeddings. To compute this probability, they relied on the masked language model objective using simple template sentences. Using a template, they tried some attribute words and computed the BERT probability for that sentence. The difference between the probabilities is considered the gender bias measure. [Tan and Celis, 2019, Guo and Caliskan, 2021] confirmed the same conclusion by assessing social biases in BERT contextualized embeddings and affirmed that human biases are encoded in the contextualized word models.

In [May et al., 2019], the authors generalized the WEAT [Caliskan et al., 2017] to the context of sentence encoders introducing the Sentence Encoder Association Test (SEAT). They used a sentence-measuring technique in which individual words from WEAT tests are inserted into sentence templates. Comparing different embedding techniques, they revealed that recent methods like Bert are more resistant to biases but still encode biases.

[Zhao et al., 2019] analyzed bias in ELMo, showing its sensitivity to gender unequally for female and male entities. They showed that ELMo contextualized embeddings inhibit gender bias. The effect of this bias can be delegated to downstream tasks, such as coreference resolution.

[Bartl et al., 2020] have measured gender bias concerning professions and shown that pretrained language models preserve biases from real-world data and amplify stereotypes. They presented a template-based corpus to measure the bias in English contextualized embeddings in BERT. They also evaluated the corpus in German as a grammatically gendered language. This corpus contributed to streamlining the visualization of gender bias in other contextualized word embedding models other than BERT. They showed that the same techniques for measuring bias in English models do not transfer to other languages like German.

3.4.2 Mitigating Gender Bias in Word Embeddings

Researchers have shown outstanding efforts in recent years to mitigate the bias in embeddings and prevent its convey to downstream applications. However, [Gonen and Goldberg, 2019] noted that gender bias has a more profound existence in word embeddings. Even when the embedded gender information is removed, gender information remains in the vector representation.

Standard word embeddings. As mentioned, [Bolukbasi et al., 2016] were the pioneer to propose a mitigation approach to the standard word embeddings, applied on Word2vec embeddings. [Zhao et al., 2018b] proposed an extension to GloVe embeddings, where the authors trained the embeddings with a loss function restricting the gender information to a specific portion of word embeddings. The gender information is declared as protected attributes, and once the training is done, the gender can be easily removed from the embedding vector, eliminating the gender bias.

Another approach in [Kaneko and Bollegala, 2019] proposed using particular training parameters to mitigate bias in the pre-trained embeddings. The primary purpose of these parameters is to reserve non-discriminative gender-related information while removing stereotypical discriminative gender biases from pre-trained word embeddings. Apart from parameters, the authors [Zhang et al., 2018] decided to use an adversarial network to mitigate bias in contextualized embeddings. For exploring gender in gendered languages apart from English, e.g., Spanish, [Zhou et al., 2019] proposed mitigation techniques to shift the bias along the semantic gender direction (the same direction as [Bolukbasi et al., 2016]) and an alignment technique for the gendered language with the bias-reduced English language.

Exciting work for removing bias from representations was introduced by [Ravfogel et al., 2020] which introduced an idea of iteratively training linear classifiers to predict the protected attribute aimed for removing (e.g., gender) from the representation. Then the representations are projected on the null-spaces of these classifiers. This approach shows the correlations between certain textual features and the model's predictions.

Contextualized Embeddings. Different interesting approaches were presented in multiple works for eliminating bias in contextualized embeddings. [Kaneko and Bollegala, 2021a] presented a method that can be applied to any pre-trained contextualized embedding models without retraining them. The method preserves the semantic information acquired by the model regarding the gender-related words while eliminating the gender biases in the model. The main advantage is employing this method to any internal architecture of any model on the level of tokens or sentences, enabling debiasing at various granularities and on different layers. The authors presented more ideas in [Kaneko and Bollegala, 2021b] for learning constraints to debias word embeddings using dictionary definitions, eliminating the need for predefined word lists. The dictionary would specify the debiased definition of any word. Using these definitions, an encoder is employed to generate debiased contextualized embeddings, retaining the same semantics of the pre-trained embeddings. The resulting embeddings remain orthogonal to the vector space spanned by biased basis vectors in the pre-trained word embedding space. Experimental results on standard benchmark datasets show that the proposed method accurately produces fairer pre-trained word embeddings while preserving essential semantics.

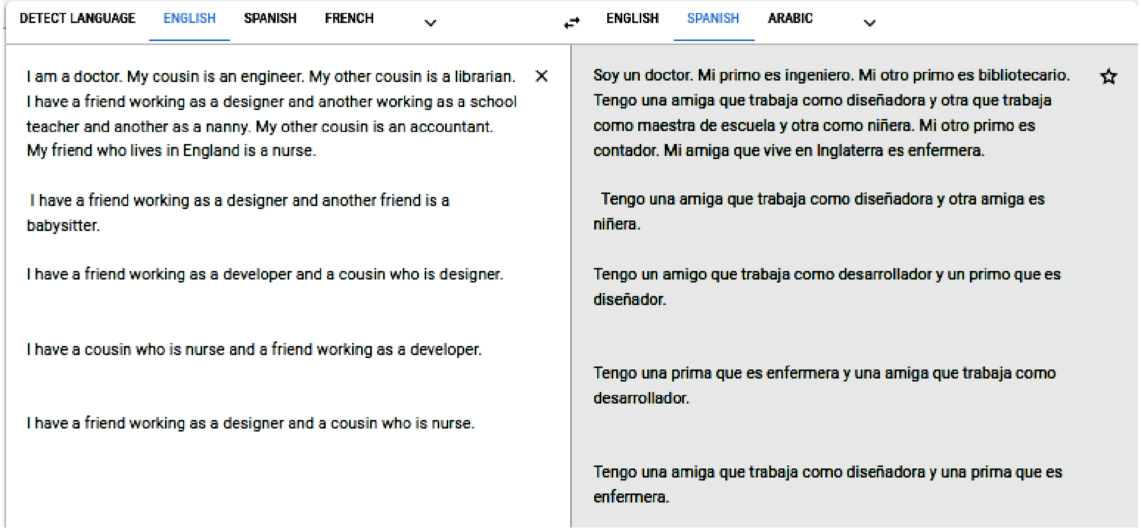
[Liang et al., 2020] employed DensRay debiasing approach on BERT. This approach is an analytical method for identifying the embedding subspace of specific linguistic features. They applied it to the attention heads and showed that the gender information is processed in all attention heads. For evaluation, the authors created a method that utilizes existing occupation datasets for assessment, relying on the templates created to evaluate contextualized language models. They also applied the Association tests for evaluation. They demonstrated that they could remove bias multilingually in English and Chinese.

In [Liu et al., 2021b], the authors proposed a framework for mitigating political bias using reinforcement learning. The rewards from a classifier or word embeddings can mainly regulate the debiased generations without retraining the system. They had two modes for computing the debias reward; the first mode is to debias the representations of the words by forcing a rule in picking the unbiased token at each step generation. This rule demands that neutral words have equal distance to groups of sensitive words. The second mode is a classifier guide debias, where they compute a classifier for the generated text to determine if the text is biased and reinforce using more neutral generated accordingly. The framework focused on three bias attributes (gender, location, and topic) and showed improvements in mitigating gender bias using direct and indirect bias metrics. The system is mainly beneficial since training large-scale LM is costly.

[Webster et al., 2020] introduced a novel analysis, DisCo, based on template and generation-based methods to discover and evaluate gender correlations in pretrained contextual representations. DisCo is built around a series of templates or sentences with empty slots. The templates have two slots, e.g., *[PERSON] studied [BLANK] at college.* They defined an evaluation framework based on classification and regression tasks, considering the model accuracy, gendered correlations in models, and the methods affecting them. Additionally, they presented dropout regularization and counterfactual data augmentation (CDA) methodologies for mitigating gender bias in pretrained language models. They showed that both techniques (together) could minimize the gender correlations, maintaining accuracy. Therefore, the techniques should help mitigate gender bias in BERT and ALBERT models. They also demonstrated that the mitigated models could resist stereotypes and gender correlations.

ADELE (Adapter-based DEbiasing of LanguagE Models) was introduced by [Lauscher et al., 2021]. This debiasing approach depends on introducing debiasing adapters. These adapters are injected into the layers of BERT and incorporate the debiasing information only in additional parameters, without changing the pretrained information. This approach proved effective in bias attenuation when the authors evaluated it on BERT and mBERT, and they seemed to outperform the debiased BERT in the previously mentioned approach [Webster et al., 2020].

3.5 Gender Bias in Machine Translation



I am a doctor. My cousin is an engineer. My other cousin is a librarian. I have a friend working as a designer and another working as a school teacher and another as a nanny. My other cousin is an accountant. My friend who lives in England is a nurse.

I have a friend working as a designer and another friend is a babysitter.

I have a friend working as a developer and a cousin who is designer.

I have a cousin who is nurse and a friend working as a developer.

I have a friend working as a designer and a cousin who is nurse.

Soy un doctor. Mi primo es ingeniero. Mi otro primo es bibliotecario. Tengo una amiga que trabaja como diseñadora y otra que trabaja como maestra de escuela y otra como niñera. Mi otro primo es contador. Mi amiga que vive en Inglaterra es enfermera.

Tengo una amiga que trabaja como diseñadora y otra amiga es niñera.

Tengo un amigo que trabaja como desarrollador y un primo que es diseñador.

Tengo una prima que es enfermera y una amiga que trabaja como desarrollador.

Tengo una amiga que trabaja como diseñadora y una prima que es enfermera.

Figure 3.1: Google translate error (February, 2022).

3.5.1 Main Challenges of Bias in NMT

A few years ago, people started to notice male dominance in translations, where masculine defaults repeatedly refer to specific nouns, despite several the presence of female pronouns in the text. Translations can exhibit a noticeable level of bias only by changing simple context, as shown in figure 3.1.

NMT is a multifaceted task resolving many gendered-related subtasks (e.g., coreference resolution, named entity recognition). Accordingly, MT inherits gender bias differently according to the language. It is hard to conceptualize gender bias similarly in all languages [Savoldi et al., 2021]. Therefore, researchers recommend that studying the related literature outside of NLP can highly impact the advancement of the field [Blodgett et al., 2020].

NMT is considered one of the most challenging tasks in mitigating gender biases. There are many challenges and sources of biases other more amplified in the task, besides the sources already mentioned in section 3.2:

Translation from Languages with Different Natures. Translation from gender-less to morphologically gendered languages forces the model to attend to contextual clues. The context is not always enough for the system to predict, added to problems that may arise from errors in coreference resolution, name entity recognition, or any sub-related task. Another issue is the abundance of English in most NLP tools, while there are few resources for other languages. The under-exposure induces challenges in different languages; thus, the NLP tasks underperform for them.

Some common problems occur due to the different nature of languages, as listed:

1. **Gender Agreement** In general, gender agreement is the main feature of gendered language and varies from one language to another. Gender agreement is the agreement between the gender of the noun, the correlated pronouns, the correlated numbers, and the correlated adjectives. Some languages require all these agreements, and some require only a subset. An example of gender agreement, *Elle est heureuse* (fr), which is translated to *She is happy* (en). *happy* is neutral and goes for males and females while *heureuse* is for females and *heureux* is for males.

2. **Neutral Possessives** These are the possessive pronouns that are expressed in neutral sound. One needs context when translating them from this form to a gendered form according to the language. German is a popular language for this case as it has female, male, and neutral pronouns to every possessive pronoun, making the translation from or to German quite challenging. An example is shown in Table 3.2, where the German language has *Meine* a posessive pronoun refers to a female noun and *Mein* refers to a male noun.

3. **Dropped Pronouns** Some languages drop the pronouns when speaking about a second person and depend on the context to understand. This is the case in Spanish.
   For example, *Trabaja en Barcelona* (es) is translated to *(He/She) works in Barcelona* (en). *Trabaja (works)* is not accompanied by any gendered pronoun, depending on the context and the referenced person from the beginning. Therefore, if this sentence stands alone without more context, its translation would be probable for both genders.

4. **Stereotyped-inflection** Stereotyped inflection is usually illustrated with the famous Google Translate example of translating the gender-neutral Turkish source sentence *o bir doktor*, which currently produces two different gendered target sentences: *She is a doctor (feminine)* and *He is a doctor (masculine)* [Tomalin et al., 2021]. The problem mainly happens when the inflected languages translate the pronouns and nouns to their stereotyped inflection reflecting the data biases and imbalances. The non-binary inflection is still an issue not addressed and lack accepted conventions [Ackerman, 2019]. There has been recent criticism of stereotyping in translations; as a result, researchers and industry have been dedicating special efforts to solving the issue. However, the efforts still fail regarding the neutrality of the source sentences.

3.5.2 Assessing Approaches

Evaluation is always the primary step towards understanding the nature of any problem. Therefore, researchers always dedicate efforts to understand and evaluate the problem and assess the reasons leading to it. Several efforts have been working towards assessing and quantifying the gender bias in NMT, and in the next section, we show the assessing approaches in the related literature.

Several approaches have been trying to assess gender bias in different systems to understand when and why it occurs. [Prates et al., 2020] have investigated the pronoun translations widely from 12 gender-less languages to English. The authors have built simple sentences using the U.S. Bureau of Labor Statistics occupations, e.g., *he/she is an engineer*. Moreover, they have focused on 22 macro-categories of occupation to see if the proportion of pronouns translated corresponds to the real-world population of gender employment. Their main observation was that Google translate tends to use masculine defaults much more frequently than the expectations of the demographic statistics. They suggested that 50:50 pronominal predictions are unrealistic, but the proportions of the predicted ones are still far from reality. The current MT systems tend to underestimate the feminine proportion more than realistic occupations data suggest.

The same conclusion has been demonstrated by [Stanovsky et al., 2019]. The authors assessed multiple industrial MT systems such as Google Translate, Microsoft Translator, Amazon Translate, and SYSTRAN. They tested the systems using a challenge test set WinoMT, which is the collection of WinoBias [Zhao et al., 2018a] and WinoGender [Rudinger et al., 2018] illustrated in details in Section 3.6. They have demonstrated the prevalence of gender bias across multiple NMT systems. The fact that the pronouns and possessives are frequently translated with the masculine defaults in the translations proves that the systems reinforce sexist tendencies in society.

[Hovy et al., 2020] studied whether there is gender stylistic bias in MT. The authors translated a corpus of online reviews [Hovy, 2015] and compared the predicted gender and age in translation to the original demographic information of the reviews. They demonstrated obvious gender stylistic bias, and the MT commercial systems make authors "sound" older and male. Such results show the model's unfamiliarity with neither as many female writings as males nor different age segments.

[Vanmassenhove et al., 2021] have done a wide range of experiments with three different MT architectures and nine metrics, among which they measure lexical frequency profile, synonym frequency analysis, and other measures. They showed that the original data usually has more lexical and morphological diversity than MT translations. They concluded that the inappropriate stereotypes are currently built-in in the systems, and the systems cannot even warn the users about the assumptions made.

3.5.3 Mitigation Approaches

Several directions are implemented for mitigating bias in NMT and we classify them to the following categories that cover most of the recent work:

Gender Tagging and Additional Context. [Vanmassenhove et al., 2018] was one of the first to suggest that gender tagging can help the problem. The authors train the model from scratch by augmenting Europarl data by appending a gender tag to the source sentence (M for male speakers and F for female speakers). Such a limited setting proved to be effective even though the concept of adding gender per sentence is not appropriate for all kinds of datasets and all translations. The authors proved an enhancement in translation due to adding control over hypothesis gender when translating from sentences that do not contain explicit gender for the speaker.

[Moryossef et al., 2019] prepend a short phrase at inference time representing an explicit gender feature for each sentence. They added a constraint of applying this approach to the text with one gender identity per sentence.

[Stafanovičs et al., 2020] and [Saunders et al., 2020] explored the use of the word-level gender tags. In the former approach, the authors train their NMT models with all source language words annotated with the target language grammatical gender. In the latter approach, they add a tag after the entities in source with either male tag <M>, or female <F> or non-binary <N>.

Domain Adaptation. Domain adaptation techniques have proved to impact the performance of translation in [Costa-jussà and de Jorge, 2020] and [Saunders and Byrne, 2020]. [Saunders and Byrne, 2020] show that finetuning the system on small gender-balanced data can improve gender translation accuracy. The main problem is that this gender-balanced data is counterfactual. [Saunders and Byrne, 2020] have built their gender-balanced dataset of English sentences following this schema: *The <profession> finished <his|her> work.* Then they automatically translated these sentences and manually checked to obtain the corpus. This follows the counter-factual augmentation approach [Lu et al., 2020] to create identical sentences with feminine/masculine forms. Consequently, they finetuned the NMT system on this new corpus using Elastic Weight Consolidation (EWC), a technique used for minimizing the forgetting during model adaptation

[Kirkpatrick et al., 2017]. Using a lattice rescoring module, they also proposed a post-processing technique for rescoring all translation alternatives for gendered marking words. After rescoring, the output would be the sentence with the highest score. The authors showed that the domain-adaptation could generate less biased translations without surrendering the overall system accuracy. This work in [Saunders et al., 2020] extended this model adaptation on gender-tagged data for controllable gender inflection and the assessment of the gender-neutral infections in grammatically gendered target languages as well.

[Costa-jussà and de Jorge, 2020] have finetuned the NMT models on natural gender-balanced dataset extracted from Wikipedia [Costa-jussà et al., 2020] instead of a synthetic one. They first analyzed the balanced dataset to show that this set only encodes much less bias than other datasets. Then they finetuned the standard MT system with this dataset and showed that this mitigated the gender bias.

[Tan et al., 2020] have exposed NMT model to inflectional adversaries; morphologically varied input. Afterward, they finetuned it on a representative adversarial training set. The system showed significant robustness to inflectional adversaries while preserving performance on the clean dataset.

Debiased Word Embeddings. The authors in [Font and Costa-jussà, 2019] have leveraged debiased pre-trained word embeddings to remove the gender associations from the representations of English gender-neutral words. The authors studied the bias in the translations using a custom test set composed of occupations. This set consists of sentences that include the ambiguous *friend* word added to the additional context in the English-Spanish translation. The word can then be translated to feminine or masculine, depending on the context. Using this methodolgy, an impact is shown on gender accuracy and BLEU scores in English-Spanish data.

Data Balancing. Given the imbalances in the training data leads to gender bias problems [Costa-jussà, 2019]. Researchers have worked toward balancing data using multiple techniques. We are discussing the different approaches for balancing the data. However, researchers argue that balancing data introduces another kind of bias in the

Researchers argue that balancing data may introduce new biases. The data-set has complicated sorts of biases, and removing the simple apparent ones can change the data statistics and introduce new kinds of biases [Hovy et al., 2020]. From the methodological point of view, removing particular kinds of bias in AI systems is not feasible and desirable.

1. **Counterfactual techniques** Counterfactual augmentation techniques create a counterfactual sentence for the opposite gender. This technique can augment the dataset with sentences for both genders, compensating for the under-represented gendered categories. It adds data with equivalent sentences of the swapped gendered version. For example, the counterfactual sentence of *she is a nurse, and I talked with her* would be *he is a nurse, and I talked with him.* As evident, this augmentation technique works better for more straightforward sentences without many coreferences. Another challenge is the large bilingual corpora required to train NMT models. Creating counterfactual augmented translation of the source sentence does not always apply for high inflected languages in the same way. It is because translating the counterfactual sentence may not be accurate. It is sometimes impossible to achieve such balancing in practice. Consequently, a balanced dataset with equal numbers of gendered entities is only balanced relative to that dichotomy [Tomalin et al., 2021]. Generally, the gender-swapping technique can be simple for English sentences, but this is not the case in grammatical languages with marked morphological articles, adjectives, verbs, and nouns. For example, languages like German require primarily applying the parts-of-speech identification task to determine which parts of the sentence should be updated.

   [Tomalin et al., 2021] have proposed three automated strategies for removing bias of data; downsampling, upsampling, and counterfactual augmentation. The main drawback was that all systems trained on these debiased datasets had declined overall translation performance than the baseline. These datasets were preferable when finetuning existing trained NMT models instead of training them from scratch. In the latter case, the systems could produce competitive translations with improved gender accuracy.

   Creating the counterfactual swapped sentence is not straightforward for highly inflected languages like German, as mentioned above. [Zmigrod et al., 2019] have introduced a Markov random field scheme for such infected rich morphological languages. This schema infers which parts of the sentence need change to preserve the syntactic gender agreement when changing the nouns.

The main problem is that the coreference information necessary to cover more than one entity is not included [Tomalin et al., 2021], making the scheme not applicable to real-life data or even WinoMT.

2. **Gender re-inflection** The problem of gender agreement in Arabic is familiar, [Habash et al., 2019] and [Alhafni et al., 2020] have proposed a preprocessing solution to face this problem. The preprocessing solution reinflects the personal references, only first-person, into male/female forms.

   [Habash et al., 2019] preprocessing process is done in two steps. The first is reinflecting the first-person references in MT output after identifying them. The second is reinflecting both forms of gender (male and female) from the MT output. The method does not require gender information of the speaker. In [Alhafni et al., 2020], the desired gender in the reinflection is needed besides the MT translation to be fed to the preprocessing stage.

Although researchers have made significant efforts to mitigate and assess MT bias, there is still no SOTA method for mitigating bias. There is no explored work on integrating all these efforts in the current MT systems. Gender bias is a challenging problem in NMT, and there is no current real solution to tackle different problems. All solutions tackle an aspect or two, depending on many factors, including the conceptualization, corpora, the targeted languages, the generalization, and context-aware interference.

3.6 Evaluation Benchmarks

Monolingual Evaluation Benchmarks These monolingual evaluation benchmarks, shown in Table 3.3, are mostly made up of syntactic examples, each with a specific phenomenon or measurement criteria for certain features. These benchmarks are primarily valuable for language modeling and its related tasks mainly. Sometimes, it is not feasible to identify if the model propagates stereotypical representations using these challenge sets.

WinoBias-WinoGender. [Rudinger et al., 2018] and [Zhao et al., 2018a] created counterfactual augmented data-sets and showed improvement in coreference resolu-

Dataset	Types of bias	Size
WinoBias	Gender	3,160
WinoGender	Gender	720
Stereoset	Gender, race, religion, and profession	17,000
CrowS-Pairs	Nine types of bias (gender, religion, race, age, ..)	1,508
GAP	Gender	8,908

Table 3.3: Monolingual different benchmarks.

tion applications when trained on these data-sets. Both corpora contain Winograd-schema [Rahman and Ng, 2012] style sentences with entities corresponding to people referred by their profession (e.g., nurse, developer, doctor, carpenter). The sets have pro-stereotypical sentences and anti-stereotypical sentences. A pro-stereotypical sentence has an explicit gender word (e.g., *he*, *she*) that corefers to a socially stereotyped gendered occupation. On the other hand, anti-stereotypical sentences oppose the gender word with the stereotype of the occupation, e.g., the doctor with female pronouns and the nurse with male ones. An example in WinoBias for an anti-stereotypical sentence is, *The doctor hired the secretary because she was overwhelmed.*, where *she* stands for the doctor and is replaced by *he* in a stereotypical sentence. An example for WinoGender is *The nurse notified the patient that her/his/their shift would be ending in an hour.*, *his* stands for the nurse and is replaced by *her* in a stereotypical sentence and *their* in a neutral sentence.

Stereoset (SS). A large-scale natural English dataset [Nadeem et al., 2021] to measure stereotypical biases in four domains: gender, profession, race, and religion. SS contains test instances both at intra-sentence and inter-sentence discourse levels. The authors claim that they contradict stereotypical bias and language modeling ability of popular models like RoBERTa, BERT, XLnet, and GPT-2. The authors showed that these models manifest stereotypical bias. Each intra-sentence should contain three possible associations for the masked token in the sentence, one being stereotypical, another being anti-stereotypical, and a third being unrelated. They proposed a score (stereotype score) of a model. This score computes the percentage of the stereotypical examples assigned a higher score to the anti-stereotypical ones. Compared to stereotypical and anti-stereotypical associations, they did the same with unrelated association examples. This score is called ‘the language modeling score’. Examples of the sentences are a stereotypical one *our housekeeper is Mexican*, an anti-stereotypical one *our housekeeper is American* and a non-related one *our housekeeper is computer.*

CrowS-Pairs (CP). Crowdsourced Stereotype Pairs benchmark (CrowS-Pairs) [Nangia et al., 2020] is a challenging dataset targeting certain protected groups in the USA. It covers stereotypes dealing with nine types of bias, like gender, religion, race, and age. Test instances in CP include sentence pairs; one is more stereotypical than the other. Annotators were responsible for creating examples that demonstrate stereotypes contrasting historically disadvantaged groups against advantaged groups. With a crowdsourced validation task, CP has test instances more reliable than the ones in SS. In CP, the likelihood probability is computed to estimate the bias of the LM to select a stereotypical sentence over a less one. They evaluated the likelihood of the stereotypical and anti-stereotypical sentences and the percentage of stereotypical examples assigned higher likelihood than the anti-stereotypical sentence. When masking tokens from the test sentences and predicting them, the systems favor advantaged groups more frequently than the disadvantaged ones in the corpora used to train the LM model. Examples of the pair sentences are a stereotypical one *people who live in trailer parks are alcoholics*, then the paired sentence would be *people who live in mansions are alcoholics*, which is the anti-stereotypical version of the first sentence.

GAP. [Webster et al., 2018] produced a gender-balanced dataset containing coreference-labeled pairs collected from Wikipedia to assess coreference resolution in practical applications. The corpus has human-annotated ambiguous pronoun-name examples, filtered through a multistage process devised to enhance quality and diversity. These examples are also attentive to the well-known gender biases. An example is *The disease is named after **Eduard Heinrich Henoch** (1820–1910), a German pediatrician (nephew of Moritz Heinrich Romberg) and **his** teacher*, where *his* is an inference sample involving the entity topicality with parenthesis, adding uncertainty to the resolution.

Multilingual Evaluation Benchmarks For gender bias analysis, NMT datasets are the only ones available in bilingual/multilingual settings due to the nature of the task [Stanczak and Augenstein, 2021]. Synthetic and natural datasets are available. However, the natural corpora are usually preferred as they quantify the actual female representations in MT in real-life scenarios and account for gender bias in natural conditions. However, benchmarks remain valuable to evaluate and monitor the model's performance and provide insights into how the model treats gender-related issues. Examples of each benchmark are given in Table 3.4.

Dataset	Size	Example
WinoMT	3,888	The *developer* argued with the designer because *she* did not like the design. (Anti) The *developer* argued with the designer because *he* did not like the design. (Pro)
SimpleGEN	1,332	That *engineer* is a funny *guy*! (Pro-MoMc) That *nanny* is a funny *lady*! (Pro-FoFc) That *mechanic* is my funny *woman*! (Anti-MoFc) My *brother* is a *nanny*. (Anti-FoMc)
Unamiguous Set	1850	My *sister* is a *carpenter* . My *nurse* is a good *father*.
Arabic Inflected	2,448	أنا طبيب(I am a male doctor) أنا طبيبة (I am a female doctor)
MuST-SHE	1,062 (En-It) 1,074 (En-Fr)	Sono nata e cresciuta a Mumbai.(I was born and brought up in Mumbai.) Je suis neé et j'ai grandi a Mumbai. (I was born and brought up in Mumbai.)
GeBioCorpus	2,000	Bridegroom was an actor and songwriter... (En) Bridegroom era un actor y compositor ...(Es) Bridegroom era un actor i compositor ...(Ca)
Google Gender set	1,471(En-Es) 1,471(En-De)	Su intento fue en vano. (Her struggle proved unsuccessful) Doch ihre Bemühungen blieben erfolglos. (Her struggle proved unsuccessful)
ReflexiveChange	4,560	The firefighter placed her/his shoes in the closet. (En-Source) Brandmanden placerede hendes sko i skabet (FEM-Danish) Brandmanden placerede hans sko i skabet (MASC) Brandmanden placerede sine sko i skabet (REFL)
BUG	108K	Hiei's captain ordered her crew to abandon ship after further damage. (The anticedent pronoun is ambiguous.)

Table 3.4: Different benchmarks for NMT, pro stands for pro-stereotypical, anti stands for anti-stereotypical. FEM stands for Feminine, MASC stands for masculine and REFL stands for reflexive in case of Reflexive Change dataset.

WinoMT. [Stanovsky et al., 2019] is the first challenge test set for evaluating gender bias in MT systems for translating from English to multiple languages. This test set is a combination of previous mentioned Winogender [Zhao et al., 2018a] and WinoBias [Rudinger et al., 2018] sets, consisting of 3,888 sentences of 1,584 anti-stereotyped sentences, 1,584 pro-stereotyped sentences, and 720 neutral sentences. Each sentence contains two personal entities, where one entity is a co-referent to a pronoun, and a golden gender is specified for this entity. An example of an anti-stereotypical sentence is demonstrated in the figure below, where *her* refers to the *doctor*. The translation tended to stereotype the professions, giving the 'doctor' male gender and the 'nurse' the female gender. The evaluation mainly depends on comparing the translated entity with the specified gender of the golden entity to correctly gendered translation. Three metrics were used for assessment: accuracy (Acc.), ΔG and ΔS. The accuracy is the correctly inflected entities compared to their original golden gender. ΔG is the difference between the correctly inflected masculine and feminine entities. ΔS is the difference between the inflected genders of the pro-stereotyped and anti-stereotyped entities.

SimpleGEN. This dataset [Renduchintala et al., 2021] focuses on two language pairs, English to Spanish (En-Es) and English to German (En-De). While the target as a gendered marking language gives gender to the occupation nouns, the source (English) lacks this phenomenon, which forces the NMT system to attend to contextual clues. The main template for constructing the set has enough contextual evidence to specify the gender of the occupation noun. Therefore, this test set has unambiguous occupation nouns. The English sentences are pro-stereotypical (pro) and anti-stereotypical (anti) types. The difference is that the pro sentences have female occupations in the female context (FOFC) and the male professions in the male context (MOMC). In contrast, the anti has female professions in the male context (FOMC) and male professions in the female context (MOFC). The male context and female context refer to the existence of unambiguous signals that the occupation noun corresponds to a male or a female person, respectively. The set contains 1,332 pro and anti sentences, 814 in the MOMC and MOFC subgroups, and 518 in the FOMC and FOFC subgroups.

For the sake of translation evaluation, the authors create an occupation-noun bilingual dictionary with synonyms of the same profession. For example (*medico, medica*) and (*doctor, doctora*) for masculine and feminine forms of the word English *physician* in Spanish.

Unambiguous Gendered Challenge Set. [Renduchintala and Williams, 2021] composed a set of English sentences to study the syntactical agreement in gender bias. The sentences are constructed such that occupation is related to its unequivocal gender trigger, e.g., *My nurse is a good father*. They designed the sentence to correctly study the translation of gender morphology in unambiguous contexts across syntactically diverse sentences. They concluded that NMT struggles to correctly predict the translation of unambiguous occupations, even in simple settings. This dataset translates from an English source into 20 languages from several different language families.

Arabic Parallel Gender Dataset. [Habash et al., 2019] have constructed the English-Arabic dataset containing 2,448 sentences, all of which have a first person singular reference to the speaker. The corpus was formed from "OpenSubtitles" natural language data [Lison and Tiedemann, 2016]. The corpus would contain words that are gendered ambiguous and can be translated to the two forms like *I'm leaving* which can be translated to أنا راحل (male form) or أنا راحلة'female form). Such sentences would contain verbs, adjectives, or nouns, which can be reinflected in both female and male genders. In this case, these sentences were translated to both gender forms, obtaining equal number of genderly annotated sentence pairs, qualitatively and quantitatively. The creation of the corpus needed extensive manual work, which makes it a beneficial resource for gender-marked natural language, as it allows for cross-gender evaluations on MT translations of the speaker's gender.

MuST-SHE Dataset. MuST-SHE [Bentivogli et al., 2020] is an exciting set allowing evaluation for MT and Speech Translation (ST) for English-French, English-Italian, and English-Spanish language pairs. It is built on Ted-talks data [Cattoni et al., 2021] with gender-balanced samples. Each dataset pair has triplets of information(audio, transcript, and translation). The existence of the source and target translations made it feasible to measure gender accuracy in BLEU added to other gender metrics. The dataset has two types of data: sentences for first-person speakers and sentences with contextual information to disambiguate gender.

GeBioCorpus. GeBioCorpus [Costa-jussà et al., 2020] is a gender-balanced set from Wikipedia biographies (GeBioCorpus) and contains 1000 sentences from male biographies and 1000 sentences from female biographies for English-Catalan and English-Spanish. It is a gender-balanced dataset with the same considered number of documents of males and females. It has the advantage of being natural data collected from biographies.

Google Gender Challenge Dataset. This dataset [Stella, 2021] is a natural one, collected from Wikipedia. English-Spanish and English-German datasets are extracted to be gendered balanced, with diversity in occupations and nationalities. Three problems were addressed in choosing the samples; pronouns dropping, neutral possessive pronouns, and gender agreement. The authors provide the dataset with extra information; the Wikipedia links and the gender of the entity of the intended biography. Such information can be helpful for different tasks other than the MT task.

Reflexive Challenge Dataset. [González et al., 2020] focused on gender-related translation errors resulting from the syntactic structure and unambiguous coreference. The authors considered these types of mistakes unforgivable as there is no real reason for ambiguity. The authors investigated non-English languages (Swedish, Russian, Chinese, and Danish). They chose these languages because their anti-reflexive possessive pronouns are gendered, but reflexives are not. The examples quantify how systems amplify gender bias in predicting pronouns with unwarranted disambiguation. For NMT, the templates of the sentences focus on the gendered pronoun's resolution in the source language, as in the example *The doctor put the book on her table.*

BUG Dataset. The authors in [Levy et al., 2021] created a large dataset of challenging grammatical patterns indicating stereotypical and non-stereotypical gender-role assignments in corpora from three domains in this work. This resulted in creating a large-scale gender bias dataset with real-world examples to translate from English to multiple languages. They examined the dataset in two downstream applications; coreference resolution and machine translation models. Both applications were shown to encode gender bias and rely much more on stereotypical examples. They showed that this set could be used for finetuning a coreference resolution model and mitigating gender bias in it.

4 Evaluation Gender Bias in Embeddings and NMT architectures

Evaluation is a critical step towards understanding the performance of a system and measuring any improvement occurring to it, specifically here for measuring gender bias. We would never know about the problem until [Bolukbasi et al., 2016] revealed that the bias is encrypted in the encodings of the embeddings. This chapter is divided into two parts; evaluation of gender bias in contextualized embeddings and evaluating how multilingual NMT architectures treat gender translation and delegate gender bias[1].

4.1 Evaluation of Gender Bias in Contextualized Embeddings

4.1.1 Motivation

Gender bias is amplified through NLP tasks; we can identify its effect in language modeling, NMT, etc. The main component of all these tasks is the word embeddings. If the word embeddings encode gender bias, it will propagate through applications and exacerbate the bias problem in downstream tasks.

Besides, the evolution of contextualized embeddings and their application in many NLP tasks raised many questions regarding encoding different biases in them. This

[1] [Basta et al., 2019, Basta et al., 2021, Costa-jussà et al., 2022]

field opened exhaustive research on how to evaluate and thus mitigate bias in embeddings. Understanding how gendered words are represented, how the bias is encrypted, and the relation between the embeddings and the gender bias issue can better guide us in understanding the bias in the NLP cycle. This study of contextual word embeddings to assess bias is essential in considering the fair context in the NLP system.

Asking the right questions towards evaluating the bias in such embeddings can help us demonstrate more understanding of many perspectives of the bias problem. Moreover, the essential part of such evaluation is that it can help us conceptualize the biased definitions and concepts in word embeddings. The lack of bias definitions has always been challenging to tackle, and investigating different research questions in analyzing and evaluating can lead to better conceptualizations.

4.1.2 Research Questions

Contextualized word embeddings still exhibit gender bias [Zhao et al., 2019]. To begin our evaluation study on contextualized embeddings, we focused on the following questions, these question will be addressed in sections 4.1.6 and 4.1.7:

1. Does the effect of gender bias propagate from the corpus level to the contextualized word-level across different domains?
2. Is gender bias more represented in the contextualized word embeddings of professions?
3. What evaluation measures can be easily applied to gendered languages such as Spanish?
4. Can we rely on particular measures of evaluation more than others?

4.1.3 Experimental Framework

For our study to understand the influence of bias in contextualized embeddings, experiments were performed with two languages, English as a neutral gendered language and Spanish as a high gendered language. Spanish is considered a highly-gendered morphological language compared to English, where the professions and adjectives also have gender associations. For example, *I am a nurse* is translated to *soy enfermero* for a male speaker, and *soy enfermera* for a female speaker.

In order to evaluate and quantify the presence of bias in contextualized word embeddings, we apply the established methodologies for classical word embeddings by [Bolukbasi et al., 2016], [Zhao et al., 2018b] and [Gonen and Goldberg, 2019], reformulating them appropriately for contextual representations. They rely on direct intrinsic measures based on probes on different gender-predicting tasks. We focus on intrinsic measures, as opposed to other bias detection on extrinsic measures, where WinoMT [Stanovsky et al., 2019] is the main representative example. However, given the tight coupling of the test to the downstream task (i.e., MT) implies multiple problems: the impact of pre-trained debiased embeddings in the resulting translations cannot be measured in isolation; apart from that, pre-trained embeddings are seldom used in MT due to the importance of learning them along with the task; furthermore, the word-level token granularity in our contextual word embeddings is not appropriate for neural MT systems, where sub-word token granularity is needed to achieve good translation quality. Therefore, we understand that intrinsic measures, like those under study in this work, are the most appropriate testing framework for learned word-level representations like ELMo's contextualized word embeddings.

Contextualized word embeddings toolkit. Contextualized representations have proven to be essential in improving the results compared to non-contextual representations (i.e., classical word embeddings) on a wide range of tasks. Among the different contextualized representation learning approaches, tokenization is a differential factor. Some approaches, like BERT [Devlin et al., 2019], use sub-word level tokens. This makes the association between word-level information, like semantics and tokens, hard to establish. On the other hand, contextualized word representation learning, like ELMo [Peters et al., 2018], enables connecting these word-level representations with their semantic traits (e.g., gender) and reasoning about such a connection. That is why we have chosen ELMo representations over BERT or other sub-word level ones.

ELMo produces multiple forms of word embeddings for every single word, which is different from traditional word embeddings. ELMo produces three layers of word embeddings for a single word. The higher layers capture different context-dependent aspects of word embeddings, and the lower-level layers capture syntax-dependent aspects. We can use only one layer of the embeddings or the concatenation of all layers to obtain the benefits of the different representations of the layers. After experimenting with different representations from the three layers, there was not much variance in performance between layers. No distinguished information was explained by trying the evaluation measures on the three different layers. Therefore, we demonstrate the results of the representations of the third layer, which is more related to the context and its semantics, and the concatenation layer, which concatenates the representations of the three layers. The ELMo embeddings for Spanish were computed with the corresponding library[2].

Domain	**TEDx**	**WMT**	**PubMed**	**EuroParl**
No. lines in direct bias	2,,894	2866	6,507	19,821
No. lines of corpus	157,895	174,441	287,811	1,965,734
Total professions in KNN	144	114	68	142
Female in KNN%	48.61	42,98	54.41	45.07
No. of biased words (Cluster)	700	640	409	740
Females in biased clustering%	51.29	42,18	53.79	46.08
No. of biased words (Classify)	3,637	3,447	2,223	3,923
Females in biased classify%	49.46	41.77	53.08	45.73

Table 4.1: Domain-specific data summary figures.

4.1.4 Experiments on the English Language

Data and Lists. We selected four domains to explore diversity's effect on contextualized representations. We chose a medical domain (Pubmed[3]), a political domain (Europarl[4]), a social domain (TEDx[5]), and a news domain (WMT[6]). From the statistics in Table 4.1, we can observe a difference in the size of each corpus, the number of existing professions, and the number of existing biased words. The TEDx and WMT corpora are more general and smaller in size, while Pubmed and EuroParl are more specific domains and larger concerning the size.

We found from the statistics that each domain considers certain professions. For example, the most dominant professions in TEDx are *student* and *teacher*, which

[2] https://github.com/HIT-SCIR/ELMoForManyLangs
[3] https://github.com/biomedicaltranslationcorpora/corpora
[4] http://opus.nlpl.eu/Europarl.php
[5] http://opus.nlpl.eu/TED2013.php
[6] http://www.statmt.org/wmt13/translation-task.html

appear 160 and 153 times, respectively, while those that appear the least or only once are *mechanic, waitress, receptionist* and *firefighter*. For those professions that appear once, the gender of their appearance will surely prevail. Therefore, the cluster and classification will treat them with the gender of their appearance. One interesting fact about TEDx is that it contains a wide diversity of professions. Europarl has evident examples of the diversity of the domain where *citizen* appears 2408 times, *advocate* occurs 1157 times, *judge* occurs 1071 times, and *minister* and *president* appear 1038 and 841, respectively. The large occurrence of certain political professions influences the computation of any measure having such a profession. As expected, in the Pubmed corpus, the highest occurring professions are *physician, nurse* and *doctor*.

To perform our English analysis, we used a set of lists from previous works [Bolukbasi et al., 2016, Gonen and Goldberg, 2019]. We refer to the list of definitional pairs[7] as the 'definitional list' (e.g., *she-he*, *girl-boy*). We refer to the list of all of the definitional pairs added to other gendered words (e.g., *lady-gentleman*, *niece-nephew*) [8] as the 'equivalent list'. We refer to the list of female and male professions [9] as the 'professional list' (e.g., *accountant*, *surgeon*). The 'biased list' is the list used in the clustering experiment, and it consists of biased male and female words (500 female-biased tokens and 500 male-biased tokens). This list is generated by taking the most biased words, where the bias of a word is computed by taking its projection on the gender direction ($\overrightarrow{he}$-$\overrightarrow{she}$) (e.g., *breastfeeding*, *bridal* and *diet* for female and *hero*, *cigar* and *teammates* for male). The 'extended biased list' is the list used in the classification experiment and contains 5000 male and female-biased tokens, 2500 for each gender, generated in the same way as the biased list[10]. The lists we used in our experiments were obtained from [Bolukbasi et al., 2016] and [Gonen and Goldberg, 2019]. However, since we used words in sentences, our corpora may not contain examples of all the words in the lists, preventing us from obtaining their contextualized embeddings.

Evaluation Measures. The English experiments are considered an extensive study for evaluating the gender bias in contextualized embeddings. The study was done in different domains. For each domain, five experiments were conducted to understand

[7] https://github.com/tolga-b/debiaswe/blob/master/data/definitional_pairs.json

[8] https://github.com/tolga-b/debiaswe/blob/master/data/equalize_pairs.json

[9] https://github.com/tolga-b/debiaswe/blob/master/data/professions.json

[10] Both the 'biased list' and 'extended biased list' were kindly provided by Hila Gonen to reproduce experiments from her study [Gonen and Goldberg, 2019]

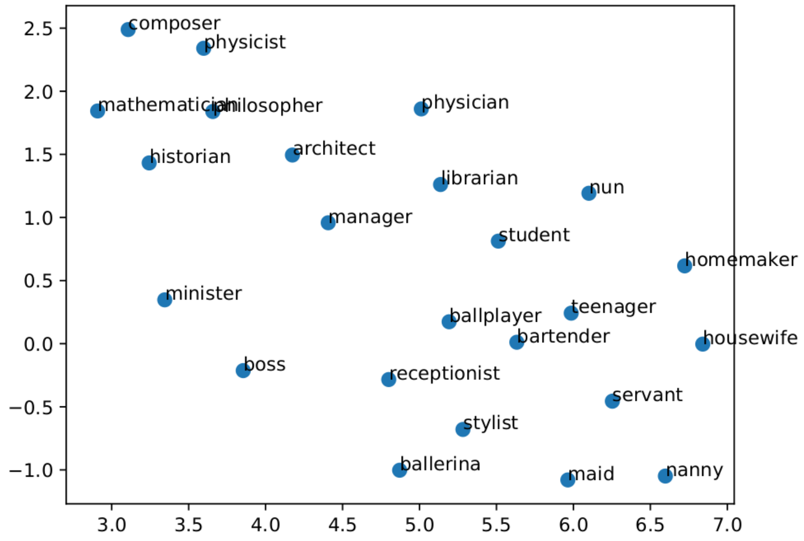


Figure 4.1: TEDx annotations of the professions.

more about the bias in each perspective. As mentioned, in our analysis, we used a set of lists from previous works [Bolukbasi et al., 2016, Gonen and Goldberg, 2019]. According to these works, [Bolukbasi et al., 2016] have referred that gender bias can be detected when one can determine the gender of non-explicitly gendered words by looking at its projection on gendered pair in the definitional list. [Gonen and Goldberg, 2019] have demonstrated that there is still gender bias when these non-explicitly gendered words directly relate to gendered or biased words.

The measures used in the first two experiments are: determining the gender direction and computing the direct bias between the profession's neutral words and this direction. The other three measures reformulated from [Gonen and Goldberg, 2019] measures to study if bias is deeply encoded in embeddings.

Word embedding association test (WEAT), the most common association test for word embeddings, has been proven to overestimate bias systematically [Ethayarajh et al., 2019]. Additionally, the work in [Kurita et al., 2019] has implied that WEAT can not be considered as a useful measure for bias in contextual embeddings. Additionally, WEAT was used on sentence embeddings in [May et al., 2019] of ELMo and BERT, but no evidence of bias was found. These conclusions guided us not to adopt WEAT in our experiments.

The main experiments carried out in our evaluation are illustrated as follows:

- **Detecting gender direction (Exp.1):** To compute the gender subspace, we followed the state-of-the-art method in [Bolukbasi et al., 2016] in a manner suitable for the contextualized embeddings. For a given corpus, we generated the corresponding gender-swapped variants, for sentences that had any instance of equivalent pairs in the equivalent list (changing *he* to *she* and vice-versa, *business-man* to *business-woman* and vice-versa, etc.). Thus, we had a sentence pair, each with a different gender for the definitional word.

  To compute the gender subspace, the representations of words were selected from randomly sampled sentences that contained words from the definitional list. We then obtained the ELMo representations of the definitional word in each sentence pair and computed their difference. On the set of difference vectors, we computed their ten principal components using Principal Components Analysis (PCA) to get the gender direction and its value from the top component.

- **Direct bias computation (Exp.2)**: Direct bias measures how close a specific set of words are to the gender vector. To compute it, we extracted the sentences that contained professional words in the professional list from the training data. We excluded the sentences with both a professional token and a definitional gender word to avoid the latter's influence over the presence of bias in the former, e.g., *he was my doctor*. Sentences with other equivalent words from the equivalent list, which are not definitional, were excluded, e.g., *I listened to the congressman*. We applied the definition of direct bias (see equation 1) from [Bolukbasi et al., 2016] to the ELMo representations of the professional words in these sentences.

  $$\frac{1}{|N|}\sum_{w \epsilon N} |cos(\vec{w}, g)| \tag{4.1}$$

  where N is the amount of gender neutral words, g the gender direction, and $\vec{w}$ is the word vector of each profession. In our case, N is the number of sentences with professional words.

For the next experiments, one representation for each word was considered, to avoid dealing with a word as male-biased and female-biased simultaneously.

- **Male and female biased clustering approach (Exp.3)**: To study how biased male and female words from biased lists cluster together when applying contextualized embeddings, we used k-means to generate two clusters of the token embeddings from the biased list. Then we computed the accuracy of clustering with the original biased version as a measure of bias. The higher the accuracy was, the more the clusters aligned with gender.

- **Classification approach (Exp.4)**: To study if contextualized embeddings learn to generalize bias from a set of gendered words to others based only on the contextualized representations and how the classifier learns from being trained on a subset of the extended biased list. We trained a radial basis function-kernel support vector machine (SVM) classifier on the ELMo embeddings of 1000 random male and female-biased words from the extended biased list. Next, we evaluated the generalization of the other 4000 biased tokens. The accuracy of classification was taken as a measure of bias. The higher the accuracy was, the more the words were classified according to gender.

- **K-Nearest Neighbors approach (KNN) (Exp.5)**: We applied the KNN on the professional list to obtain the nearest 100 neighbors to each profession. For each token on the profession list, a randomly sampled sentence is used to get a contextualized representation. After applying the KNN algorithm to each profession, we computed the percentage of female and male stereotyped professions among the 100 nearest neighbors of each profession target token. Then, we computed the Pearson correlation of this percentage with the original bias of each profession.

One key factor of the experiments is randomization, as it considerably influences the experiments. ELMo provides a different representation of a word according to its context in a sentence. We randomized the sentence chosen for such representation to choose a particular representation for a word. Moreover, experiments 3-5 were repeated ten times and averaged to guarantee this randomness.

A difference should be noted between the professional list used in the direct bias experiments and the list used in the KNN experiment, Exp.2 and Exp.5. In Exp.2, we considered all of the sentences that contained the words of the professions. However, in Exp.5, we only considered 200 professions, including the 100 top female-biased professions and 100 top male-biased professions, and one random representation for each profession was considered.

4.1.5 Extension to the Spanish Language

Data and Lists. The corpus used in our evaluation is the Spanish version of the news corpus in WMT13, translation task from Spanish to English, as the English version is used in our English experiments. Consequently, the Spanish corpus has 174,441 lines.

The definitional list, equivalent list, and professional list were translated from English to Spanish. Native speakers were asked to revise all of the translations. The biased and extended biased lists were created from scratch by including the top biased female and male words concerning the grammar and semantic directions, as explained in section 4.1.5.

Evaluation Measures. Extending our evaluation to the Spanish language had considerable challenges. To begin with, we had to swap gender in sentences containing the equivalent pairs of the equivalent list. Given the properties of the Spanish language, adjectives and professions had to be swapped to the other gender added to swapping the equivalent pairs. The articles also had to be considered in the swapping procedure. The articles and the equivalent-pairs swapping were done automatically, but the rest was done manually to ensure that the whole sentence had the same gender. This manual check consumed time and resources, which prevented us from applying the experiments to different domains. We had to check all the swapped sentences (5,848 lines) to ensure each sentence was grammatically correct. We made sure not to swap the gender in sentences with a proper name. For example, *presidente Barack Obama*, was not swapped to *presidenta Barack Obama*.

Answering positively to research question 4 from section 4.1.2, we adapted the experiments to be suitable for the Spanish language and to give us insight about the bias in it. We applied the following:

- **Gender directions (Exp.6)**: We adopted the idea of obtaining different gender directions, including the semantic direction [Bolukbasi et al., 2016] and the grammar direction [Zhou et al., 2019].

Semantic direction $\overrightarrow{d_{pca}}$: We followed the same procedure previously mentioned in Exp.1 by using the PCA approach over the differences between male and female definitional contextualized word embeddings, from the sentences that have these definitional words and their swapped variants.

Grammar direction $\overrightarrow{d_g}$: We extracted the nouns (feminine and masculine) from the corpus, approximately 7,000 nouns for each gender, using Spacy parts of speech library[11] to extract these nouns. Next, in order to learn the grammar direction, since there are no equivalent pairs, linear discriminant analysis (LDA) for dimension reduction was applied. We applied LDA on 3,000 random sets of nouns of each gender multiple times. We tried random contextualized representations for these nouns. The range of the accuracy of learning the grammar direction was between 0.45-0.65. When the ELMo representations of these nouns were plotted, they were scattered in the subspace, as shown in Figure 4.8. There is no discrimination between a feminine subspace and a masculine subspace with these nouns.

Following the literature, the grammatical gender component in the computed gender direction is projected out to make the semantic gender direction $\overrightarrow{d_s}$ orthogonal to the grammatical gender direction:

$$\overrightarrow{d_s} = \overrightarrow{d_{pca}} - \langle \overrightarrow{d_{pca}}, \overrightarrow{d_g} \rangle \overrightarrow{d_g}, \tag{4.2}$$

where $\overrightarrow{d_s}$ is the semantic gender direction, which will be used in our experiments.

- **Direct bias (Exp.7)**: For the professional list in Spanish, we obtained two translations for each profession, a male-gendered and a female-gendered translation. The number of lines with professions is 4,987, with 2,198 being feminine and the rest masculine. We separately computed the direct bias on the male and female lists and then on their concatenated version. We computed the direct bias on the semantic direction $\overrightarrow{d_s}$ computed from Exp.6.

- **Clustering and classification experiments (Exp.8 and Exp.9)**: With respect to the biased list and the extended biased list, we performed the fol-

[11] https://spacy.io/models/es

lowing procedure to obtain the 500 and 5000 masculine and feminine biased words for these experiments:

- We downloaded the Spanish Word2vec embeddings[12], which is trained on one billion words.
- For these embeddings, the semantic gender direction was derived using the PCA method on the definitional standard word embeddings, and then the grammar gender direction was derived following the previously described method in Exp.6. Following equation 4.2, the semantic gender direction was computed.
- We obtained the top male and female biased words, with respect to the grammar direction and the semantic direction, respectively. Top 500 male-biased and 500 female-biased for clustering, both female and male biased words are considered the 'biased list', semantic biased list, and grammar biased list. For classification experiment, 5,000 female-biased and 5000 male-biased were gathered, and both together were considered the 'extended biased list', semantic biased list, and grammar biased list.

The clustering experiment, following the description in Exp.3, was applied to the semantic biased list and the grammar biased list. The classification experiment, following the description in Exp.3, was applied to the grammar and semantic extended biased list. As the Word2vec embeddings were trained on different words, not all the words in the lists are available. For the semantic biased lists, only 254 were available from the biased list, and 1881 were available from the extended biased list. Whereas for the grammar biased lists, 457 were available from the biased list, and 4,339 were available from the extended biased list.

4.1.6 Discussion

We will discuss the English and Spanish experiments separately in order to focus on different aspects.

[12] https://github.com/dccuchile/spanishwordembeddings

TEDx	Layer 2	Layer Concatenation
Direct bias	0.031	0.031
Clustering (Acc.%)	67.9% (2%)	68.4% (2%)
Classification (Acc.%)	87.1% (2%)	87.3% (3%)
KNN (Pearson Cor.)	**0.160 (0.3)**	0.501 (0.35)

Table 4.2: Results of TEDx experiments 2-5. Less biased in bold, the higher, the worse. Numbers between brackets show the difference between the maximum and the minimum numbers acquired from the ten experiments.

WMT	Layer 2	Layer Concatenation
Direct bias	0.028	0.026
Clustering (Acc.%)	**66.4% (3%)**	66.9 (3%)
Classification (Acc.%)	83% (2%)	85.4% (3%)
KNN (Pearson Cor.)	0.971 (0.02)	0.975 (0.01)

Table 4.3: Results of WMT experiments 2-5, the higher, the worse. Numbers between brackets show the difference between the maximum and the minimum numbers acquired from the ten experiments.

4.1.6.1 English Results

Figures 4.2-4.5 show results of Exp.1 for all domains and layers. Tables 4.2-4.5 show the results of experiments Exp.2, Exp.3, Exp.4 and Exp.5. Regarding the last three experiments, the average of ten experiments is shown for each domain for the third and concatenation layers. The numbers in brackets show the difference between the maximum and the minimum of the ten experiments.

Propagation of Gender bias from the corpus to the contextualized word representations: The variability in the number of lines of the corpus, the diversity of the professions, the existing biased words, and the percentage of feminine biased words are factors influencing our analysis and conclusions. Accordingly, our analysis is not based on the comparison between domains. Still, it relies on deriving conclusions about the gender bias propagation across the domains from the corpus level to the contextualized word representations level. Understanding the effect of gender bias on a particular domain leads to awareness of its impact on training neural models on such domain in different tasks.

From Exp.1, shown in Figures 4.2-4.5 for the plots for the percentage of variance explained by the ten gender pairs of the definitional list, PCA can derive a dom-

Pubmed	**Layer 2**	**Layer Concatenation**
Direct bias	0.021	0.021
Clustering (Acc.%)	79.4% (22%)	**77.3% (17%)**
Classification (Acc.%)	85.2% (3%)	**84.9% (4%)**
(Pearson Cor.)	1	1

Table 4.4: Results of Pubmed experiments 2-5,the higher, the worse. Numbers between brackets show the difference between the maximum and the minimum numbers acquired from the ten experiments.

Europarl	**Layer 2**	**Layer Concatenation**
Direct bias	0.05	**0.048**
Clustering (Acc.%)	**68.4 (1%)**	68.6 (2%)
Classification (Acc.%)	86% (2%)	**85.9 (2%)**
KNN (Pearson Cor.)	0.919 (0.03)	0.906 (0.03)

Table 4.5: Results of Europarl experiments 2-5, the higher, the worse. Numbers between brackets show the difference between the maximum and the minimum numbers acquired from the ten experiments.

inant subspace, as the subspace of gender-flipped vectors contain less informative dimensions. After using the PCA, the first component appears to have dominant information, as it explains more variance than the other components, whereas, in the Europarl domain, the first two components explain more variance, not only the first.

As shown in Tables 4.2-4.5, the direct bias of professions computed with gender direction demonstrates the propagation of gender bias in professions across the domains. To understand the impact of gender bias propagating from the corpus to the contextual word representation, we applied the clustering and classification techniques which associates male and female nouns as concept words and their stereotypical clustering and classification. Clustering experiments, illustrated in Figure 4.6 and Tables 4.2-4.5, show that male-biased words cluster together and so do female-biased words with accuracy more than 60% for all domains. Therefore, clusters seem to align with gender across domains. Classification experiments (Tables 4.2-4.5) demonstrate that bias is generalized from some gendered words to others, based only on their contextualized representations, with >80% accuracy across the four domains. Therefore, the classifier learns bias from gendered biased words. Accordingly, bias tends to propagate from the corpus level to the encoding level, which directly answers the first question in research questions (section 4.1.2).

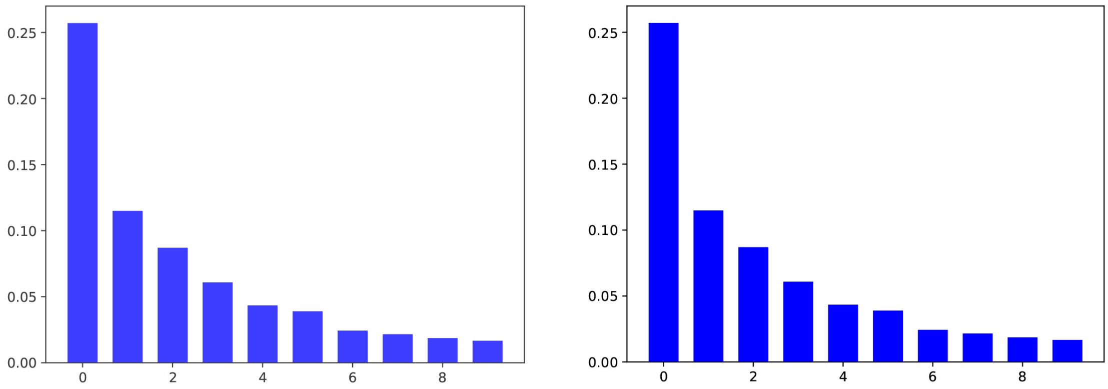


Figure 4.2: X-axis refers to the ten PCA components and Y-axis refers to the percentage of variance explained by the ten principal components in TEDx Exp.1 in layer 2 (left) and layer concatenation (right).

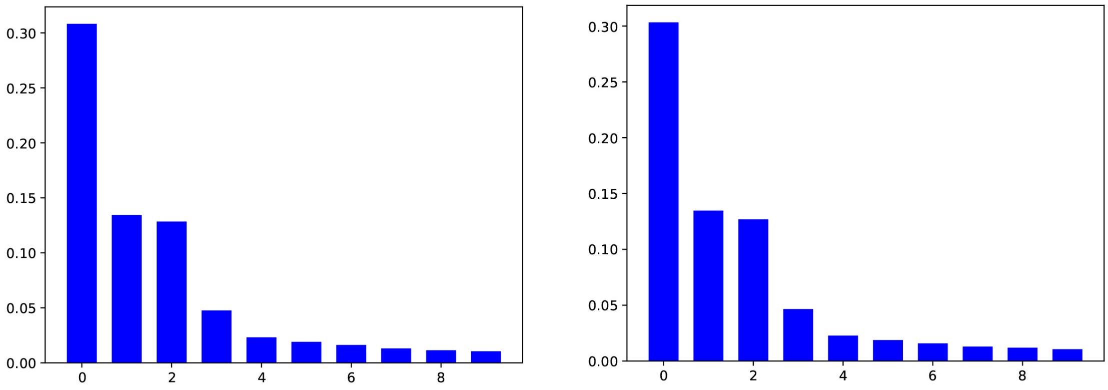


Figure 4.3: X-axis refers to the ten PCA components and Y-axis refers to the percentage of variance explained by the ten principal components in WMT Exp.1 in layer 2 (left) and layer concatenation (right).

Layer concatenation vs. layer 2: The main objective of experimenting on different layers was to deduce which layer results in less biased representations. Experimenting on the three different layers led to slight differences; thus, we cautioned against definitive conclusions. By experimenting on the two different layers, the layer concatenation and the last ELMo layer (layer 2), we observe varying results. Since the last ELMo layer captures different semantic aspects of word embeddings, layer 2 encodes less bias in the case of more general domains (TEDx and WMT) in experiments 2-5. The concatenation layer also benefits from the syntax and semantic aspects of the three ELMo layers in the more specific domains (Pubmed and Europarl). However, the difference between the results of the two layers in the case of Europarl is not significant. The difference in means is 0.002 in direct bias, 0.2% in

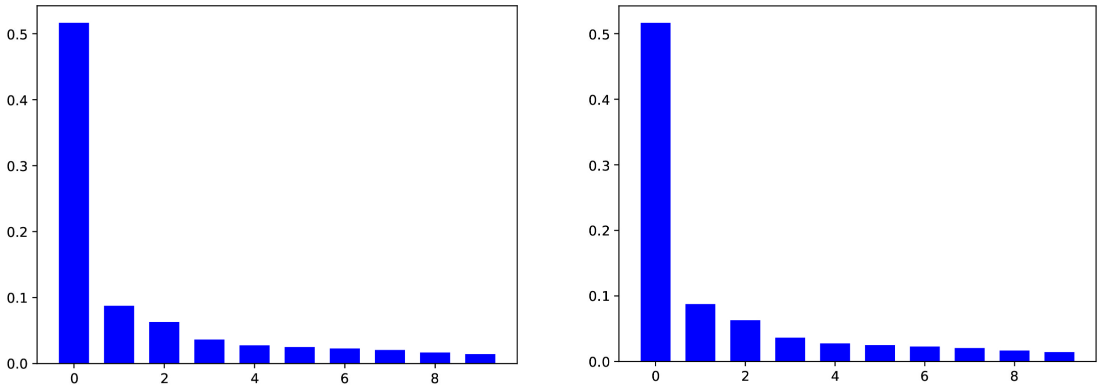


Figure 4.4: X-axis refers to the ten PCA components and Y-axis refers to the percentage of variance explained by the ten principal components in Pubmed Exp.1 in layer 2 (left) and layer concatenation (right).

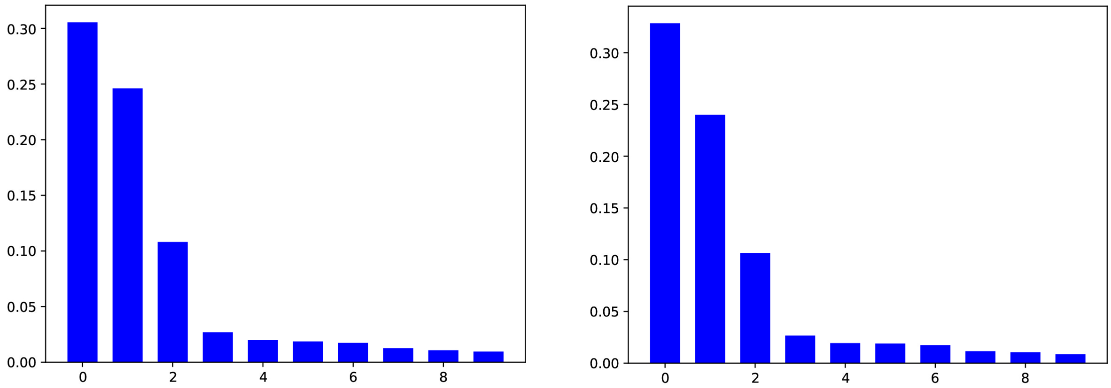
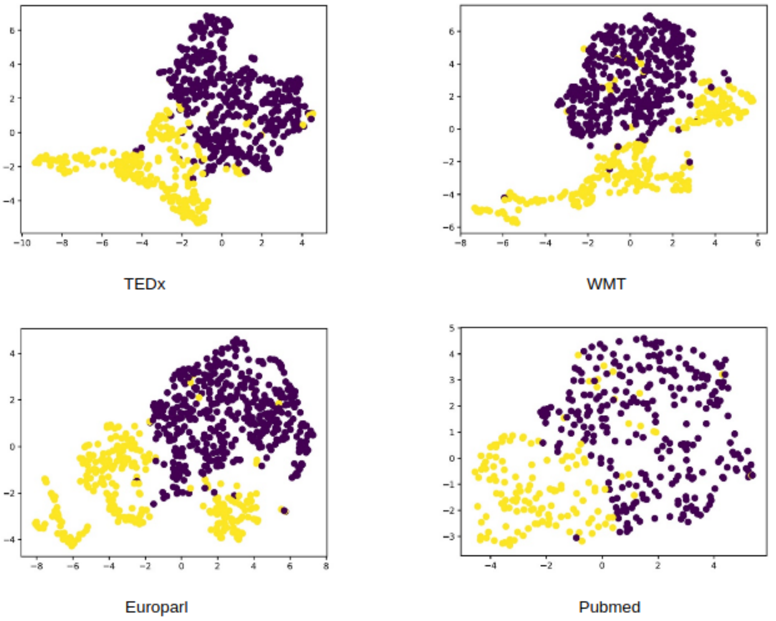


Figure 4.5: X-axis refers to the ten PCA components and Y-axis refers to the percentage of variance explained by the ten principal components in Europarl Exp.1 in layer 2 (left) and layer concatenation (right).

clustering experiment, 0.1% in classification, and 0.013 in KNN. From the different results, we can conclude that using the representation of words from different ELMo layers is not distinguished concerning gender bias. Accordingly, choosing the layer should depend on factors other than gender bias.

Professions perpetuates serious bias: Responding to the second research question (and also to first), Tables 4.2-4.5, direct bias computation and KNN experiments show obvious kinds of bias in most domains, except for TEDx, where less bias from KNN experiment is demonstrated.The bias of professions is evident in Pubmed that frequently associates medical occupations with male gender pronouns.



TEDx

WMT

Europarl

Pubmed

Figure 4.6: Clustering experiments for TEDx, WMT, Europarl and Pubmed for representations from layer 2, male clusters are in violet and female clusters are in yellow.

Effect of randomization is higher for clustering and KNN: Randomization of the ELMo embeddings of used words has led to the most varying results when repeating the clustering and KNN experiments. The corpus size and the number of word occurrences can also affect the randomization. Some of the ten experiments have yielded a wide range of maximum and minimum results. The differences between the minimum and maximum in Exp.3 in Pubmed have reached 22% in layer 2 and 17% in layer concatenation. The wide range of differences can be attributed to the randomized representations of the words, which has resulted in different clustering. Pubmed is a large corpus that has a different context for the biased words used in the clustering experiment. Similarly, Exp.5 in TEDx is highly affected by randomization, showing differences of 0.3 in layer 2 and 0.35 in layer concatenation. Although the mean of the experiments sometimes implies a similar bias between corpora, lower numbers are shown in the results in one corpus than in the other.

Measures to be considered and more reliable than others: Exp.1 and Exp.2 can be considered the most reliable measures across the four domains because they directly observe gender nature in the domains. Exp.3 and Exp.4 can be regarded as related measures. They are synchronized and can reflect the bias of the representations in the four domains.

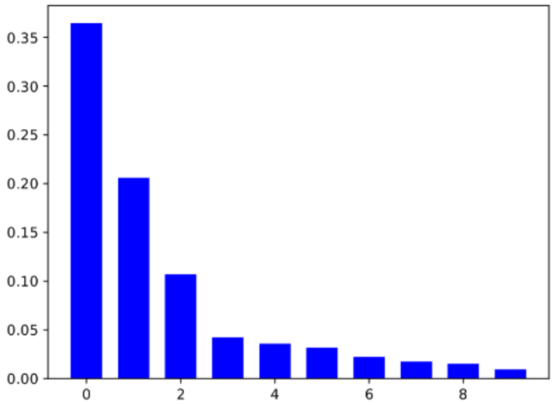


Figure 4.7: X-axis refers to the ten PCA components and Y-axis refers to the percentage of variance explained by the ten principal components of definitional pairs' embeddings of Spanish.

KNN can be discarded when dealing with domains of low representation of professions. It is unreasonable to compute less than 100 KNN for each profession to understand how neighbors go together. Therefore, the KNN measure is inconsistent in the case of the corpus with less biased professions. This can be applied to the Pubmed corpus, where only 64 professions out of 200 are present, and the correlation is always 1. This conclusion directly answers the third question from section 4.1.2.

4.1.6.2 Spanish Results

Again, Exp.8 and Exp.9 mainly used randomization and were repeated ten times, and their mean was calculated. Randomized representations were also used in Exp.6 to extract the nouns for the grammar direction.

Spanish semantic direction (results from Exp.6): By applying the PCA experiment on the embeddings of gender definitional words in original and swapped sentences, the percentage of variance represented from the PCA components of definitional vector difference was obtained (see Figure 4.7).

Additionally, for Spanish, we observe that the first component represents the most significant percentage of the variance of the ten PCA components, reaching 0.36, and this top component determines the gender direction. After projecting out the

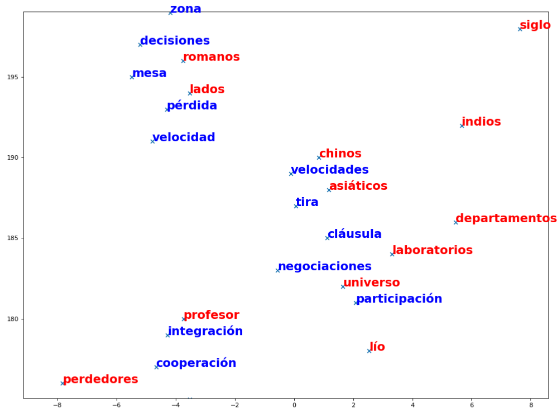


Figure 4.8: Plotting Spanish representations of nouns on gender direction.

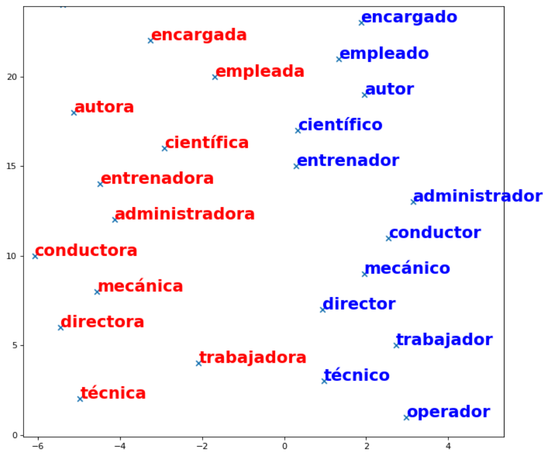


Figure 4.9: Plotting Spanish representations of occupations on gender direction.

grammar direction from the semantic direction, we found a slight decrease in the vector's percentages of variance determining the semantic direction.

Direct bias is higher for female professions: Direct bias is studied in both the grammar and the semantic gender directions, as described in Exp.7. Figure 4.8 and Figure 4.9 and illustrated that plotting the professions, rather than the nouns, appears more segregated with gender (feminine vs. masculine). After calculating the direct bias on feminine and masculine professions separately, as shown in Table 4.6, the former case shows a higher direct bias. This means that the feminine professions are closer to the semantic direction and, consequently, more biased.

Clustering and classification have to be performed on semantic biased words: Grouping the embeddings of masculine and feminine words does not always indicate bias due to grammatical gender [Zhou et al., 2019]. The clustering and classification accuracy, noted in Table 4.7, is higher with words that are grammar biased. This is normal because nouns will be clustered and classified according to gender. On the other hand, the accuracy is still high in clustering and classifying the semantic biased words. Thus, clustering according to gender and generalization of learning bias occur too.

Direct bias	Semantic direction
Female-version of professions	0.1215
Male-version of professions	0.0572
Male and Female together	0.098

Table 4.6: Direct bias of Spanish professions with semantic direction.

	Classification (Acc.%)	Clustering (Acc.%)
Semantic biased words	94.25%	84.48%
Grammar biased words	99.05%	93.27%

Table 4.7: WMT Spanish clustering and classification experiments.

4.1.7 Conclusions

This section makes the following contributions: first, we have extended existing analyses of gender bias to state-of-the-art ELMo contextual word models and indicate that such bias exists in these models. This highlights the scope of the problem of fairness in state-of-the-art models for language processing. We have provided evidence that gender bias is encoded strongly in contextual word models in professions and stereotypical nouns.

Second, this study understands the effect of domains on contextualized word representations. Domains differ in statistics and nature and in representing gender bias in contextualized word embeddings. This shows that such unsupervised methods perpetuate bias in downstream applications, and our work forms the basis of evaluation. Additional contribution is analyzing the gender bias represented in Spanish contextualized word embeddings. This research reminds us that languages other than English have different properties that need further treatment. Finally, we have compared various measures to understand which ones to rely on to help mitigate gender bias in these embeddings. The extensive analysis is consistent with previous studies [Dev et al., 2020]. The techniques used to measure or mitigate bias in standard embeddings do not necessarily succeed for contextualized embeddings.

One advantage of the gender direction and direct bias evaluation measures is being more generalized and based on less specific lists that are not domain or language dependent. On the other hand, direct bias seems to be less discriminating (see Tables 2-5). While the clustering and classification seem more discriminating (again, responding to the fourth research question from section 4.1.2, see Tables 2-5), the disadvantage is strongly dependent on the existing vocabulary and less generalized to different domains. This can be attributed to studying the clustering and classification of embeddings of biased words that are biased in the original Word2vec embeddings. At the same time, each corpus may have its own set of different biased words. Again, applying clustering and classification of the biased words of each corpus would not be comparable from one domain to another; therefore, obtaining the biased set from the original embeddings is still more reasonable. KNN can be neglected as a measure when there are few biased professions. As professions do not have enough neighbors, it is difficult to evaluate whether they are truly biased or just a matter of lacking neighbors.

4.2 Evaluation Gender Bias in Multilingual Machine Translation

4.2.1 Motivation

There are various sources of gender bias; one is model bias. The model architecture can impact the behavior of the model towards gender bias. The bilingual NMT approaches are studied in various studies [Stanovsky et al., 2019, Saunders et al., 2020, Saunders and Byrne, 2020] showing that these models amplify bias when translating to stereotypes. The research never studied the same trend in multilingual architectures. Multilingual neural machine translation architectures mainly differ in the number of sharing modules and parameters applied among languages. That can help us study whether the chosen architecture, when trained with the same data, influences gender bias from an algorithmic perspective. This work is a part of more extensive research presented in [Costa-jussà et al., 2022]. We only mention what is related to our thesis and our study.

4.2.2 Research Questions

To start the study, we are motivated to answer a number of research questions concerned with multilingual architecture. These questions are answered in sections 4.2.4 and 4.2.5:
1- What is the effect of parameters sharing on multilingual gender biased translations?
2- How gender information is encoded in the embeddings in multilingual architectures and what is their effect on the translation accuracy?

4.2.3 Experimental Framework

In this section, we report the details of the experiments including data and training architecture and parameters.

4.2.3.1 Architectures

The architectures used in this experiment are described as follows:

Bilingual Encoder-Decoder. Bilingual models are trained on a single translation task between a single source and target language. This approach would be taken as a reference in our experiments, as such architectures devote the entire representation capacity of the model to a single task, capturing specific features and correlations of the language pair.

Shared Encoder-Decoder. [Johnson et al., 2017] trained a single encoder and decoder with multiple input and output languages. Given a language set, a shared architecture has a universal encoder and decoder fed with all initial language pairs at once. The model shares vocabulary and parameters among languages to ensure that no additional ambiguity is introduced in the representation. By sharing a single model across all languages, the system can represent all languages in a single space. This allows translation between language pairs never seen during the training process, which is known as zero-shot translation.

Language-Specific Encoders-Decoders. This work uses the no sharing approach [Escolano et al., 2021a] since it is the most contrastive to the shared encoder-decoder. This language-specific approach involves training independent encoders and decoders for each language. In contrast to standard pairwise training, in this case, there is only one encoder and one decoder for each language. Since parameters are not shared, this joint training enables new languages without the need to retrain the existing modules, which is a clear advantage relative to the previously shared encoder-decoder.

4.2.3.2 Data and Parameters

Experiments are performed on EuroParl data [Koehn, 2005] for English, German, Spanish and French with parallel sentences among all combinations of these four languages and with approximately 2 million sentences per language pair. Systems are trained in English, German, Spanish, and French with parallel sentences among all four languages. We also built pairwise bilingual systems (based on the transformer) on the corresponding language pair data. As validation and test sets, we use *newstest2012* and *newstest2013* from WMT[13]. All data are preprocessed using standard Moses scripts [Koehn et al., 2007]. Experiments are performed using the approach provided by Fairseq[14]. We use six layers, each with eight attention heads, an embedding size of 512 dimensions, and a vocabulary size of 32k subword tokens with byte pair encoding [Sennrich et al., 2016] (per pair). Dropout is set as 0.3 and trained with an effective batch size of 32k tokens for approximately 200k updates using the validation loss for early stopping. We use Adam [Kingma and Ba, 2015] as the optimizer, with a learning rate of 0.001 and 4000 warmup steps. We report gender bias evaluation using WinoMT with metrics accuracy (Acc.), ΔG, ΔS and **M:F** proposed by [Saunders and Byrne, 2020].

[13] `http://www.statmt.org`
[14] Release v0.6.0 available at `https://github.com/pytorch/fairseq`

Language Set		en,de,es,fr				
Lang	System	BLEU↑	Acc↑	ΔG↓	ΔS↓	**M:F** ↓
ende	bil	21.61	**64.10**	**5.7**	8.30	**1.84**
	shared	21.39	53.86	23.59	8.33	3.87
	lang-spec	**22.01**	56.28	17.45	**7.83**	2.92
enes	bil	25.82	46.00	22.90	**2.40**	**3.13**
	shared	28.08	51.67	24.77	5.49	4.09
	lang-spec	**29.53**	**54.19**	**20.73**	7.64	3.66
enfr	bil	26.73	42.18	**21.59**	14.16	**2.67**
	shared	28.43	45.55	24.99	**0.06**	3.88
	lang-spec	**29.74**	**45.81**	28.45	5.64	4.63

Table 4.8: Results in terms of BLEU and Gender Accuracy (Acc.): Bilingual (bil), Shared (shared) and Language-Specific (lang-spec). In bold, best global results. Underlined, best results between multilingual systems.

4.2.4 Results

We report the results in terms of translation quality and gender accuracy. Table 4.8 reports the results in terms of BLEU and gender accuracy for the architectures described in section 2.3.2. When comparing bilingual vs. multilingual architecture, and consistently with previous studies [Johnson et al., 2017], multilingual systems improve upon bilingual systems in terms of translation quality. However, we cannot conclude the same in terms of gender accuracy. The multilingual architecture improves upon the bilingual architecture for two of the three language pairs in terms of gender accuracy and ΔS. Regarding the rest of the gender measures, the bilingual system tends to be better, especially for **M:F**.

When comparing the multilingual architectures, we observe that the language-specific architecture shows consistent gains in BLEU of approximately 0.4-3.6%. Such superiority of the language-specific system is kept in terms of gender accuracy. The conclusions are similar when comparing ΔG and **M:F** values, with the language-specific system showing gains of up to 6% and clearly superior in 2 out of 3 language pairs. Since WinoMT is divided into 46,97% male, 46,86% female and 6.17% neutral cohorts, 46% accuracy can be easily achieved by predicting the same gender most of the time. For the shared architecture, we observe that the high ΔG is explained by having a strong preference for predicting male gender.

Regarding ΔS, the results tend to be better for the shared architecture. These differences in ΔS are attributable to the fact that the accuracy of the shared system, for both pro- and anti-stereotypical occupations, is much lower than the language-specific system, which derives from fewer differences. Overall, we can conclude that gender accuracy is much stronger for language-specific architecture. These results and conclusions give us answers to question one in section 4.2.2.

4.2.5 Interpretability Analysis

Gender information in source embeddings. Studying how source contextual embeddings codify gender information can promote understanding about how gender is predicted in translations and answer question two in section 4.2.2. We followed our classification approach in the previous evaluation study in contextualized embeddings (section 4.1), which uses embeddings to train an SVM [Cortes and Vapnik, 1995] and classify in two groups. We applied the same classification measure on two word types for source embeddings classification by using the information provided by WinoMT to measure how gender information is reflected in their contextual embeddings, determiners (*The*) and occupations. The first category is initially neutral, as it is equally employed in all categories. Therefore, all gender information present in these embeddings must come from the context of the sentence. For each system and word type, we trained an SVM classifier with a radial basis function kernel on 1000 randomly selected sentences from WinoMT and tested the remaining 2888 sentences from the set. Words are represented as their first subword in case they are split in the vocabulary.

We performed ten independent experiments to guarantee the randomization of token representations. Achieving more accuracy in the classification results means that more information on gender is encoded in the source embeddings. Figure 4.10 shows the results for this classification for all bilingual and multilingual systems (from left to right) for both determiners and occupations.

Bilingual systems show that the target language substantially impacts the amount of gender information encoded in the contextual representations. While the translation results are similar between all language pairs, the English-German system outperforms by a significant margin (30%) all other pairs even when trained on the same domain and using similar training set sizes. These results correlate with the gender accuracy illustrated in Table 4.8 showing that the systems that encode

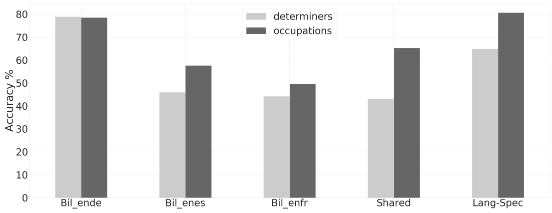


Figure 4.10: Classification results, from left to right: Bilingual (English-to-German/Spanish/French), Shared and Language-Specific. Determiner in light, occupations in dark.

more gender information on their contextual representations produce more accurate gender translations.

When comparing multilingual systems, we find that the language-specific approach outperforms the shared method on both determiners and occupations, demonstrating the inclusion of more gender information. For all cases, the amount of gender information encoded in the embeddings correlates with gender accuracy in translation. With this, we answer the second research question in 4.2.2.

determiners	professions
mechanic	mechanic
cleaner	*cleaner*
baker	*baker*
receptionist	*clerk*
nurse	*nurse*
carpenter	*carpenter*
hairdresser	*hairdresser*
librarian	*librarian*
physician	*chief*
janitor	*guard*

Table 4.9: List of the 10 most common misclassified occupations by the SVM models trained with determiners and professions. In italics, the errors in common with the manual evaluation.

Table 4.9 reports the list of the 10 most common misclassified occupations by our classifier. We report in italics the errors in common with the manual evaluation, reported later in this thesis. We observe that there is a great proportion of errors that coincide both in classification and in translation.

4.2.6 Manual Analysis

In this section, we perform a manual analysis of occupation errors across languages. Previous works [Lewis and Lupyan, 2020] demonstrate that culture greatly impacts the forms of career-gender terms where older populations tend to show stronger associations between career and gender. Such an impact affects male/female representations in the data [Madaan et al., 2018] where some occupations are represented with the masculine form only, or a higher proportion of males is represented. Figure 4.11 shows that mistranslated occupations vary from one language to another. Our study covers occupations incorrectly predicted in 35%[15] of the sentences containing them and in bilingual, shared and language-specific systems. In what follows, we offer a non exhaustive explanation covering an appropriate proportion of the errors shown in Figure 4.11.

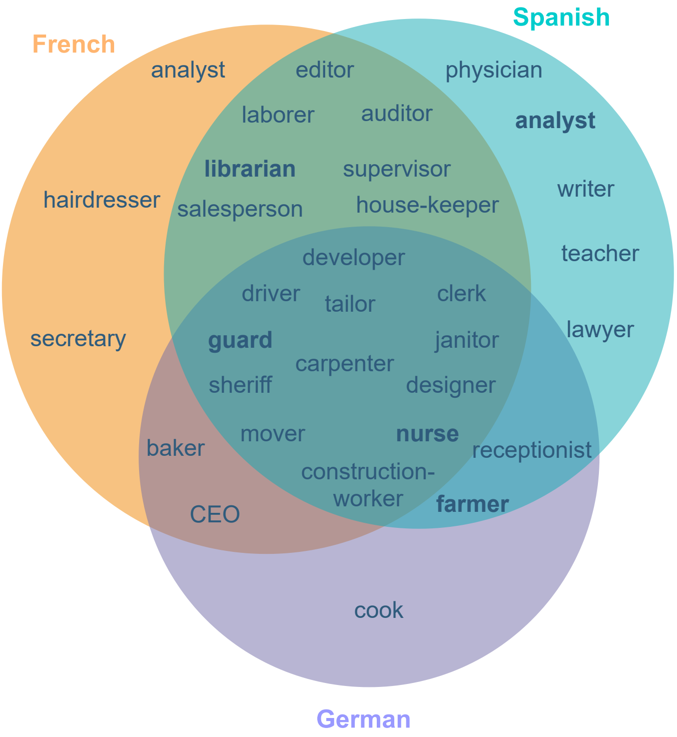


Figure 4.11: Misclassified occupations in terms of gender. Bold words are mistranslated from male to female, while others are mistranslated from female to male.

In bold, we show the occupations that are wrongly predicted to female, whereas the rest are occupations that are wrongly predicted to male. We observe that most errors come from associating occupations to male rather than to female. This may

[15] This was a trade-off between the percentage of errors and number of sentences enabling us for a manual analysis

be because of having a higher male representation in our data [Madaan et al., 2018]. This conclusion is consistent with previous studies [Stanovsky et al., 2019]. More than this, we see that the occupations that are wrongly translated vary with the language. However, when comparing Romance languages (Spanish and French) common errors in occupations raise up. As follows we try to come up with some linguistic/cultural explanation of why we are obtaining these common errors.

Regarding German errors, *nurse* tends to be assigned the feminine form (*Krankenschwester = sick + sister*), which is mostly used in everyday language. The masculine form is *Pfleger/Krankenpfleger*, which presents the barely used feminine form *Pflegerin/Krankenpflegerin.*

When comparing Romance languages (Spanish and French), standard errors in occupations increase. Because the default gender in Spanish and French was masculine in the past [Frank et al., 2004], such errors relate to linguistics and culture together. In French culture, masculine forms are predominantly used as gender neutral, and only the article may vary for some occupations, such as *présidente/président* (CEO), even in cases where the feminine form exists. Thus, some speakers say e.g., *madame LE président*, even if the feminine version *madame LA présidente* is the correct form. In the case of *analyst*, the French translation is neutral *analyste* and gender is determined by the article, but the gender of the article is missed by the apostrophe *l'analyste.*

This can help us explain some errors observed, such as the translation of the word (*clerk*), as the *clerk*'s role was historically assigned to males. Consequently, both languages have only the masculine form, although suitable feminine/masculine translations would be possible. Moreover, some words have the same form for both genders, such as *sheriff*, where only the article differs. An interesting example of a feminine mistranslation is the word *guard.* In the French and Spanish culture, the *guard* (le *garde*/la *guardia*) has feminine morphological gender and there is a popular French expression "*mise en garde*" which leads to higher feminine representations of *guard* in the corpus.

4.2.7 Conclusions

By evaluating the different architectures with WinoMT evaluation metrics and then analyzing the gender information in the embeddings, we can understand why the multilingual NMT architecture impacts gender accuracy. Our interpretability analysis shows that source embeddings in the language-specific architecture retain more information on gender. Comprehending this along with the performance of the systems in the synthetic benchmark of WinoMT, we can conclude that the language-specific model outperforms the shared one.

Finally, a manual analysis shows that most errors are made by assuming a masculine occupation instead of a feminine occupation. In contrast, the inverse error occurs when a feminine version of a word with another meaning is possible.

4.3 Final Thoughts on the Chapter

Evaluation is the key. Evaluating different architectures, different data domains, and representations of embeddings are the key to explaining the nature of bias, conceptualizing it, and understanding its impact on our system. Evaluation is the most critical step toward interpreting how our systems deal with it.

Contextualized embeddings encodes bias. Gender bias is encoded strongly in contextual word models in professions and stereotypical nouns. While gender information can be helpful for specific tasks, biased embeddings can inflect stereotypes and biased forms in other tasks. For the task of NMT, It is worth mentioning that a neutral word in a neutral context should not have any gendered information, e.g., *my friend is a doctor*, while the gender information in neutral words would be useful if there is already a gender hint in the sentence, e.g., *my friend and her cousins are managers*.

Architectures matter. Bias tends to be attributed to data [Costa-jussà, 2019]. However, our evaluation study shows that algorithms amplify the bias and that the system's architecture impacts the behavior. This conclusion can be considered in research/deployment by systematically evaluating our algorithms regarding bias.

5 Towards Mitigation Approaches of Gender Bias in Machine Translation

This chapter proposes methods for mitigation of amplification of bias NMT, these methods are adapted from our papers [Basta et al., 2020, Basta et al., 2022]. Each section will discuss one of the following methods with its details:

1. Mitigation by adding the previous sentence and the speaker gender identifier, in section 5.1.
2. Mitigation using contextual embeddings and relaxed positioning conditions, in section 5.2.

5.1 Mitigation by Adding the Previous Sentence and the Speaker Gender Identifier

5.1.1 Motivation

There are multiple problems with current paradigms in NMT; one of them is operating on a sentence-by-sentence basis. This is a structural limitation of our systems [Läubli et al., 2018], mainly when translating a coherent context, due to the current sentence's need for the previous one for better translation and specifically for better gender understanding. Another problem is missing information when translating from different morphological languages. One of this information can be the gender of the speaker leading to a lack of agreement on gender with the subject.

Both problems arouse a lack of fluency and adequacy of the translated sentence and gender-related errors.

Gender-related errors are not only causing harm to translations, but also perpetuating a male bias amounting to female discrimination in society. Such problems motivated us to study the effect on translations of two directions; adding more context and aggregating a gender tag related to the speaker. We studied the approaches' potentiality in general translations' accuracy and gendered inflections.

5.1.2 Research Questions

Given the previous two motivated problems, we sought to answer the following research questions, later the answers are given in sections 5.1.5 and 5.1.6 :

1. What effect does adding a previous sentence or a speaker's gender tag to a sentence have on the translation and gender accuracy?
2. Other than gender accuracy, can we get further benefits from these techniques?

5.1.3 Proposed Methodology

To understand the impact of adding previous sentence or speaker gender identifier on the gender accuracy of NMT, we chose a different NMT baseline [Fonollosa et al., 2019]. As follows, we describe the baseline system and the techniques.

Baseline System. Neural Machine Translation with joint source-target self-attention is an alternative architecture to the standard transformer [Vaswani et al., 2017]. It is a more simplified architecture[1] by [Fonollosa et al., 2019], which only uses the decoder block and it adopts the idea of language modeling for translation task, instead of having both encoder and decoder. The joint source-target representations are learned in the early layers, and positional embeddings are applied independently to the source and target. In the system, language embeddings are employed to represent the language of the source and the target separately. Unlike the self-attention in standard transformers, the

[1] https://github.com/jarfo/joint

Methods	Examples
Baseline	I have only done this once before.
+**PreSent**	I have only done this once before. <sep> This is not a joke.
+**SpeakerId**	MALE I have only done this once before.

Table 5.1: Methodologies examples

authors propose a locally constrained attention to attend only to a token's locality and form a reduced receptive field. A simplified version always leads to fewer parameters and better memory usage.

Enriching Methodologies. We consider the following techniques, where different kind of information is added in each way, Table 5.1 shows examples of both techniques.

- **Adding the previous context sentence (PreSent):** This method is mainly concatenating two consecutive sentences with a separator token. The main idea is increasing related context adopting the method from [Junczys-Dowmunt, 2019].

- **Incorporating the speaker gender identification (SpeakerId):** Incorporating the information of the gender of the speaker in NMT by adding the gender tag before each sentence [Vanmassenhove et al., 2018]. This approach is beneficial when translating from a less inflected language to a more inflected one, e.g., English to Spanish.

5.1.4 Experimental Setup

Data. Spanish has two features accounting for its suitability for our task. The first is that it is a high grammatical gendered language, and the second is that omission of pronouns makes the translation in English-Spanish task more challenging. These are core motivations to choose to test our work on this particular English-to-Spanish task. However, our conclusions may be extendable to other English-to-Romance languages such as Italian, French, and Portuguese. For training, we have chosen the data motivated by the fact that the data contains the information of the gender of the speaker added to the document-level information. Europarl data, compiled in the previous work by [Vanmassenhove et al., 2018], meets these requirements. The size of the English-Spanish dataset is considered moderate, with 1,419,507 sentences. For testing, we have used two test datasets: a random set of Europarl (2000 sentences) following the work of [Vanmassenhove et al., 2018] and the gender-balanced set from Wikipedia biographies (GeBioCorpus) [Costa-jussà et al., 2020] that contains 1000 sentences from male biographies and 1000 sentences from female biographies. The gender of the main character in the biography article is used as the gender tag.

Parameters. The model is built on top of fairseq[2] library and the parameters are customized as follows: Adam optimizer, 30K training steps, 14 layers, 512 as embedding dimensionality, feedforward expansion of dimensionality 2048 and 8 attention heads, based on best performing parameters from previous work [Fonollosa et al., 2019].

5.1.5 Results

Methods	Europarl	GeBio
Baseline	44.01	36.34
+PreSent	**45.10**	**36.55**
+SpeakerId	44.18	36.51

Table 5.2: Results on the Europarl and on GeBioCorpus test sets. Best results in bold.

BLEU Results (Table 5.2). These results have been acquired by testing the Europarl test set and GeBioCorpus. Adding the previous sentence has a higher impact in Europarl (+1.09) than in GeBioCorpus (+0.21) since documents in GeBioCorpus are not coherent (all sentences belong to the same document, but some sentences are not sequent to each other). Adding the gender tag shows the same effect in GeBioCorpus than in Europarl (+0.17), even if the speaker identification differs between Europarl and GeBioCorpus sets. In the former, the speaker tag relates to the speaker, whereas it relates to the main biography character in the latter. By this we answer question one in section 5.1.2 regarding the translation quality.

Evaluating on WinoMT (Table 5.3 and Figure 5.1). Evaluating on WinoMT dataset by [Stanovsky et al., 2019] system relies on judging the accuracy of the translated gender of certain entities in the sentences compared to the gold gender of these entities. It is important to note that we are evaluating the behavior of the trained

[2] https://github.com/pytorch/fairseq

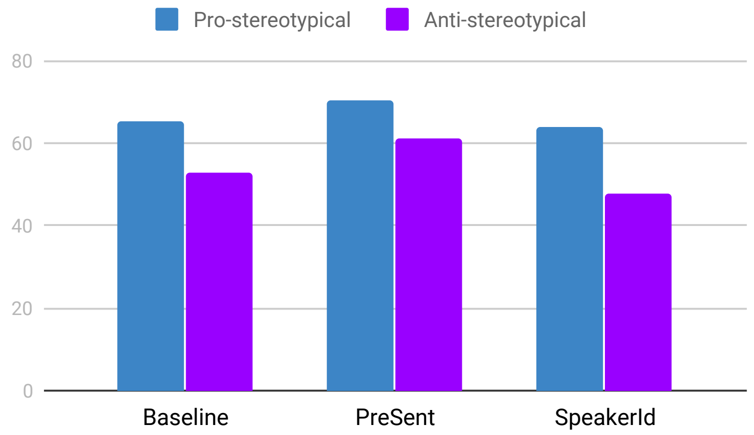


Figure 5.1: Acc.% on gender translation with respect to pro-stereotypical entities and anti-stereotypical entities in WinoMT.

systems without adding any additional information to the WinoMT sentences. As shown in Figure 5.1, the systems are performing better on the pro-stereotyped portion of WinoMT than on the anti-stereotyped one.

Among the three translation systems (baseline, PreSent and SpeakerId), as shown in Table 5.3 and Figure 5.1, the PreSent has the highest accuracy in detecting the gender. It shows better performance, whether in pro-stereotyped or anti-stereotyped translations, with 61% accuracy and 12.2% ΔG, the lowest f1-score difference between them pro-stereotyped male and female translations. The PreSent has the lowest ΔS of 9.2, which shows that it has the best performance in translating anti-stereotyped occupations with a difference of 2.8 with the baseline.

On the other hand, the SpeakerId system has the least translation accuracy of 52.5% with higher ΔG and ΔS, demonstrating that it favors the pro-stereotyped translations in general, especially the male ones.

As mentioned, WinoMT is a test set that does not contain information at the level of the document nor speaker identification, so translation with our methodologies is done without this information. However, we can note that adding the information of the previous sentence makes the system more robust even when making inferences without such information. Reaching that, we get the answer to the last part of question one in section 5.1.2 regarding gender quality with these methodologies.

Methods	Acc.↑	ΔG↓	ΔS↓
Baseline	56.0	18.7	12.0
+PreSent	**61.0**	**12.2**	**9.2**
+SpeakerId	52.5	22.2	15.5

Table 5.3: WinoMT evaluation results. Acc. indicates gender accuracy (% of instances the translation had the correct gender), ΔG denotes the masculine/feminine difference in F1 score and ΔS notes the difference in accuracies between pro-stereotyped translations and anti-stereotyped ones. In bold, best results are introduced.

5.1.6 Manual Analysis

To answer the second question in section 5.1.2 regarding other advantages when using the approaches, we had to do manual analysis to inspect the translations thoroughly. In Table 5.4 and Table 5.5), we report some translation examples for both PreSent and SpeakerId techniques. We have observed the following advantages:

Helping Towards Name Entity Disambiguation (in terms of gender). Both SpeakerId and PreSent techniques seem to impact name disambiguation positively. The challenge of resolving names occurs when the translation system can not predict its gender correctly, as it did not adequately learn it during training. Therefore, if the trained system does not recognize the names, they can be wrongly translated, and most of the time, to the male-gendered translation. Both approaches improve the translations in this case, even in the case of GeBio test-set, as shown in Table 5.4 and Table 5.5).

Improvement in Morphological Agreement and Translation Quality. Adding the contextual info (PreSent) improves the morphological agreement of the subject and its related nouns. In Table 5.4 , we can see that adjective (*catalana*) and nouns (*analista*, *activista*) agree with the female subject (*Míriam*). We can also note an improvement in the style of translation, giving a better arrangement of words *defensora española* in the last example in Table 5.4.

	Named Entity Disambiguation
Source	María del Carmen Pérez ...is **a Spanish Egyptologist , curator and researcher**.
Baseline	María del Carmen Pérez ...es **un egipcio pintor , curador e investigador español**.
+PreSent	María del Carmen Pérez ...es **una ciudadana española egipcia , curadora e investigadora**.
	Better dealing with articles
Source	Míriam Hatibi ... is **a Catalan data analyst and activist**.
Baseline	Míriam Hatibi ... es **un analista de datos catalán y un activista**.
+PreSent	Míriam Hatibi ... es **una analista y activista catalana en materia de datos**.
	Better style of translations
Source	Helena Maleno Garzón ... is a Spanish human rights **defender , journalist**, **researcher , documentalist and writer**.
Baseline	Helena Maleno Garzón ... es **un defensor** de los derechos humanos **español**, **periodista**, **investigador , documentalista y escritor**.
+PreSent	Helena Maleno Garzón ....es **una defensora española** de los derechos humanos, **periodista, investigadora , documentalista y escritora**.

Table 5.4: Baseline vs PreSent examples from GeBioCorpus.

5.1.7 Conclusion

In this work, the primary goal is to study whether we can exploit contextual or external information, PreSent and SpeakerId methodologies, in which we add either the previous sentence or the gender tag, to help mitigate gender bias in NMT systems.

It has been shown that PreSent methodology allows more accurate translations and resolves ambiguous names. Furthermore, the PreSent methodology achieves the best performance regarding gender translation. This conclusion remains valid even in the case of making inferences without the previous sentence information, mainly when applied on WinoMT, achieving the highest accuracy of 61%.

While other researchers debate that removing the gender information will be beneficial for some tasks [Elazar and Goldberg, 2018], which have to deduce decisions without considering gender. We show that the SpeakerId methodology, which adds the gender information as a tag at the beginning of the source sentence, can help remove the speaker's ambiguity and give better translations from a neutral language to a gendered language. However, the improvement of SpeakerId methodology gets surpassed by PreSent methodology, implying that adding more context achieves better performance.

	Named Entity Disambiguation
Source	Bianca Maria Piccinino ... is an **Italian writer , journalist and television hostess**.
Baseline	Bianca Maria Piccinino ... es **un escritor italiano , periodista y centro** de televisión.
+SpeakerId	Bianca Maria Piccinino ... es **una escritora italiana , periodista y anfitriona** de televisión.

Table 5.5: Baseline vs SpeakerId examples from GeBioCorpus.

5.2 Mitigation using Contextual Embeddings and Relaxed Positioning Conditions

5.2.1 Motivation

Neural machine translation (NMT) models struggle to generate gender inflections in translations correctly. This struggle is a consequence of many factors; one of them is due to the current sentence-based schemes, and another factor relies on strict alignment with source tokens.

To approach these two problems, we are motivated to examine two independent approaches; adding contextual-level information and relaxing conditions by removing residual connections. The first approach takes advantage of the Longformer architecture [Beltagy et al., 2020] to add information at the document level (previous and following sentences), which produces more informed translations. The second approach is to relax conditions by removing skip connections in some layers, leading to a less position-dependent and more flexible grammatical structure. The primary purpose is to investigate whether the system benefits more from appending contextual information or excluding information that is not relied on in all scenarios.

5.2.2 Research Questions

For exploring the approaches feasibility in our gender bias problem, we are motivated to answer the following research questions, these questions should be answered in section 5.2.5 and section 5.2.6:

1. Is it helpful to embed contextual information with the sentence embeddings regarding translation and gender accuracy?
2. Should we maintain the strict alignment between inputs and outputs ensured by the skip connections to preserve the translation quality?
3. What should be our preference when choosing between the two independent techniques?

5.2.3 Proposed Methodology

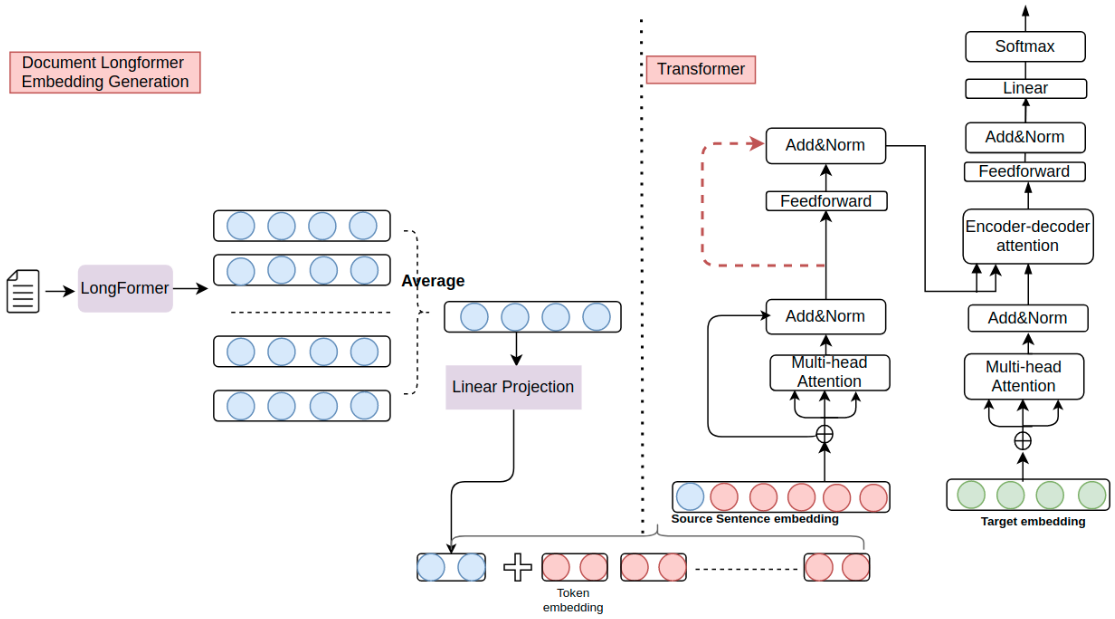


Figure 5.2: Proposed architecture.

This section describes our system, mainly the two contributions: adding Longformer document embedding and relaxing residual connections.

Longformer Contextual Embedding (LF). [Beltagy et al., 2020] presented Longformer, a modified transformer architecture with a self-attention operation that scales linearly with the sequence length, making it versatile for processing long documents. Longformer proposes an attention mechanism that scales linearly with the sequence length utilizing a sliding window that attends to a subset of tokens, relying on the importance of local context. The authors also have introduced the dilated sliding window, a variant of the sliding window, which allows each token's attention range to be increased without increasing the complexity. In most NLP tasks, special tokens such as [CLS] and [SEP] are used to attend to the entire sequence, which is not possible with the sliding window. The authors approach this problem by introducing task-specific, global attention on special tokens. The global attention is symmetric, which means that the special token can attend to every token in the sequence, and every token can attend to it.

We use LongFormer for preparing the contextual embedding; which is mainly considered to be Longformer document embedding. We feed full documents to Longformer, enabling the document embedding to contain information from the whole document context, including those sentences that appear after the current translated one. As shown in Figure 5.2, each document is tokenized and passed to the Longformer model. Sequentially, we obtain the document representation embeddings with the size of all tokens. Following [Macé and Servan, 2019], we consider the document embedding to be the averaging of all the tokens of the document, with Equation 5.1, where x is the row representing the token embedding, N is the size of tokens, and k is the document number.

$$Doc_k = \frac{1}{N}\sum_{j=1}^{N} x_{i,k} \tag{5.1}$$

The resulting mean of the embeddings is projected to the projection layer, which is a linear layer with embedding size as the output dimensionality. The resulting projected mean is then concatenated to the token embeddings, representing the input for the encoder of the NMT model.

Modifying Residual Connections (SkipRS). Residual connections are applied in every layer for both the multi-head attention and the feed-forward layer in transformers [Vaswani et al., 2017]. The connections can facilitate the flow of information through layers; they also impose one-to-one alignment between inputs and outputs. As suggested by [Liu et al., 2021a], relaxing this condition of strict alignment can cause the encoder outputs to be less position-dependent and dependent on language-specific alignments. Moreover, setting one or two encoder layers free from the constraint of the positional correspondence enables the encoder to create its own ordering instead of one-to-one mapping with the input. They applied this relaxation in zero-shot translation to see its impact on the translation accuracy in such cases.

We study the impact of relaxing positional constraints, the dashed red residual connection in the encoder in Figure 5.2, in one or two encoder layers. Relaxing the alignment condition can stimulate diverting attention enforcing the attention mechanism to attend to further tokens for context, leading to better gender generation and decreasing gendered correlation with prior data seen in training. We study this impact on translation accuracy and on mitigating gender bias in a bilingual setting.

5.2.4 Experimental Setup

In this section we introduce the datasets used for training our systems and we go through the data filtering process that was applied in each one of them.

Datasets. When choosing the language pair, we chose English-Spanish after accounting that Spanish is a high grammatical gendered language and that it has pronoun dropout features leading to the omission of pronouns. Such features make translation in English-Spanish tasks more challenging.

To have document-level contextual embedding, we have the requirement of document annotation on the training and evaluation dataset. The training and evaluation need to have defined document boundaries; thus, each sentence would be able to have a global document tag number. For the data domain diversity, we chose the News-commentary and TED-talks datasets. For the News-commentary, the training documents were obtained from WMT[3], considering newstest2015 as the valid set and newstest2016 as the test set. In addition, the TED-talks (IWSLT) were obtained from IWSLT16 competition from the wit3 site[4], considering IWSLT16-dev2010 as the valid set and IWSLT16-test2010 as the test set. For the comparative analysis, we used the same filtered data in the LF and SkipRS experiments.

Data Preparation. Data preparation with Longformer can process sequence lengths up to 4,096 tokens. Knowing that the new lines and the beginning of paragraphs represent tokens, we hypothesize that 250 lines per document can be considered an average number for getting less than 4,096 tokens. Therefore, we filtered the documents folder, removing documents exceeding 250 lines. In addition, we filtered the documents that had the wrong language than the specified pair in ei-

[3] `http://data.statmt.org/news-commentary/v14/documents.tgz`
[4] `https://wit3.fbk.eu/2016-01-d`

Corpora	Partition	Original Files	Filtered Files	No. Lines
News	Train	6,845	6,665	248,025
	Valid	99	99	3,003
	Test	52	52	3,000
TED-Talks	Train	1,820	1,795	207,256
	Valid	8	8	887
	Test	11	11	1,570

Table 5.6: Statistics for EN-ES datasets.

ther the target or source. The filtering steps reduced the total documents, as shown in Table 5.6.

To collect the training data from the files in News-commentary and the talks in TED-talks, we kept only the text data, removing any extra information. Each document is tagged with a unique ID so that each sentence in the document is mapped to this ID. The same steps are done for the valid and test files.

Document Representations Preparation. To prepare the document representation, it is tokenized using Longformer Tokenizer and passed to the Longformer Sequence Classifier model, both from huggingfaces[5]. Global attention is applied to the first token in the document sequence, the start token. After averaging the tokens' embeddings, the embedding dimension is transformed using linear projection to be mapped to the embedding dimension of the encoder's input in the transformer. A map of the documents' IDs and their representations is prepared for the training process. These decisions were made after experimenting with different Longformer models, language models, and sequence classifier models. In addition, embeddings of different tokens were included in experiments aside from the average, and the averaging was observed to perform better.

Training Details. The experiments are performed on a transformer [Vaswani et al., 2017], Fairseq[6] library, with 6 encoder and decoder layers. We use 4 attention heads, an embedding size of 512, an inner size of 1,024, a dropout rate of 0.1, and a label smoothing rate of 0.1. We use the learning rate of 0.001 and the inverse square root schedule from [Vaswani et al., 2017], with 4,000 warmup steps. During the training, the mapping between the sentence and its document is

[5] https://huggingface.co/transformers/model_doc/longformer.html
[6] https://github.com/pytorch/fairseq

retrieved and then the documentation is replaced and concatenated to the source tokens.

To relax the positioning constraints, the residual connections of different layers were modified. We modify the residual connections in the middle of the feedforward layer, as advised by [Liu et al., 2021a], in our case layer three and layer four out of the six layers. Several configurations of layers were tested, and we report only the best performing configuration. It is worth noting that configurations using noncontiguous skip connections lead to nontrainable models.

We apply the classifiers as a probing task of the gender information contained on contextual embeddings proposed in sections 4.1 and 4.2 for more results interpretability. For these gender classification experiments, classifiers are implemented using SVM with a radial basis function using the *Sk-learn* [Buitinck et al., 2013] implementation with default parameters using the WinoMT dataset. One thousand random sentences are extracted for training, using the remaining 2,888 sentences as the test. To ensure that the results are not conditioned on sampling, all experiments are performed ten times, reporting their average results.

5.2.5 Results

When treating gender bias in NMT, a tradeoff between gender accuracy and general translation accuracy can occur [Renduchintala et al., 2021]. In this section, we analyze the results regarding both to understand the impact of our approaches on both accuracies. The WinoMT gender bias evaluation framework is widely used among researchers in evaluating gender bias in NMT [Saunders et al., 2020, Saunders and Byrne, 2020, Costa-jussà and de Jorge, 2020, Stafanovičs et al., 2020]. This framework can reveal how systems resolve or amplify gender-stereotyped translations.

Moreover, we used our systems to translate the Google gendered challenge set [Stella, 2021] with many challenging patterns. If increased accuracy is observed, this shows that the systems can overcome some of the problems introduced in the patterns. This set has the advantage of being a natural dataset, showing how the system can affect gendered real-life examples. To apply the Longformer experiments for the WinoMT and Google sets, we created the Longformer representations for the set, considering that each sentence is the document representation for itself. This approach also may be helpful to compare when the system considers a whole document as the contextual embedding versus the sentence itself.

Domain	System	BLEU↑	Acc↑	ΔS↓	ΔG↓	GBLEU↑
News	Base	27.48	46.5	3.7	35.6	32.8
	Base+LF	27.87	45.9	2.9	33.2	33.2
	Base+SkipRS	27.67	46.6	1.0	33.2	33.5
TED-Talks	Base	38.86	45.4	6.2	37.8	28.6
	Base+LF	39.12	45.2	5.6	34.4	29.0
	Base+SkipRS	39.12	45.1	4.0	36.2	28.7

Table 5.7: Results in terms of BLEU and Gender accuracy: Base stands for Baseline, LF stands for the system with longformer representations, SkipRS stands for the system with modified residual connection. GBLEU is the bleu score for google challenge dataset. The results are for the two domains TED and News.

Translation Accuracy (Table 5.7). In both domains, News and TED-Talks, we can observe that our approaches can lead to a consistent enhancement in the translation accuracy, considering the BLEU metric. In the News domain, LF outperforms the baseline by 0.39 BLEU points and SkipRS surpasses it by 0.19 BLEU points. On the other hand, in the TED-Talks domain, the two techniques increase the translation accuracy by 0.26 BLEU points. To guarantee the consistency of this increase regardless of the random initialization, we repeated the experiments three times in each approach in the two domains.

Gender Bias Results (Table 5.7). LF has a tradeoff of the general gender accuracy with the male dominance of translations, LF could achieve weaker preference for predicting male and increased accurate anti-stereotyped translating, sacrificing 0.6% accuracy in total gendered accuracy compared to the baseline. This difference is mainly the difference of correct pro-stereotyped instances in the case of the baseline. When examining the translations, we observed that the difference between WinoMT accuracy in the baseline is higher due to translating fewer instances of pro-stereotyped correctly. In addition, we noticed that the WinoMT framework can give higher accuracy to the correct gender; however, the occupation is not correctly translated. Nevertheless, LF shows an improvement in translating anti-stereotypes by 0.8 in ΔS and less prevalent male-gendered instances by 2.4 ΔG. On the other hand, the superiority of SkipRS in gender accuracy can be observed in the News domain, with gains in all WinoMT metrics. Translating anti-stereotype entities resulted in an enhancement of 2.7 ΔS with a lower prevalence of males in translation by 2.4 ΔG.

In the TED-Talks domain, we noticed improvements regarding the anti-stereotyped translations in LF systems and SkipRS, with differences of 0.6 and 2.2 in ΔS compared with the baseline. The decrease in ΔG in the two systems is a result of less preference for male translations.

Regarding the Google dataset evaluation, LF sustains the same improvement of 0.4 BLEU points, compared to the baseline. On the other hand, SkipRS has a better impact in the News domain, with an improvement of 0.7 BLEU points compared to the baseline. Reaching this, we have answered the first two questions in section 5.2.2 regarding the performance of the two techniques in translation and gender accuracy.

The difference of results between the two datasets may be attributed to the nature of the data; TED-Talks are mainly inspiring talks, which may have less occurrence of multiple professions. Moreover, as shown in Table 5.6, all data in TED-Talks cover only 26.9% of documents compared to News, and the broader coverage of News may lead to better usage of professions and different gendered patterns. In the case of SkipRS, the News domain can benefit more from relaxing the alignment and positioning conditions.

Gender Classification Results (Table 5.8). In this task, we aim to measure the gender information contained in the contextual embeddings of our translation models by training SVM classifiers on determiners and professions tokens. We observe significant differences between domains, especially on tokens that correspond to professions, with a 24.11% accuracy gap between baseline systems. These results may explain the observed differences in ΔS and GBLEU, as models seem to express gender differently according to the domain used for training.

Our modifications show that the LF vector introduces gender information to the system, with classification accuracies 10% higher than the other tokens encoded by the same system. This is consistent in both domains, especially on News, where we observe a more consistent representation of gender between determiners and professions than the baseline system.

Domain	System	Det.	Profs.	LF Vec.
News	Base	50.01	73.35	-
	Base+LF	63.45	65,13	76.81
	Base+SkipRS	47.06	45.10	-
TED-Talks	Base	44.42	49.24	-
	Base+LF	43.15	42.15	55.77
	Base+SkipRS	43.81	42.51	-

Table 5.8: Results in terms of Gender classification accuracy: Base stands for Baseline, LF stands for the system with longformer representations, SkipRS stands for the system with modified residual connection. The results are for the two domains TED and News. Det., Prof. and LF Vec. stand from determiners, professions and Longformer vector representation, respectively.

On the other hand, applying SkipRS reduces the positional information encoded by our systems, making them less reliant on specific tokens associated with gender. The results show this behavior in both domains, reducing the accuracy gap between determiners and professions to less than 2%. Our intuition is that by not focusing attention on specific tokens that encode most of the gender information, the model is bound to attend to longer dependencies, focusing on a broader context and reducing the impact of learned biases. This emphasizes the answers we got from the previous results regarding the first two questions in section 5.2.2.

5.2.6 Conclusion

Throughout this work, we have focused on the impact of adding information to the NMT architecture or relaxing connections in training to better generalize, showing that both approaches can improve our models by addressing different aspects of the problem. We have particularly applied the two different techniques to mitigate gender bias in the NMT transformer model.

We found that LF document embedding in the LF model incorporates gender information, disambiguating gendered professions in WinoMT, and increasing correct translations of the Google gendered challenging patterns. The LF model's main limitation is lacking scalability to all datasets, as it depends on document-level information.

Regarding the SkipRS model, we concluded that positional information allows the system to associate certain patterns with specific tokens in the sentence. Therefore, reducing such information enforces wider cross-attention, attending to the context of the sentence instead of stereotypical words. However, strict positioning may be essential for specific pairs of languages, whose nature demands such positioning.

We conclude that both techniques help mitigate gender biases, and we advise making the choice between the techniques while taking into account the nature of languages and datasets. Now, we reach to answer the last question in section 5.2.2 regarding how to decide between the two techniques.

5.3 Final Thoughts on the Chapter

This chapter investigates ways to mitigate the effect of gender bias in downstream applications; NMT task. We can conclude the following from our studies:

- Gender bias is a multi-faceted problem that is challenging to mitigate in a downstream application. However, a series of actions and considerations can help mitigate the amplification of gender bias in downstream applications, which is also helpful for fairer NLP applications.
- Mitigation gender bias can be defined differently from one task to another. This results from different effects of gender bias in different downstream applications. Therefore, defining how to mitigate a task can help approach the problem differently.
- Additional context can enable the system to understand the gender better and resolve the ambiguity related to genders in many cases, mainly translating from a neutral gendered language to a high grammatical gendered language.
- Depending on the nature of datasets and languages, mitigating bias does not always rely on adding features or information. Rethinking the model and relaxing some conditions that do not affect the general performance can lead to the same or better effect on mitigating the amplification of bias.

6 Towards Creating Balanced Datasets

To certify and qualify the existence and scale of gender bias across several languages, researchers have been working on dedicated benchmarks for this purpose. In this chapter, we are proposing two different datasets. The first is a speech version of WinoMT [Costa-jussà et al., 2022b], especially for evaluating gender bias in speech translation systems. The second proposes a toolkit for building training and evaluation multilingual datasets, balanced in gender per occupation [Costa-jussà et al., 2022a] for training and evaluation. Besides, we are presenting a balanced multilingual dataset for evaluation.

6.1 WinoST challenge set

6.1.1 Motivation

Biases have been shown in the NMT task when translating from neutral or less grammatical to high grammatical gendered languages [Stanovsky et al., 2019, Saunders and Byrne, 2020]. Additionally, Automatic Speech Recognition (ASR) has demonstrated biases having a higher error rate for female voices than males [Tatman, 2017]. Speech Translation (ST) intersects ASR and MT tasks, perpetuating biases from both tasks. Therefore, the problem of gender bias in the case of ST is more challenging.

Such a challenge motivates us to benefit from the WinoMT [Stanovsky et al., 2019] evaluation protocol in favor of ST task, allowing us to measure how biased our

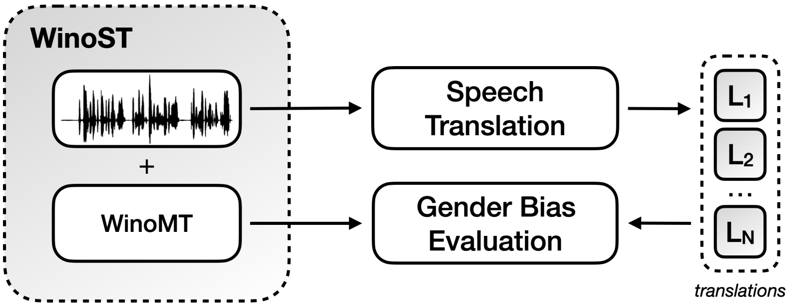


Figure 6.1: WinoST evaluation block diagram for speech translation.

ST systems are. This evaluation protocol enables evaluating from English to any language, complementing the work of MuST-SHE [Bentivogli et al., 2020], which has three language pairs for evaluation. Another difference, MuST-SHE contains naturally occurring gender phenomena; WinoST is a synthetic challenge set. This difference has several implications, not only in terms of the size of the resources themselves (generating synthetic data is somehow easier than collecting them in the wild) but also in terms of their applicability in realistic evaluation settings. Both types of resources are valuable and much needed.

6.1.2 Proposed Gender Evaluation Set

WinoST is the speech version of WinoMT, recorded in off-voice by an American female speaker, and consists of 3, 888 speech audios in English. By nature, sentences from WinoST contain information in the utterance content, not in gender information in the speaker's voice. An example of these sentences is *The developer argued with the designer because she did not like the design.*, where *she* refers to *developer*, meaning that the *developer* is actually a female.

WinoST serves as an input of the ST system to be evaluated, and the output text of the systems follows the same evaluation protocol as WinoMT. Figure 6.1 shows the block diagram of this procedure. As a side-product, and not shown in the figure, WinoST can also be used as a challenge set for evaluating ASR gender bias.

Further technical details on WinoST are reported in Table 6.1, including the number of files, total hours/words, audio recording, and format. The voice mastering process we applied to the recordings includes dynamic voice processing, broadcasting, equalization, and filtering. WinoST is available under the MIT License[1] with the limitation that recordings cannot be used for speech synthesis, text to speech, voice conversion, or other applications where the speaker's voice is imitated or reproduced.

# Files	$3,888$
# Hours	~ 6
# Words	$\sim 50,500$
Audio format	WAV (48 KHz, 16-bit)

Table 6.1: WinoST details.

6.1.3 Experiments

In this section, we are describing the first experiments with WinoST. We describe the baseline ST systems we are using and the results we obtain in gender accuracy. We limit our experiments to four language pairs, but WinoST is extendable to any language pair with English as a source language. The only requirement is to have a part-of-speech for the target language.

Data preprocessing. Before training the model, we preprocessed both speech and text data. We extracted 40-dimensional log-Mel spectrograms from the audio files, using a window size of 25 ms and hop length of 10 ms, with XNMT [Neubig et al., 2018].[2] We normalized the punctuation from text data, tokenized it, and de-escaped special characters using the Moses scripts.[3] Furthermore, in the case of transcriptions, we lowercased them and removed the punctuation. We used the BPE algorithm [Sennrich et al., 2016] for encoding translation texts, using a vocabulary size of 8000 for each language but a character-level encoding in the case of transcriptions.

Speech Translation System. We trained a ST system to evaluate its gender bias with the methodology we are presenting. We used an end-to-end ST approach that directly translates the utterance without obtaining the intermediate transcriptions. This approach was introduced by [Bérard et al., 2016], and recently it had

[1] `https://github.com/gabrielStanovsky/mt_gender/blob/master/LICENSE`
[2] https://github.com/neulab/xnmt
[3] https://github.com/moses-smt/mosesdecoder

a growing interest in the research community [Weiss et al., 2017, Vila et al., 2018, Liu et al., 2019b]. The data we used to train it is the MuST-C corpus consisting of speech fragments from TED Talks, transcriptions, and translations into 8 European languages [Cattoni et al., 2021].

The architecture we used is the S-Transformer, a popular adaptation of the Transformer for ST [Gangi et al., 2019]. It applies a stack of convolutions and self-attention layers to process the log-Mel spectrograms extracted from the speech utterances. The two bidimensional (2D) convolutional layers are in charge of capturing local patterns in the spectrogram, in both time and frequency dimensions. Moreover, they reduce the feature maps by four, which is crucial to avoid memory issues when feeding the Transformer with long sequences. Then, the two 2D self-attention layers introduced by [Dong et al., 2018] model long-range dependencies that convolutional layers cannot capture. Finally, the self-attention layers of the Transformer encoder also include a logarithmic distance penalty that biases them towards the local context [Sperber et al., 2018].

Following a common approach, we pre-trained the S-Transformer encoder for ASR to improve the performance of the final ST system, as introduced by [Bérard et al., 2016] and recommended by the authors of the S-Transformer.

System Details. The model we used has two convolutional layers with a kernel size of 3, 64 channels, and a stride of 2. The Transformer has an embedding size of 512, 6 layers at the encoder and decoder, 8 self-attention heads, and a feed-forward network hidden size of 1024. We trained the S-Transformer with an Adam optimizer, with a learning rate of $5 \cdot 10^{-3}$, and an inverse square root scheduler. The training has a warm-up stage of 4000 updates, in which the learning rate grows from $3 \cdot 10^{-4}$. We used a cross-entropy loss with label smoothing by a factor of 0.1. Moreover, a dropout of 0.1 and a gradient clipping to 20 was applied. Furthermore, we generated the outputs with a beam search of size 5. We loaded 8 sentences per update, with a frequency of 64, which supposes an effective batch size of 512. Those audios longer than 14 seconds and sentences with more than 300 tokens weren't used during training.

	Source transcription	[Target lang.] Predicted translation	Phenomenon
1	**The nurse** sent the carpenter to the hospital because **he** was extremely cautious	[ES] **El enfermero** envió el carpintero al hospital porque era extremadamente consciente	The gender of "nurse" is correctly inferred from the coreference.
2	**The construction worker** asked the nurse for an examination because **she** was injured	[ES] **El trabajador de la construcción** le pidió a la enfermera una examinación porque estaba **herida**	"Construction worker" is stereotyped to its male inflection, with a mismatched adjective gender.
3	The construction worker asked **the nurse** for an examination because **he** was the best	[ES] El trabajador de la construcción le pidió a **la enfermera** una examinación porque **él** era **el mejor**	"Nurse" is stereotyped to its female inflection, with a mismatched pronoun and adjective gender.
4	**The farmer** asked the designer what **she** could do to help	[FR] **L'agriculteur** a demandé au designer ce qu'**elle** pouvait faire pour aider	Although the pronoun is translated correctly, "farmer" has a biased translation form.
5	The writer wrote a book about **the carpenter** because **her** story is very moving	[FR] L'écrivain a écrit un livre sur **le charpentier**, parce que **son** histoire est très émouvante	Biased form for "carpenter" with a neutral possessive gender

Table 6.2: Examples of outputs for different sentences from the WinoST corpus. Words in **blue**, **red**, and **orange** indicate male, female and neutral entities, respectively.

6.1.4 Results

This section describes the results of evaluating the ST system on WinoST and its performance in terms of gender. We are also interested in evaluating ASR English transcriptions and perceive if they contain any gender bias.

General ASR and ST Evaluation. We use the standard WER and BLEU measures to report the ASR and ST performance, respectively in Table 6.4. Our results concur with the results in [Cattoni et al., 2021].

Gender Bias Evaluation in ST. Our main objective is to evaluate the accuracy of the systems for each of the language pairs. The high accuracy demonstrates that the

ST	Acc. (↑)	ΔG (↓)	ΔS (↓)
en-de	51.0	1.7	1.5
en-es	45.2	25.7	12.3
en-fr	43.2	13.7	14.5
en-it	37.3	23.6	5.6

Table 6.3: WinoMT Gender Evaluation for four language pairs. Acc.(% of instances the translation had the correct gender)(the higher the better) ΔG notes difference in F1 score between masculine and feminine sentences (the higher the worse) and ΔS notes difference in accuracy between pro/anti stereotypical sentences (the higher the worse).

Language	ASR (WER ↓)	ST (BLEU ↑)
en-de	24.24	17.8
en-es	24.76	21.9
en-fr	23.98	28.2
en-it	24.18	18.3

Table 6.4: WER and BLEU (%) scores for the MuST-C corpus.

system is able to translate the gender of the entities correctly. We also report ΔG and ΔS in Table 6.3. Ideally, these values should be close to 0. High ΔG indicates that the system translates males better, and high ΔS denotes that the system tends to translate pro-stereotypical entities better than anti-stereotypical entities.

The English-to-German (en-de) system has the highest accuracy 51%. This system also shows the minor difference in treating males and females translations (lowest ΔG, 1.7) and the minor difference in the pro-stereotypical and the anti-stereotypical entities (lowest ΔS, 1.5). The surprising behavior comes with the English-to-Italian (en-it) system, which has the lowest accuracy of 37.3%. Still, it performs reasonably towards the anti-stereotypical entities translations, with the second lowest ΔS difference (5.6). However, the system still favors the male translations with a high ΔG difference (23.6). Both English-to-Spanish (en-es) and English-to-French (en-fr) have similar accuracies (45.2 and 43.2, respectively). However, there is a big difference in the ΔS, which is much higher in the case of en-es (25.7), showing a higher bias towards male translations. With these accuracy results, we are showing that the four translation directions present a significant amount of bias, and they are far from approaching gender parity in performance. Moreover, after manually investigating the translation outputs, we observe that some professions are not correctly translated. *nurse* is always translated to the female version in en-es and en-it. Similarly, *developer* is always translated to the male version in en-it and en-fr, showing that stereotypes are perpetuated in ST. As illustrated in Table 6.2, many inflection errors can occur due to these stereotyped translations. Example 1 shows an anti-stereotypical co-reference case of 'nurse'. One of the common errors happens when translating the gendered adjective or pronoun correctly according to the context while referencing the wrong gendered stereotyped profession, as shown in examples 2 and 4. Another common problem is mismatched pronouns; the translation of the noun contradicts the profession's translation due to biased translation in one of them, e.g., example 3. Example 5 shows a biased translation with a neutral pronoun.

Gender Bias in ST vs MT. Using the S-Transformer, the gender accuracy in the four languages is lower than the reported accuracy of MT commercial systems in the original WinoMT paper [Stanovsky et al., 2019]. The best reported accuracies from commercial systems reached 74.1% in en-de, 59.4 % in en-es, 63.6% in en-fr, and 42.4% in en-it, while in ST case, it is lower for all language pairs as shown in Table 6.4.

This may be since ST is much more challenging than MT, and lower system performance implies higher biases. This big gap is reduced when comparing in terms of ΔG and ΔS. In this case, ST becomes closer to MT (when comparing in absolute terms), showing even better results in: ΔG for en-it (in MT, 27.8); ΔS for en-de (in MT, 12.5) and en-it (in MT, 9.4).

Gender Bias Evaluation in ASR. ASR systems have demonstrated gender biases for female speakers outputs [Tatman, 2017]. However, gender bias associated with the context has not been studied in ASR yet, and WinoST allows this analysis. We may expect that ASR is less prone to show gender bias in contextual patterns because of the nature of the task, which inherently combines the purpose of acoustic and language modeling. The acoustic part does not consider long context information, but it tends to benefit from local context information [Sperber et al., 2018]. However, the language modeling part considers the long-range context, and thus it may induce bias [Bordia and Bowman, 2019, Basta and Costa-jussà, 2021b].

For further analysis of employing WinoST for ASR gender bias evaluation, it is required to distinguish between the gender's errors in transcriptions and the general ones. Therefore, we computed the global accuracy in WinoST for the ASR best system in Table 6.4, en-fr, and got a 74.5% accuracy. However, this global accuracy includes 680 misspelled professions. Discarding these misspelling errors, we obtained a 98.72% accuracy in predicting pronouns, showing that the amount of gender bias at the context level is relatively low in ASR.

6.1.5 Conclusions

This thesis presents a new freely available[4] challenge set for evaluating gender bias in ST. This challenge set, WinoST, can benefit from the evaluation protocol widely used for MT. Our set is only based on evaluating the gender inaccuracies in translations in ST systems, mainly relying on the gender information extracted from the context and not from the audio signal.

We used an S-Transformer end-to-end ST system and evaluated their accuracy in terms of gender bias with this new challenge set. Results show that gender accuracy is much lower for ST than for MT, but we have to consider that ST also has a lower quality than MT, which may impact the gender translations as well. Finally, we show that ASR can exhibit a slight gender bias at the contextual level.

WinoST shares similar limitations as WinoMT, which is the fact of using a synthetic challenge set. Having a synthetic set is positive because it provides a controlled evaluation and is also harmful because we might be introducing some artificial biases. Therefore, further work could find templates in the wild transcriptions (with parallel speech utterances) that hold the valuable patterns designed in WinoMT, following [Levy et al., 2021].

[4] Freely available in Zenodo (10.5281/zenodo.4139080) `https://zenodo.org/record/4139080#.YlQo3bxBxH4`

6.2 GENOCC Toolkit: Building Real-world Multilingual Balanced Parallel Data

6.2.1 Motivation

There is an urge to generate different evaluation sets to tackle judging our systems regarding gender bias. Currently, mechanisms to extract balanced datasets are limited to narrow languages. Additionally, minimal balanced test sets concerning gender within occupation exist. Such facts motivate us to create the GENOCC toolkit customized according to research and development needs regarding languages and gender definition (beyond binary), to create training and evaluation datasets. Our motivation to create balanced datasets for training is that previous works have shown that fine-tuning with balanced data mitigates gender bias. While creating balanced datasets for evaluation aligns to further progress in responsible artificial intelligence evaluation[5].

Our extracted datasets are balanced in gender within occupations because they have the same number of Wikipedia articles in all genders under consideration for each particular occupation entity, with the same total number of sentences per gender for each occupation. For example, for the case of the *politician* occupation, if limiting to binary categorization of gender (male and female), we would have N number of articles for female politicians and N for male with M male sentences and M female sentences.

6.2.2 Proposed Data Collection and Curation Methodology

Our proposed methodology (Figure 6.2) involves multiple stages; data collection, mining strategy, data alignment (based on previous work [Schwenk and Douze, 2017]) and balancing.

[5] https://ai.facebook.com/blog/facebooks-five-pillars-of-responsible-ai

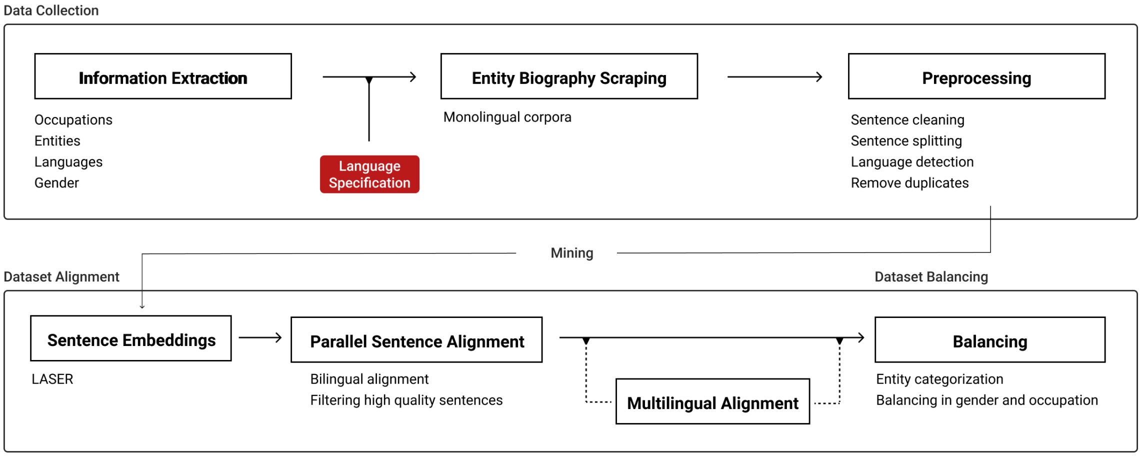


Figure 6.2: Pipeline overview. First step is data collection (top), which includes collecting and preprocessing data. Within this step, we have to define which languages our final dataset will contain. Followed by mining then dataset alignment and balancing (bottom), with an optional step for multilingual alignment in case of need.

Data Collection On the one hand, our data are collected from Wikidata[6], a well-known knowledge base known for its quality. Wikidata is a project that maintains its data quality by monitoring methods and evaluations to guarantee that it suits users' needs. Briefly, our data contains a set of people (*from now on* entities) with their occupation(s), gender, and Wikipedia links in all available languages. On the other hand, our monolingual data are extracted from Wikipedia[7], similar to [Costa-jussà et al., 2020, Stella, 2021]. Monolingual data are related to the textual data of the entity's biography for one language from Wikipedia.

Information Extraction. In this first step, we extract data from Wikidata. Mainly, our data relate a set of entities with their working occupations, gender, and biographies from Wikipedia in all available languages. Figure 6.3 shows a schema of the information extraction procedure, described as follows:

1. We extract all the occupations present in the knowledge base.
2. For each occupation, we gather the data of every entity that works in the related occupation.
3. For each entity from the previous step, we determine the gender information and related Wikipedia links in all available languages (biographies).

[6] https://www.wikidata.org/wiki/Wikidata:WikiProject_Data_Quality
[7] https://www.wikipedia.org

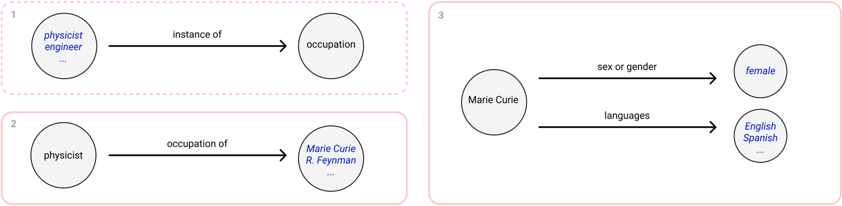


Figure 6.3: Extraction schema. Each step is depicted in a triplet format: ⟨subject,predicate,object⟩. *Blue* (italics) information is the information extracted at each step. For each step outlined with a dotted rectangle (−−), the information extracted is the subject; otherwise, the information extracted is the object.

Afterward, we consider cleaning details. We remove the occupations that do not have related entities and entities that lack gender information. We checked the language tags in each entity, which leads to the removal of entities that do not have an ISO language code[8] nor special language codes from Wikimedia[9].

Entity Biography Scraping. At this step, we specify the languages included in our final dataset. We consider different criteria in choosing the languages, conditioned by the number of Wikipedia biographies, the family of languages, and the number of multi-languages intended. Consequently, the size of the corpus at the end of the pipeline will be heavily influenced by the type and number of selected languages. For instance, high-resource languages are more likely to have more biographies; nevertheless, a multilingual dataset with high-resource languages may significantly reduce the number of sentences compared to a bilingual dataset, which may not be as detrimental. As a result, there are implicit trade-offs between high-resource and low-resource languages and between bilingual and multilingual datasets. By specifying a set of ISO language codes to the system, we scrape all the monolingual data from the corresponding Wikipedia biography for entities with a link for all the given languages. We can scrape Wikipedia directly because we have previously gathered the Wikipedia biographies' link for each entity in all available languages.

Preprocessing. As follows, we describe the steps used to preprocess monolingual data for alignment afterward.

- **Sentence cleaning.** Regex expressions are applied to remove the information

[8] http://www.lingoes.net/en/translator/langcode.htm

[9] https://meta.wikimedia.org/wiki/Special_language_codes

between brackets and parenthesis, which is mainly related to phonetics, dates, and references in Wikipedia articles.

- **Sentence splitting.** Monolingual data are split into sentences; consequently, the sentences are prepared for alignment individually.
- **Language detection.** Sentences are checked by a language detection module to exclude those that are not from the corresponding language, as Wikipedia pages can mix sentences from several languages. This step ensures that all sentences are from the intended language.
- **Remove duplicates.** Duplicated sentences are removed, ensuring we have unique sentences for each entity.

Dataset Alignment and Balancing Our mining strategy is the process of preparing the data in the way considering each entity's data individually. Accordingly, the Wikipedia data should be prepared for the purpose of mining each entity individually. This facilitates the next steps to perform the sentences embeddings of each language independently and then computes the candidates between a source language and a target language on each entity individually (parallel alignment). Then multilingual alignment is performed to obtain the final set of aligned sentences between all chosen languages.

Sentence Embeddings. We obtain sentence embeddings through a multilingual sentence encoder based on the architecture [Schwenk, 2018] in which semantically similar sentences are closer to each other, independent of their language [Schwenk et al., 2021]. This allows for a common ground for sentences from different languages. It facilitates the use of the multilingual encoder to extract parallel sentences relying on distance-based metrics to perform the next step parallel sentence alignment.

Parallel Sentence Alignment. Parallel sentence alignment follows the margin-based criterion introduced in [Artetxe and Schwenk, 2019] as a metric to execute the nearest neighbor. The margin criterion between two candidate sentences x and y is defined as the ratio between the cosine distance between the two embedded sentences and the average cosine similarity of its nearest neighbors in both directions:

$$margin(x, y) = \frac{\cos(x,y)}{\sum\limits_{z \in NN_k(x)} \frac{\cos(x,z)}{2k} + \sum\limits_{z \in NN_k(y)} \frac{\cos(y,z)}{2k}},$$

where $NN_k(x)$ denotes the k unique nearest neighbors of x in the other language and $NN_k(y)$ denotes the same for y. This alignment step allows for getting parallel bilingual candidates which are sorted according to their margin scores, and a threshold is applied to get the desired quality of parallel sentences. This step is performed on each pair of languages independently. If a multilingual alignment approach is required, we consider one language as the target (pivot) to all other languages, and perform this step for each pair of languages, e.g., for a case of multilingual alignment of English, French and German, we can consider English as the target language and perform parallel alignment for French-English, and German-English.

Multilingual Alignment. The steps are: (1) parallel alignment of all language pairs with one pivot language, (2) intersection of all common sentences in these aligned parallel sentences according to a certain threshold for getting the desired quality. To extract multilingual parallel sentences in more than two languages, i languages, we use a greedy approach with pivot language L_1. We detect all the parallel sentences in pairs L_1–L_i and then extract the intersection of sentences between the language pairs mainly depending on the similar L_1 sentences in all pairs (same pivot sentences).

Balancing

At this point, we have multi-parallel sentences corresponding to different entities annotated with the corresponding gender and occupation.

Entity Categorization. Entities could have more than one occupation. We categorize entities by the number of occupations they include. Such categorization informs us about the multiplicity of occupations and their corresponding entities in our data. This information enables the choice of categories intended for balancing later. For example, category one represents the entities that have one occupation.

Balancing in Gender within Occupation. The output of this step will be a balanced set with respect to numerous occupations. Each occupation will be represented by a similar number of gender entities, and the total numbers of sentences per gender will be the same. There might be an occupation's name that refers to a single gender, but the data within this occupation will be balanced regarding all the genders (e.g. actor). During balancing, we balance each category (i.e., number of occupations) separately and incrementally, for example, balancing category one (i.e., one occupation) followed by category two (i.e., two occupations). Balancing higher categories (i.e., with multiple occupations) means excluding occupations that already exist in previous lower categories. For example, in the case of extracting category two (i.e., two occupations), if we have an entity with occupations of doctor and politician, this entity is excluded if either doctor or politician or both were included occupations in category one. This guarantees that the balancing of the new occupations is not conditioned by balancing the occupations of the last category.

At this point, we continue with gender balancing. We compute the number of gender entities for each occupation and the sum of each gender sentence from all the corresponding entities. For each occupation, balancing will be carried out according to the minimum number of entities and sentences. For example, for binary gender (male and female), if an occupation has four females and five total related sentences and seven males and ten total sentences, then four entities and five sentences are the maximum intended values for each gender in this occupation. Consequently, this step excludes occupations that have one gender representation (female or male). We prioritize the male and female entities that have a similar number of sentences with a higher degree of similarity among languages (i.e., this similarity is based on the margin criterion defined in this section. Details are illustrated in Algorithm 1 for the case of using binary gender as we do in our use cases.

6.2.3 Use-case Study

In this section, we report the experimental details of our methodology by including details on two use cases (high- and low-resource languages). We provide details for data collection, dataset alignment, and balancing.

High-resource Languages The top-7 languages with the largest number of entities are English, German, French, Spanish, Russian, Italian, and Arabic. Among these top languages, there are four linguistic families, Germanic, Latin, Slavic, and Semitic, and we choose one language representing each family. We are limiting gender to binary (male and female), relying on the tagged category of the perceived gender from our sources. We extract multiparallel data among the high-resource languages that cover different linguistic families, including Semitic (Arabic), Germanic (English), Slavic (Russian), and Latin (Spanish). The motivation of this use case is to have a balanced dataset in languages that are well-studied in the community. Nonetheless, this dataset can also be used in conjunction with other existing benchmarks that may contain occupational stereotypes or unbalances in gender. Hereinafter, we alternatively refer to this use case either as the high-resource or en-es-ru-ar use case. The latter specifically mentions the covered languages.

Algorithm 1: *Balancing gender within occupations.*

```
Input  : $U_{dic}$ // An unbalanced dictionary containing the information about occupations,
    entities in each gender, and aligned sentences with their alignment score.
Output: $B_{dic}$ // A balanced dictionary where each occupation has the same number of sentences
    in balanced male and female entities.
$Occs$; // A list of occupations in $U_{dic}$.
$Em_i$; $Ef_i$; // A list of male and female entities with $i$th occupation, respectively.
$Sm_i$; $Sf_i$; // Number of sentences in $Em_i$ and $Ef_i$, respectively.
$B_{dic} = \{\}$; // Initialize the empty dictionary to store the balanced information.
for $i \leftarrow 0$ to $len(Occs)$ do
    if $len(Em_i) == len(Ef_i)$ then
        • Balance the entities from $Em_i$ and $Ef_i$ such that $Sm_i$ is equal to $Sf_i$ and update $B_{dic}$;
    else
        $Emin = \min(len(Em_i), len(Ef_i))$;
        if $Emin == len(Em_i)$ then
            • Select only $Emin$ female entities with high-quality sentences from $Ef_i$;
            • Balance the entities from $Em_i$ and $Ef_i$ such that $Sm_i$ is equal to $Sf_i$ and update $B_{dic}$;
        else
            • Select only $Emin$ male entities with high-quality sentences from $Em_i$;
            • Balance the entities from $Em_i$ and $Ef_i$ such that $Sm_i$ is equal to $Sf_i$ and update $B_{dic}$;
        end if
    end if
end for
return $B_{dic}$
```

6.2.4 Implementation Details

We extract data from Wikidata using a Python SPARQL wrapper. For entity biography scraping, we implement an algorithm that works with Beautiful Soup[10], whose purpose is pulling data from HTML content. After that, the following preprocessing techniques are implemented to improve the outcome of our collection process:

- We use regex expressions to clean the collected monolingual corpora.
- We use the nltk[11] sentence tokenization package[12] to split sentences across all languages except Arabic, which uses a sentence splitter wrapper[13] for CoreNLP[14].
- We apply language detection to sentences using Compact Language Detector 3[15], which can identify up to 108 languages, to remove sentences that are not labeled with the appropriate language.
- We also remove sentences repeated within a document; to ensure that sentences within a biography are unique.

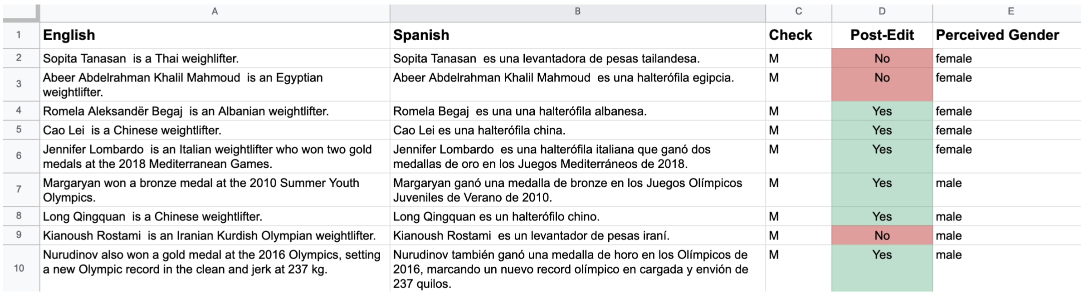
We prepare the text for each entity individually. Then, to execute parallel sentence alignment, we utilize LASER [Schwenk and Douze, 2017], which provides multilingual sentence embeddings. After embedding the sentences, the aligned parallel sentences in a language pair are computed using the distance in the embedding space. The candidates are sorted according to the order of the similarity between sentences. When aligning multiple languages, we consider English to be the pivot language.

6.2.5 Postediting

Given the high-resource and low-resource datasets, we postedit them to have curated datasets that can be used for evaluation in machine translation. We use English as the anchor language and distribute sentences in a spreadsheet (see Figure 6.4) for native annotators in non-English languages in which English is the second language.

Each language set of sentences was split into 2 to 4 subsets addressed by different annotators. The annotation guidelines are as follows:

[10] https://www.crummy.com/software/BeautifulSoup/bs4/doc/
[11] https://www.nltk.org
[12] https://github.com/Mottl/ru_punkt
[13] https://github.com/chaojiang06/CoreNLP_sentence_splitter
[14] https://stanfordnlp.github.io/CoreNLP/index.html
[15] https://github.com/google/cld3



	A	B	C	D	E
1	**English**	**Spanish**	**Check**	**Post-Edit**	**Perceived Gender**
2	Sopita Tanasan is a Thai weighlifter.	Sopita Tanasan es una levantadora de pesas tailandesa.	M	No	female
3	Abeer Abdelrahman Khalil Mahmoud is an Egyptian weightlifter.	Abeer Abdelrahman Khalil Mahmoud es una halterófila egipcia.	M	No	female
4	Romela Aleksandër Begaj is an Albanian weightlifter.	Romela Begaj es una una halterófila albanesa.	M	Yes	female
5	Cao Lei is a Chinese weightlifter.	Cao Lei es una halterófila china.	M	Yes	female
6	Jennifer Lombardo is an Italian weightlifter who won two gold medals at the 2018 Mediterranean Games.	Jennifer Lombardo es una halterófila italiana que ganó dos medallas de oro en los Juegos Mediterráneos de 2018.	M	Yes	female
7	Margaryan won a bronze medal at the 2010 Summer Youth Olympics.	Margaryan ganó una medalla de bronze en los Juegos Olímpicos Juveniles de Verano de 2010.	M	Yes	male
8	Long Qingquan is a Chinese weightlifter.	Long Qingquan es un halterófilo chino.	M	Yes	male
9	Kianoush Rostami is an Iranian Kurdish Olympian weightlifter.	Kianoush Rostami es un levantador de pesas iraní.	M	No	male
10	Nurudinov also won a gold medal at the 2016 Olympics, setting a new Olympic record in the clean and jerk at 237 kg.	Nurudinov también ganó una medalla de horo en los Olímpicos de 2016, marcando un nuevo record olímpico en cargada y envión de 237 quilos.	M	Yes	male

Figure 6.4: Spreadsheet for annotators. Complete example for the Spanish language.

Given a sentence in English (first column (A)) and Arabic/Russian/Spanish (second column (B)), do the minimum number of edits (in the same column (B)) to the Arabic/Russian/Spanish sentence to match the English sentence. If the Arabic/Russian/Spanish sentence contains more information than the English sentence, then remove it. If the English sentence contains more information, add this information translated into Arabic/Russian/Spanish. Mark each sentence that is edited (add M to the third column (C)). If you do not know how to postedit without changing the meaning, mark "NM" in the third column (C). Mark if postediting was necessary (yes) or not (no) in the fourth column (D). Pay special attention to gender; if it is ambiguous, please check the entity-perceived gender from the fifth column (E).

After annotating the entire dataset in each language, there was an additional annotator for each language who reviewed the entire set. Annotators were volunteers, and they are acknowledged at the end of this work.

6.2.6 Experimental Results

In this section we report the experimental results of the use cases that we proposed in the previous section. We report data statistics for our datasets and machine translation results using standard state-of-the-art models.

6.2.6.1 Data Statistics

Entities per language. Figure 6.5 shows the number of entities for each language and gender. English is the language with the highest number of entities, and Arabic is the one with the lowest. In general, there is a large difference between gendered representations. In particular, we observe very few entities that are not male or female. The figure also shows that all languages have three times more male representations on average than female representations.

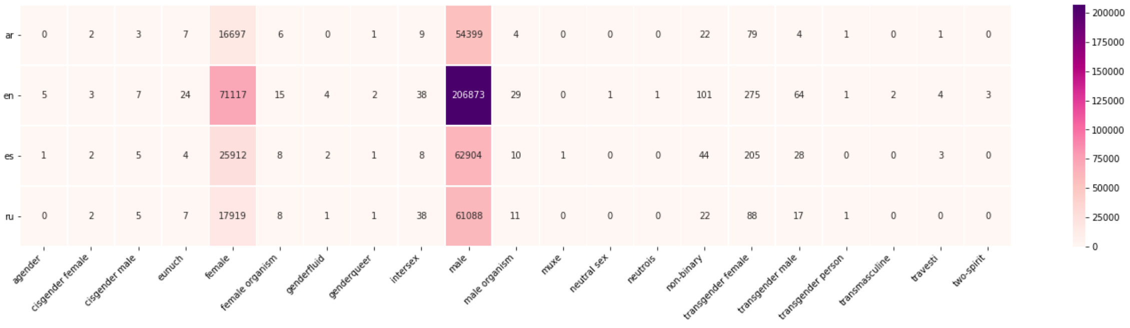


Figure 6.5: Distribution of entities' gender across languages.

Number of Entities, Occupations and Sentences through the Pipeline. Table 6.5 shows the number of entities, occupations, and sentences at the different stages of our pipeline (Figure 6.2): entity biography scraping, preprocessing, alignment (multilingual or bilingual), and balancing. These statistics show how the number of entities and occupations is reduced at each step for our user cases. They show that alignment and balancing steps have a great impact in reducing the number of occupations and entities, illustrating the reason for losing entities and corresponding sentences when choosing more languages to align and balance. The numbers of sentences can only be provided from the alignment step onward, since sentences per language can only be noted individually before this step. As predicted in section 6.2.2, among the balanced occupations in the high-resource use case, we found occupations' names characterizing only the male gender, such as pornographic actor or monarch.

		Entity biography scraping	Preprocessing	Alignment$^{(*)}$	Balancing
en-es-ru-ar	entities	15421	14635	2436	286
	occupations	644	256	203	59
	sentences	-	-	6732	524

Table 6.5: Evolution of the number of entities and occupations through the pipeline. $^{(*)}$Multilingual alignment for the en-es-ru-ar (high-resource) use case.

Entity Categorization. Figure 6.6 shows the number of entities (y-axis) with different numbers of occupations (x-axis). This categorization is used to balance occupations. In the case of high-resource languages, we used entities with three occupations at most.

Details on Post-editing. Figure 6.7 (left) shows the number of postedited sentences from high and low-resource languages. Note that sets from Spanish, Russian and Arabic are comparable. We see that for high-resource languages, the sentences that need to be post-edited is around 50%. Figure 6.7 (right) shows the Translation Edit Rate (TER)

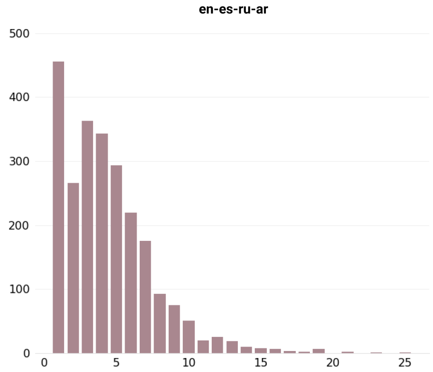


Figure 6.6: Number of entities with different amounts of occupations regarding our use-cases: en-es-ru-ar (high-resource).

results, computed with Huggingface's Whitespace tokenization. Results are coherent with the previously post-edited sentences. This TER results give us an idea of the amount of error that our toolkit can introduce when extracted data is not post-edited. The values that we observe in TER, which are not superior to 30% in all cases, show that the amount of post-edition is moderately low. Moreover, assuming this amount of error, our toolkit can be considered for training purposes without requiring post-edition.

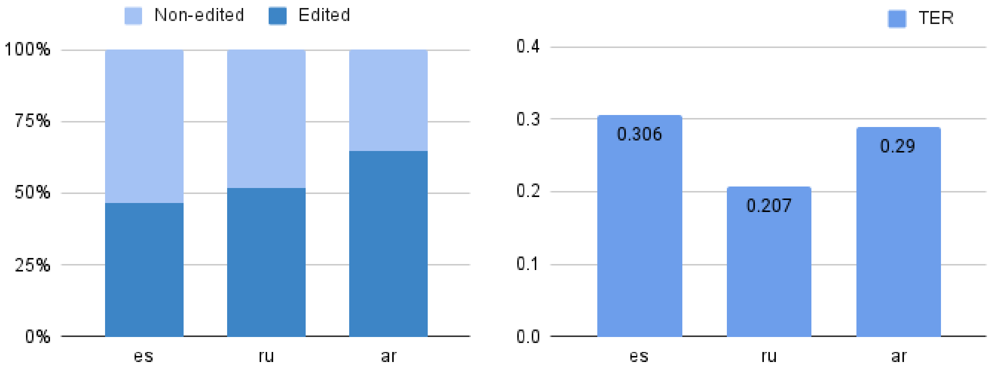


Figure 6.7: Percentage of post-edited sentences per language (left). Translation Edit Rate (TER) per language (right)

6.2.6.2 Machine Translation

System Description and Implementation. To evaluate our dataset, we use the downstream task of machine translation (MT). We used three multilingual models that include the languages from our use cases: M2M_100 [Fan et al., 2021], mBART50_m2m [Tang et al., 2020] and Opus-MT [Tiedemann and Thottingal, 2020]. These systems are transformer-based models [Vaswani et al., 2017], and they use SentencePiece-based segmentation [Kudo and Richardson, 2018]. M2M_100, supporting translation between any direction for 100+ languages, is the model that includes many-to-many supervised training covering thousands of language directions. mBART50_m2m supports translation between any direction for 50+ languages, but it is only trained with supervised translation from and to English. Opus-MT supports 1200+ translation directions for 150+ languages, implemented with MarianNMT[16]. A part from the characteristics of the Open-MT project is the constant collaboration with the Wikipedia Foundation. For the three systems, we used the default implementation from EasyNMT[17].

Results. Figure 6.8 (top) shows the heatmap of BLEU results for different language pairs and (bottom) the average BLEU across language directions for the high-resource languages use case. In general, we see the best performance for the Opus-MT model, with few exceptions (English-Russian, Spanish-Russian, and Spanish-to-English) on which M2M_100 is better. mBART50_m2m has the lowest performance, especially in directions that do not involve English, which makes sense because it is unsupervised. The best results are obtained when translating to English, and the worst results are obtained when translating to Arabic.

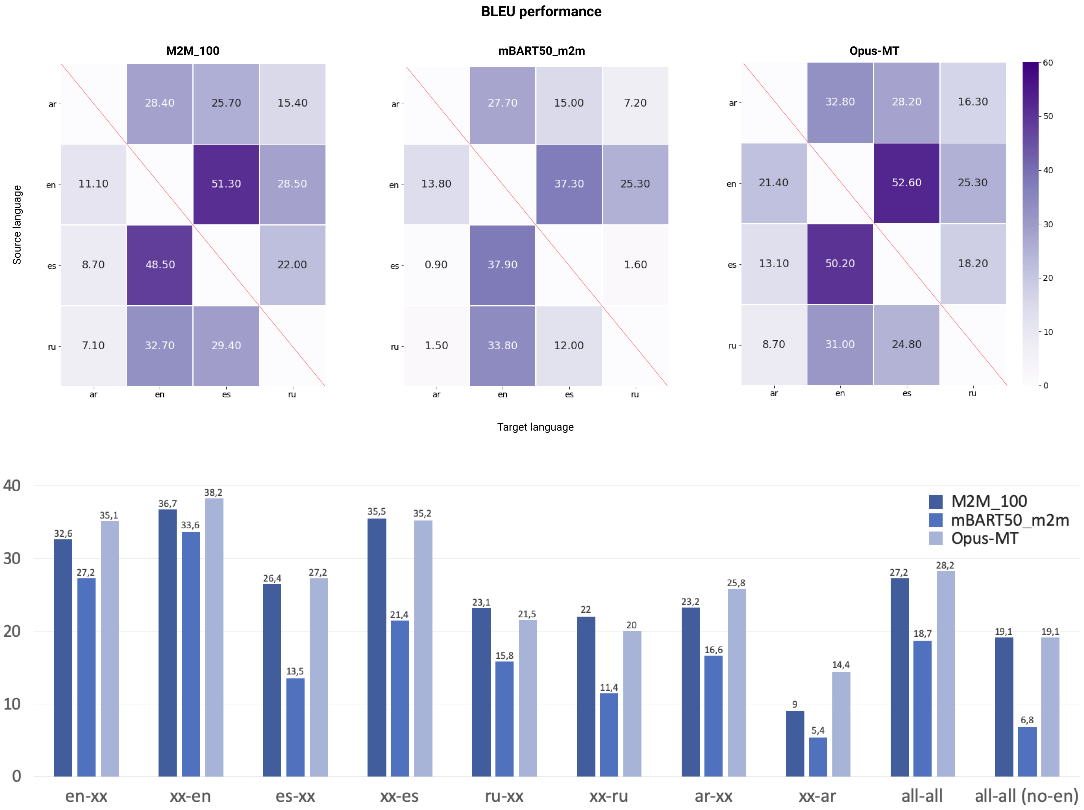


Figure 6.8: High-resource language results. (Top) Heatmap for BLEU scores between languages with M2M_100, mBART50_m2m and Opus-MT models. (Bottom) Average BLEU for all language directions.

[16] https://marian-nmt.github.io
[17] https://github.com/UKPLab/EasyNMT

Figure 6.9 reports the results in terms of BLEU for the two female/male subcorpa from our high-resource use case benchmark. For English and Arabic (both directions) and translating to Russian, the performance is better (or similar) in the male set for all models. In the case of translating from Russian, the performance of the female subcorpus is better than that of the male subcorpus in all models except for the mBART50_m2m model. When translating to Spanish, this improvement only holds for the M2M_100 model.

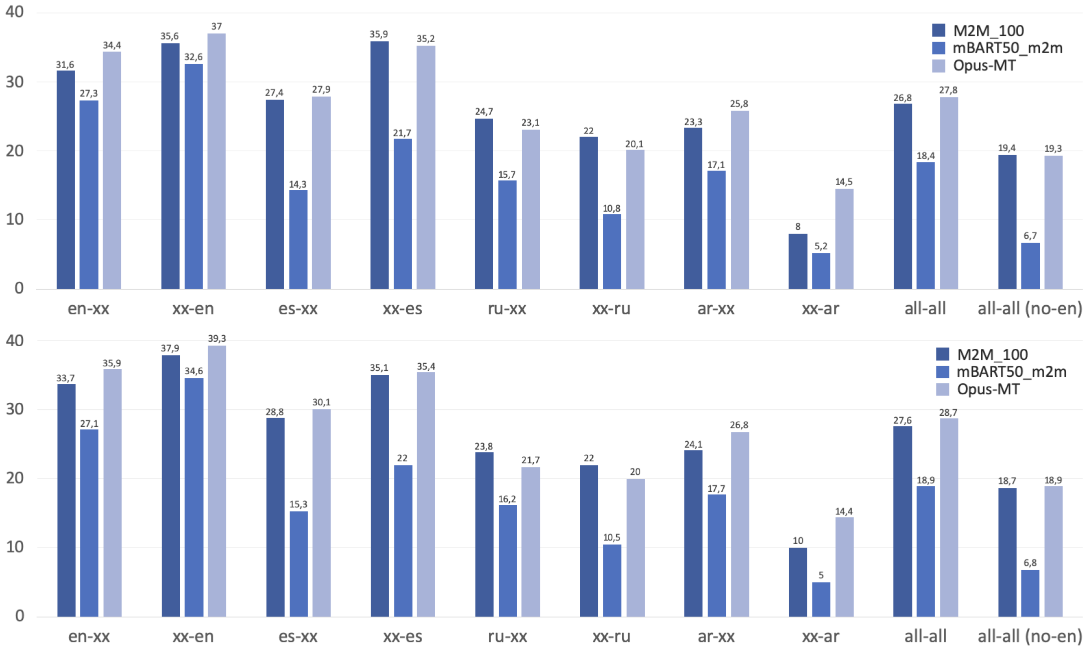


Figure 6.9: High-resource language results. Average BLEU across language directions with M2M_100, mBART50_m2m and Opus-MT models. (Top) Female (Bottom) Male.

Discussion. An intriguing pattern suggests that male English translations impact the overall (overall) performance more than female translations. Figure 6.9 (female (top) and male (bottom)) shows that male translations influence the overall BLEU performance more (Figure 6.8, bottom) since they have the highest BLEU among both genders. However, in the overall case without English, the trend radically alters, with female translations contributing more than male translations to "all-all, no-en" (Figure 6.8 bottom), which is a subset of the all-all dataset. Thus, if we analyze performance with English either on the source or the target side on both genders (Figure 6.9), we found that the difference between the performances of male and female translations using English on either side is the greatest for gender comparisons in any other high-resource language pair, which explains the change in translation performance with and without English. This reveals that the performance of these models is skewed towards male English translations in any direction.

When looking at the translation direction (e.g., language A to all or all to language A), we observe English and Spanish exhibit different behaviors than Russian and Arabic. The former languages exhibit a higher BLEU performance when the languages are on the target side rather than on the source side. For the same languages, on the target side, the performance on the female set tends to be lower than the male set. We hypothesize that English and Spanish have a solid language generation. This strong language generation might be partly because they are written in Latin script, which is shared among many high-resource languages (i.e., Italian, French, and Portuguese). Sound language generation may help achieve higher performance when on the target side. However, it may also contribute to paying less attention to the source sentence, and more attention to the previous words in the target sentence [Ferrando and Costa-jussà, 2021], which may explain why the performance in the female set did not improve. Less attention given to the source sentence and more attention given to the previously generated words in the target can overgeneralize to the most frequent gender, which tends to be male. This suggests that when translating to English, even if source languages (Spanish, Russian, and Arabic) have high morphological information, the performance on the male set is higher than that on the female set. For the latter language set, Russian and Arabic, the performance of the translation direction does not vary as it does in the previous case. This may be because even high-resource languages have a different script, which is not typical of other higher-resource languages. This may explain the poorer language generation.

6.2.7 Conclusions

We propose the GENOCC toolkit to generate monolingual, bilingual, and multilingual balanced datasets in gender within occupations. This toolkit can be customized in languages and gender. In addition, we present a high-resource benchmark. The former includes a multilingual English, Spanish, Russian, and Arabic corpus. Finally, we show experiments using these benchmarks to evaluate state-of-the-art machine translation models. Our balanced sets allow for the analysis of gender performance with standard evaluation methods and without requiring new ones. We provide an accurate analysis of performance behavior for the particular case of binary gender. We conclude that female translations tend to be worse for high-resource languages with a high-quality language generation model. We hypothesize that, in these cases, the model gives less attention to the source words than the target context words, and using the target context may overgeneralize to most frequent patterns (which tend to be male patterns) rather than producing an accurate translation.

6.3 Final Thoughts on the Chapter

In general, we need well-established benchmarks as a means of conducting diagnostic tests for gender bias, to study if the systems have a high positive and low negative predictive value for the presence of gender bias [Stanczak and Augenstein, 2021]. The current research on gender bias in speech translation is limited. Consequently, the community needs to pay more attention to studying, evaluating, and mitigating bias in such systems. This fact motivates us to exhibit more efforts in adapting the WinoMT dataset to speech translation and generating the WinoST dataset, the multilingual evaluation dataset for speech translation. Although this dataset is synthetic, it can still demonstrate the presence of gender bias in a system.

Another problem arises from the current natural or synthetic datasets in only English. Only in the case of MT a few sets are available in the high-resource languages such as Spanish or German [Stanczak and Augenstein, 2021]. This led us to work on a toolkit to generate a natural monolingual, bilingual or multilingual benchmark balanced in gender within occupations. This tool can also be used for creating balanced training data to finetune the systems.

Overall, we strongly encourage further research to establish evaluation benchmarks for the different models and tasks.

7 Conclusion of the Thesis

7.1 Reflections and Insights

The current situation is promising, given research and industry's interest in addressing gender bias. We interviewed scientists[1] with expertise in the field to have different insights into the current and future situation of NLP.

Regarding the current situation, *Nizar Habash* commented with the following: "In earlier efforts on NLP, there was much attention on system accuracy. Now, as the systems are getting more mature thanks to new technologies and advances, there is an opportunity to focus on gender and social bias problems. The priorities are adjusted, and different needs are raised. We are in a better place to start considering such problems and dedicating more efforts to mitigate the biases." We should also comment that the low-resource languages still suffer from a lack of accuracy, which makes studying such issues harder for them. On the other hand, *Eva Vanmassenhove* commented that researchers should be aware of the biases that can happen due to proxiesin the models, and she gave an interesting example of predicting high social standards based on images of people who own dogs. When understanding such proxies of models and their effect on our predictions, we can comprise that impact. *Ryan Cotterell* focused on an essential aspect of the current situation, mentioning that we are currently focusing on English without much attention to other languages. This would solve the bias problem from a Western-centric view. Low-resource languages are not given enough attention for the quality or tackling such issues.

Regarding the future research situation in this area, *Eva Vanmassenhove* mentioned two perspectives that should be considered. More collaboration with experts from social and linguistic fields is highly needed. We need to have an honest conversation with linguists and social psychologists to define gender bias better. We then can understand the real harms and the impact on affected people. The other perspective is more related to the inter-

[1] Acknowledgments to *Nizar Habash*, NYU Abu Dhabi, *Eva Vanmassenhove*, Tilburg University and *Ryan Cotterell*, ETH Zurich.

pretability and controllability of our models. It is considered the facility of understanding what is not comprehended by the machines and what is lacking to make a satisfactory decision. *Nizar Habash* referred to the need to pay attention to uncover implicit user preferences in our systems to optimize serving the user, such as asking the user's pronouns. Briefly, the users should be represented and treated in their desired way. *Ryan Cotterell* revealed the fact that a real solution to the problem would not happen without teaching our generations about the problem and its impact. He quoted this fact: "Academics should keep fostering creative individual people and encouraging them to think differently."

7.2 Main Contributions

In this thesis, we covered three directions toward building fairer NLP models: evaluating and understanding the biases inherited in existing models and mitigating the effect of bias in NLP models (MT task),

Chapter 4 adds new insights in bias detection. We aimed to identify and bias in a multilingual setting given the lack of multilingual settings in literature. We adapted bias evaluation techniques to the contextualized embeddings and extended the evaluation to another language (Spanish) rather than English. We revealed that contextualized embeddings could amplify the bias in the training data. Embeddings represent an important component in any NLP system and the same for the architectures. Therefore, we proposed studying the interpretation of bias in multilingual NMT architectures. We showed that architectures matter even if trained on the same biased data. Some NMT architectures can amplify the bias more than others attributing to the parameters and features shared in models. Regarding more proper evaluation, we found that the existing literature lacks multilingual benchmarks for evaluating the bias in different models. To fill in this gap, we presented different datasets in chapter 6. We built the multilingual evaluation WinoST, the speech version of WinoMT, for the speech translation task in section 6.1. We found that speech translation is less mature than textual translation and has low accuracy, which prevented us from making accurate conclusions regarding gender bias. Furthermore, we proposed a toolkit for generating monolingual, bilingual, and multilingual datasets balanced in gender within occupations in section 6.2. This tool should be helpful in the research community in multiple tasks.

The next step was to start exploring methods to mitigate gender bias in models; we explored how to mitigate the bias in the NMT task in chapter 5, seeing that it is a multifaceted, challenging, and affecting a wide range of users. We adapted two methodologies for our problem; increasing context and adding the speaker tag in section 5.1. In section 5.2, we proposed two new methodologies, and we aimed to study each individually, adding the contextual embeddings and relaxing the positioning by removing skip connections. We showed that adding contextual information through a sentence or a contextual embedding helps have more gender information. We also demonstrated that we should always consider looking at the architecture and figuring if relaxing a condition may give better gender results preserving the model performance.

7.3 Final Thoughts and Closure

Our thesis is concerned only with binary gender, and this is a significant limitation in our work and most of the current research in tackling gender bias in NLP systems. Unfortunately, there are few resources to study other genders, limiting the studies. Moreover, we should mention that the thesis is tackling the gender bias problem from a linguistic perspective more than a social perspective

The problem of gender bias is more than solving a problem in methodology or architecture, and it has been rooted in the communities for a long history. The cultural and languages difference reinforce such biases. However, we are in a better place nowadays. There is more awareness in the young generations, and they have already started to learn and address this problem.

Generally, awareness of gender bias and fairness issues has been cultivated recently. Wider research has been carried out to tackle different perspectives of the problem. Full workshops[2] and special tracks in conferences[3] are dedicated to this. Still, more collaborations are needed, especially between different parties concerned with the problem. Tackling gender bias is not optional. In order to use NLP models in real-world applications, they should not exhibit and amplify any detrimental bias and not marginalize any group. The models should be developed more safely and responsibly by removing biases. For future work, the field can benefit from a simple metric that is scalable to languages without any additional overhead. We understand that standardizing an evaluation approach for different tasks is challenging since it depends on the task and its related bias definition. However, we also believe that this is very important step for creating fair applications.

[2] Workshop on Gender Bias in Natural Language Processing
`https://genderbiasnlp.talp.cat/`

[3] AAAI (AI for Social Impact)
`https://aaai.org/Conferences/AAAI-22/aiforsocialimpactcall/`

Bibliography

[Ackerman, 2019] Ackerman, L. M. (2019). Syntactic and cognitive issues in investigating gendered coreference. *Glossa*.

[Alhafni et al., 2020] Alhafni, B., Habash, N., and Bouamor, H. (2020). Gender-aware reinflection using linguistically enhanced neural models. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 139–150.

[Artetxe and Schwenk, 2019] Artetxe, M. and Schwenk, H. (2019). Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

[Bahdanau et al., 2015] Bahdanau, D., Cho, K., and Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. In Bengio, Y. and LeCun, Y., editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

[Bartl et al., 2020] Bartl, M., Nissim, M., and Gatt, A. (2020). Unmasking contextual stereotypes: Measuring and mitigating BERT's gender bias. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 1–16, Barcelona, Spain (Online). Association for Computational Linguistics.

[Basta and Costa-jussà, 2021a] Basta, C. and Costa-jussà, M. R. (2021a). Impact of covid-19 in natural language processing publications: a disaggregated study in gender, contribution and experience. In *Proceedings of the First Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 1–6.

[Basta and Costa-jussà, 2021b] Basta, C. and Costa-jussà, M. R. (2021b). Impact of gender debiased word embeddings in language modeling. *arXiv preprint arXiv:2105.00908*.

[Basta et al., 2019] Basta, C., Costa-jussà, M. R., and Casas, N. (2019). Evaluating the underlying gender bias in contextualized word embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics:GeBNLP 2019*.

[Basta et al., 2021] Basta, C., Costa-jussà, M. R., and Casas, N. (2021). Extensive study on the underlying gender bias in contextualized word embeddings. *Neural Computing and Applications*, 33(8):3371–3384.

[Basta et al., 2020] Basta, C., Costa-Jussà, M. R., and Rodríguez Fonollosa, J. A. (2020). Towards mitigating gender bias in a decoder-based neural machine translation model by adding contextual information. In *Proceedings of the The Fourth Widening Natural Language Processing Workshop*, pages 99–102. Association for Computational Linguistics.

[Basta et al., 2022] Basta, C., Escolano, C., and Costa-jussà, M. R. (2022). To add or relax? examining approaches for mitigating gender bias in machine translation. In *Under Submission*.

[Beltagy et al., 2020] Beltagy, I., Peters, M. E., and Cohan, A. (2020). Longformer: The long-document transformer. *CoRR*, abs/2004.05150.

[Bentivogli et al., 2020] Bentivogli, L., Savoldi, B., Negri, M., Di Gangi, M. A., Cattoni, R., and Turchi, M. (2020). Gender in danger? evaluating speech translation technology on the MuST-SHE corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6923–6933, Online. Association for Computational Linguistics.

[Bérard et al., 2016] Bérard, A., Pietquin, O., Besacier, L., and Servan, C. (2016). Listen and Translate: A Proof of Concept for End-to-End Speech-to-Text Translation. In *NIPS Workshop on end-to-end learning for speech and audio processing*.

[Blodgett et al., 2020] Blodgett, S. L., Barocas, S., Daumé III, H., and Wallach, H. (2020). Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

[Bolukbasi et al., 2016] Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., and Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In Lee, D. D., Sugiyama, M., Luxburg, U. V., Guyon, I., and Garnett, R., editors, *Advances in Neural Information Processing Systems 29*, pages 4349–4357. Curran Associates, Inc.

[Bordia and Bowman, 2019] Bordia, S. and Bowman, S. R. (2019). Identifying and reducing gender bias in word-level language models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics:SWR*.

[Buitinck et al., 2013] Buitinck, L., Louppe, G., Blondel, M., Pedregosa, F., Mueller, A., Grisel, O., and et al., V. N. (2013). API design for machine learning software: Experiences from the scikit-learn project. In *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*.

[Caliskan et al., 2017] Caliskan, A., Bryson, J. J., and Narayanan, A. (2017). Semantics derived automatically from language corpora necessarily contain human biases. *Science*, 356:183–186.

[Cao and Daumé, 2021] Cao, Y. T. and Daumé, Hal, I. (2021). Toward Gender-Inclusive Coreference Resolution: An Analysis of Gender and Bias Throughout the Machine Learning Lifecycle*. *Computational Linguistics*, 47(3):615–661.

[Casas et al., 2019] Casas, N., Fonollosa, J. A., Escolano, C., Basta, C., and Costa-jussà, M. R. (2019). The talp-upc machine translation systems for wmt19 news translation task: Pivoting techniques for low resource mt. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 155–162.

[Cattoni et al., 2021] Cattoni, R., Di Gangi, M. A., Bentivogli, L., Negri, M., and Turchi, M. (2021). Must-c: A multilingual corpus for end-to-end speech translation. *Computer Speech Language*, 66:101155.

[Cho et al., 2014] Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014). Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar. Association for Computational Linguistics.

[Cortes and Vapnik, 1995] Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20(3):273–297.

[Costa-jussà et al., 2022a] Costa-jussà, M. R., Basta, C., Domingo, O., and Niyongabo,

R. A. (2022a). Occgen: Selection of real-world multilingual parallel data balanced in gender within occupations. In *Under Submission*.

[Costa-jussà et al., 2022b] Costa-jussà, M. R., Basta, C., and Gállego, G. I. (2022b). Evaluating gender bias in speech translation. In *LREC*.

[Costa-jussà and de Jorge, 2020] Costa-jussà, M. R. and de Jorge, A. (2020). Fine-tuning neural machine translation on gender-balanced datasets. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 26–34, Barcelona, Spain (Online). Association for Computational Linguistics.

[Costa-jussà et al., 2020] Costa-jussà, M. R., Li Lin, P., and España-Bonet, C. (2020). GeBioToolkit: Automatic extraction of gender-balanced multilingual corpus of Wikipedia biographies. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4081–4088, Marseille, France. European Language Resources Association.

[Costa-jussà, 2019] Costa-jussà, M. R. (2019). An analysis of gender bias studies in natural language processing. *Nature Machine Intelligence*, 1(11):495–496.

[Costa-jussà et al., 2022] Costa-jussà, M. R., Escolano, C., Basta, C., Ferrando, J., Batlle, R., and Kharitonova, K. (2022). Interpreting gender bias in neural machine translation: Multilingual architecture matters. *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 11855–11863.

[Dev et al., 2020] Dev, S., Li, T., Phillips, J. M., and Srikumar, V. (2020). On measuring and mitigating biased inferences of word embeddings. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7659–7666.

[Devlin et al., 2019] Devlin, J., Chang, M., Lee, K., and Toutanova, K. (2019). BERT: pre-training of deep bidirectional transformers for language understanding. In Burstein, J., Doran, C., and Solorio, T., editors, *Proceedings of , NAACL-HLT*, pages 4171–4186. Association for Computational Linguistics.

[Dong et al., 2018] Dong, L., Xu, S., and Xu, B. (2018). Speech-transformer: A no-recurrence sequence-to-sequence model for speech recognition. In *ICASSP*, pages 5884–5888.

[Dwork et al., 2012] Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. S.

(2012). Fairness through awareness. In Goldwasser, S., editor, *Innovations in Theoretical Computer Science 2012, Cambridge, MA, USA, January 8-10, 2012*, pages 214–226. ACM.

[Elazar and Goldberg, 2018] Elazar, Y. and Goldberg, Y. (2018). Adversarial removal of demographic attributes from text data. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 11–21, Brussels, Belgium. Association for Computational Linguistics.

[Elman, 1990] Elman, J. L. (1990). Finding structure in time. *Cognitive science*, 14(2):179–211.

[Escolano, 2022] Escolano, C. (2022). *Learning Multilingual and Multimodal Representations with Language-Specific Encoders and Decoders for Machine Translation*. PhD thesis, Universitat Politècnica de Catalunya.

[Escolano et al., 2021a] Escolano, C., Costa-jussà, M. R., Fonollosa, J. A. R., and Artetxe, M. (2021a). Multilingual machine translation: Closing the gap between shared and language-specific encoder-decoders. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 944–948, Online. Association for Computational Linguistics.

[Escolano et al., 2021b] Escolano, C., Ojeda, G., Basta, C., and Costa-jussà, M. R. (2021b). Multi-task learning for improving gender accuracy in neural machine translation. In *ICON 2021: 18th International Conference on Natural Language Processing*.

[Escolano et al., 2021c] Escolano, C., Tsiamas, I., Basta, C., Ferrando, J., Costa-jussà, M. R., and Fonollosa, J. A. (2021c). The talp-upc participation in wmt21 news translation task: an mbart-based nmt approach. In *Proceedings of the Sixth Conference on Machine Translation*, pages 117–122.

[Ethayarajh et al., 2019] Ethayarajh, K., Duvenaud, D., and Hirst, G. (2019). Understanding undesirable word embedding associations. In Korhonen, A., Traum, D. R., and Màrquez, L., editors, *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 1696–1705. Association for Computational Linguistics.

[Fan et al., 2021] Fan, A., Bhosale, S., Schwenk, H., Ma, Z., El-Kishky, A., Goyal, S., Baines, M., Celebi, O., Wenzek, G., Chaudhary, V., Goyal, N., Birch, T., Liptchinsky, V., Edunov, S., Auli, M., and Joulin, A. (2021). Beyond english-centric multilingual machine translation. *J. Mach. Learn. Res.*, 22:107:1–107:48.

[Ferrando and Costa-jussà, 2021] Ferrando, J. and Costa-jussà, M. R. (2021). Attention weights in transformer NMT fail aligning words between sequences but largely explain model predictions. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 434–443, Punta Cana, Dominican Republic. Association for Computational Linguistics.

[Firat et al., 2017] Firat, O., Cho, K., Sankaran, B., Vural, F. T. Y., and Bengio, Y. (2017). Multi-Way, Multilingual Neural Machine Translation. *Computer Speech and Language, Special Issue in Deep learning for Machine Translation*, 45:236–252.

[Fonollosa et al., 2019] Fonollosa, J. A. R., Casas, N., and Costa-jussà, M. R. (2019). Joint source-target self attention with locality constraints. *CoRR*, abs/1905.06596.

[Font and Costa-jussà, 2019] Font, J. E. and Costa-jussà, M. R. (2019). Equalizing gender bias in neural machine translation with word embeddings techniques. In *Proceedings of the First ACL Workshop on Gender Bias in Natural Language Processing*, pages 147–154, Florence, Italy.

[Frank et al., 2004] Frank, A., Hoffmann, C., Strobel, M., et al. (2004). Gender issues in machine translation. *Univ. Bremen.*

[Gangi et al., 2019] Gangi, M. A. D., Negri, M., and Turchi, M. (2019). Adapting Transformer to End-to-End Spoken Language Translation. In *Interspeech*, pages 1133–1137.

[Goldberg, 2017] Goldberg, Y. (2017). Neural network methods for natural language processing. *Synthesis lectures on human language technologies*, 10(1):1–309.

[Gonen and Goldberg, 2019] Gonen, H. and Goldberg, Y. (2019). Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 609–614, Minneapolis, Minnesota. Association for Computational Linguistics.

[Gonen and Webster, 2020] Gonen, H. and Webster, K. (2020). Automatically identifying gender issues in machine translation using perturbations. In *Findings of the Association* *for Computational Linguistics: EMNLP 2020*, pages 1991–1995, Online. Association for Computational Linguistics.

[González et al., 2020] González, A. V., Barrett, M., Hvingelby, R., Webster, K., and Søgaard, A. (2020). Type b reflexivization as an unambiguous testbed for multilingual multi-task gender bias. In *In Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP).*

[Graves, 2012] Graves, A. (2012). Sequence transduction with recurrent neural networks. *CoRR*, abs/1211.3711.

[Guo and Caliskan, 2021] Guo, W. and Caliskan, A. (2021). Detecting emergent intersectional biases: Contextualized word embeddings contain a distribution of human-like biases. In *AAAI/ACM Conference on AI, Ethics, and Society, Virtual Event (AIES)*, pages 122–133. ACM.

[Gygax et al., 2019] Gygax, P. M., Elmiger, D., Zufferey, S., Garnham, A., Sczesny, S., von Stockhausen, L., Braun, F., and Oakhill, J. (2019). A language index of grammatical gender dimensions to study the impact of grammatical gender on the way we perceive women and men. *Frontiers in Psychology*, 10:1604.

[Habash et al., 2019] Habash, N., Bouamor, H., and Chung, C. (2019). Automatic gender identification and reinflection in arabic. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 155–165.

[Hardmeier et al., 2022] Hardmeier, C., Basta, C., Costa-jussà, M. R., Stanovsky, G., and Gonen, H., editors (2022). *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, Seattle, Washington. Association for Computational Linguistics.

[Hochreiter and Schmidhuber, 1997] Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural Comput.*, 9(8):1735–1780.

[Hovy, 2015] Hovy, D. (2015). Demographic factors improve classification performance. In *Proceedings of the 53rd annual meeting of the Association for Computational Linguistics and the 7th international joint conference on natural language processing (volume 1: Long papers)*, pages 752–762.

[Hovy et al., 2020] Hovy, D., Bianchi, F., and Fornaciari, T. (2020). “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*, pages 1686–1690, Online. Association for Computational Linguistics.

[Hovy and Prabhumoye, 2021] Hovy, D. and Prabhumoye, S. (2021). Five sources of bias in natural language processing. *Language and Linguistics Compass*, 15(8):e12432.

[Howard and Ruder, 2018] Howard, J. and Ruder, S. (2018). Universal language model fine-tuning for text classification. In *Proceedings of the 576h Conference of the Association for Computational Linguistics, ACL*, pages 328–339, Melbourne, Australia.

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

[Junczys-Dowmunt, 2019] Junczys-Dowmunt, M. (2019). Microsoft translator at WMT 2019: Towards large-scale document-level neural machine translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 225–233, Florence, Italy. Association for Computational Linguistics.

[Kaneko and Bollegala, 2019] Kaneko, M. and Bollegala, D. (2019). Gender-preserving debiasing for pre-trained word embeddings. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL*, pages 1641–1650, Florence, Italy.

[Kaneko and Bollegala, 2021a] Kaneko, M. and Bollegala, D. (2021a). Debiasing pre-trained contextualised embeddings. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1256–1266, Online. Association for Computational Linguistics.

[Kaneko and Bollegala, 2021b] Kaneko, M. and Bollegala, D. (2021b). Dictionary-based debiasing of pre-trained word embeddings. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 212–223, Online. Association for Computational Linguistics.

[Kharitonova et al., 2021] Kharitonova, K., Escolano, C., Costa-jussà, M. R., Basta, C., and Armengol-Estapé, J. (2021). Neutralizing gender bias in neural machine translation by introducing linguistic knowledge. In *EMNLP 2021, WiNLP*.

[Kingma and Ba, 2015] Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization. In *Proceedings of ICLR.*

[Kiritchenko et al., 2021] Kiritchenko, S., Nejadgholi, I., and Fraser, K. C. (2021). Confronting abusive language online: A survey from the ethical and human rights perspective. 71:431–478.

[Kirkpatrick et al., 2017] Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., et al. (2017). Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526.

[Koehn, 2005] Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86. Citeseer.

[Koehn et al., 2007] Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., et al. (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the ACL: Demo Papers*, pages 177–180.

[Kudo and Richardson, 2018] Kudo, T. and Richardson, J. (2018). SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

[Kurita et al., 2019] Kurita, K., Vyas, N., Pareek, A., Black, A. W., and Tsvetkov, Y. (2019). Measuring bias in contextualized word representations. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 166–172.

[Läubli et al., 2018] Läubli, S., Sennrich, R., and Volk, M. (2018). Has machine translation achieved human parity? a case for document-level evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4791–4796, Brussels, Belgium. Association for Computational Linguistics.

[Lauscher et al., 2021] Lauscher, A., Lueken, T., and Glavaš, G. (2021). Sustainable modular debiasing of language models. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4782–4797, Punta Cana, Dominican Republic. Association for Computational Linguistics.

[Levy et al., 2021] Levy, S., Lazar, K., et al. (2021). Collecting a large-scale gender bias dataset for coreference resolution and machine translation. In *EMNLP findings, 2021.*

[Lewis and Lupyan, 2020] Lewis, M. and Lupyan, G. (2020). Gender stereotypes are reflected in the distributional structure of 25 languages. *Nat Hum Behav 4, 1021–1028.*

[Liang et al., 2020] Liang, S., Dufter, P., and Schütze, H. (2020). Monolingual and multilingual reduction of gender bias in contextualized representations. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5082–5093, Barcelona, Spain (Online). International Committee on Computational Linguistics.

[Lison and Tiedemann, 2016] Lison, P. and Tiedemann, J. (2016). OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 923–929, Portorož, Slovenia. European Language Resources Association (ELRA).

[Liu et al., 2021a] Liu, D., Niehues, J., Cross, J., Guzmán, F., and Li, X. (2021a). Improving zero-shot translation by disentangling positional information. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1259–1273, Online. Association for Computational Linguistics.

[Liu et al., 2019a] Liu, N. F., Gardner, M., Belinkov, Y., Peters, M. E., and Smith, N. A. (2019a). Linguistic knowledge and transferability of contextual representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1073–1094, Minneapolis, Minnesota. Association for Computational Linguistics.

[Liu et al., 2020] Liu, Q., Kusner, M. J., and Blunsom, P. (2020). A survey on contextual embeddings. *arXiv preprint arXiv:2003.07278.*

[Liu et al., 2021b] Liu, R., Jia, C., Wei, J., Xu, G., Wang, L., and Vosoughi, S. (2021b). Mitigating political bias in language models through reinforced calibration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14857–14866.

[Liu et al., 2019b] Liu, Y., Xiong, H., Zhang, J., and et al. (2019b). End-to-End Speech Translation with Knowledge Distillation. In *Interspeech*, pages 1128–1132.

[Lu et al., 2020] Lu, K., Mardziel, P., Wu, F., Amancharla, P., and Datta, A. (2020).

Gender bias in neural natural language processing. In *Logic, Language, and Security*, pages 189–202. Springer.

[Lu et al., 2018] Lu, Y., Keung, P., Ladhak, F., Bhardwaj, V., Zhang, S., and Sun, J. (2018). A neural interlingua for multilingual machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 84–92, Brussels, Belgium. Association for Computational Linguistics.

[Macé and Servan, 2019] Macé, V. and Servan, C. (2019). Using whole document context in neural machine translation. In *Proceedings of the 16th International Conference on Spoken Language Translation*, Hong Kong. Association for Computational Linguistics.

[Madaan et al., 2018] Madaan, N., Singh, G., Mehta, S., Chetan, A., and Joshi, B. (2018). Generating clues for gender based occupation de-biasing in text. *ArXiv*, abs/1804.03839.

[May et al., 2019] May, C., Wang, A., Bordia, S., Bowman, S. R., and Rudinger, R. (2019). On measuring social biases in sentence encoders. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 622–628, Minneapolis, Minnesota. Association for Computational Linguistics.

[McGuire et al., 2020] McGuire, L., Mulvey, K. L., Goff, E., Irvin, M. J., Winterbottom, M., Fields, G. E., Hartstone-Rose, A., and Rutland, A. (2020). Stem gender stereotypes from early childhood through adolescence at informal science centers. *Journal of applied developmental psychology*, 67:101109.

[Mehrabi et al., 2021] Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., and Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6):1–35.

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

[Nadeem et al., 2020] Nadeem, A., Abedin, B., and Marjanovic, O. (2020). Gender bias in ai: A review of contributing factors and mitigating strategies. *ACIS 2020 proceedings.*

[Nadeem et al., 2021] Nadeem, M., Bethke, A., and Reddy, S. (2021). StereoSet: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5356–5371, Online. Association for Computational Linguistics.

[Nangia et al., 2020] Nangia, N., Vania, C., Bhalerao, R., and Bowman, S. R. (2020). CrowS-pairs: A challenge dataset for measuring social biases in masked language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1953–1967, Online. Association for Computational Linguistics.

[Neubig et al., 2018] Neubig, G., Sperber, M., Wang, X., Felix, M., Matthews, A., Padmanabhan, S., Qi, Y., Sachan, D., Arthur, P., Godard, P., Hewitt, J., Riad, R., and Wang, L. (2018). XNMT: The eXtensible neural machine translation toolkit. In *AMTA*, pages 185–192.

[Papineni et al., 2002] Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

[Peters et al., 2018] Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep contextualized word representations. In *Proc. of the ACL(Long Papers)*, pages 2227–2237, New Orleans, Louisiana.

[Peters et al., 2019] Peters, M. E., Ruder, S., and Smith, N. A. (2019). To tune or not to tune? adapting pretrained representations to diverse tasks. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, pages 7–14, Florence, Italy. Association for Computational Linguistics.

[Prates et al., 2020] Prates, M. O. R., Avelar, P. H. C., and Lamb, L. (2020). Assessing gender bias in machine translation – a case study with google translate. *Neural Computing and Applications, 32*, page 6363–6381.

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

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

[Rahman and Ng, 2012] Rahman, A. and Ng, V. (2012). Resolving complex cases of definite pronouns: the winograd schema challenge. In *EMNLP*, pages 777–789. Association for Computational Linguistics.

[Ravfogel et al., 2020] Ravfogel, S., Elazar, Y., Gonen, H., Twiton, M., and Goldberg, Y. (2020). Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7237–7256, Online. Association for Computational Linguistics.

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

[Renduchintala and Williams, 2021] Renduchintala, A. and Williams, A. (2021). Investigating failures of automatic translation in the case of unambiguous gender. *arXiv preprint arXiv:2104.07838*.

[Rudinger et al., 2018] Rudinger, R., Naradowsky, J., Leonard, B., and Van Durme, B. (2018). 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)*, pages 8–14, New Orleans, Louisiana. Association for Computational Linguistics.

[Saunders and Byrne, 2020] Saunders, D. and Byrne, B. (2020). Reducing gender bias in neural machine translation as a domain adaptation problem. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7724–7736. Association for Computational Linguistics.

[Saunders et al., 2020] Saunders, D., Sallis, R., and Byrne, B. (2020). Neural machine translation doesn't translate gender coreference right unless you make it. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 35–43, Barcelona, Spain (Online). Association for Computational Linguistics.

[Savoldi et al., 2021] Savoldi, B., Gaido, M., Bentivogli, L., Negri, M., and Turchi, M. (2021). Gender bias in machine translation. *Transactions of the Association for Computational Linguistics*, 9:845–874.

[Schwenk, 2018] Schwenk, H. (2018). Filtering and mining parallel data in a joint multilingual space. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 228–234, Melbourne, Australia. Association for Computational Linguistics.

[Schwenk et al., 2021] Schwenk, H., Chaudhary, V., Sun, S., Gong, H., and Guzmán, F. (2021). WikiMatrix: Mining 135M parallel sentences in 1620 language pairs from Wikipedia. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1351–1361, Online. Association for Computational Linguistics.

[Schwenk and Douze, 2017] Schwenk, H. and Douze, M. (2017). Learning joint multilingual sentence representations with neural machine translation. In *Proceedings of the 2nd Workshop on Representation Learning for NLP (Rep4NLP)*, pages 157–167. Association for Computational Linguistics.

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

[Shah et al., 2020] Shah, D. S., Schwartz, H. A., and Hovy, D. (2020). Predictive biases in natural language processing models: A conceptual framework and overview. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5248–5264, Online. Association for Computational Linguistics.

[Smith, 2020] Smith, N. A. (2020). Contextual word representations: putting words into computers. *Communications of the ACM*, 63(6):66–74.

[Sperber et al., 2018] Sperber, M., Niehues, J., Neubig, G., Stüker, S., and Waibel, A. (2018). Self-Attentional Acoustic Models. In *InterSpeech*.

[Stafanovičs et al., 2020] Stafanovičs, A., Pinnis, M., and Bergmanis, T. (2020). Mitigating gender bias in machine translation with target gender annotations. In *Proceedings of the Fifth Conference on Machine Translation*, pages 629–638, Online. Association for Computational Linguistics.

[Stanczak and Augenstein, 2021] Stanczak, K. and Augenstein, I. (2021). A survey on gender bias in natural language processing. *arXiv preprint arXiv:2112.14168*.

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

[Stella, 2021] Stella, R. (2021). A Dataset for Studying Gender Bias in Translation.

[Sun et al., 2019] Sun, T., Gaut, A., Tang, S., Huang, Y., ElSherief, M., Zhao, J., Mirza, D., Belding, E., Chang, K.-W., and Wang, W. Y. (2019). Mitigating gender bias in natural language processing: Literature review. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1630–1640, Florence, Italy. Association for Computational Linguistics.

[Suresh and Guttag, 2021] Suresh, H. and Guttag, J. V. (2021). A framework for understanding sources of harm throughout the machine learning life cycle. *Equity and Access in Algorithms, Mechanisms, and Optimization.*

[Sutskever et al., 2014] Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

[Tan et al., 2020] Tan, S., Joty, S., Kan, M.-Y., and Socher, R. (2020). It's morphin' time! Combating linguistic discrimination with inflectional perturbations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2920–2935, Online. Association for Computational Linguistics.

[Tan and Celis, 2019] Tan, Y. C. and Celis, L. E. (2019). Assessing social and intersectional biases in contextualized word representations. In *Advances in Neural Information Processing Systems*, pages 13209–13220.

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

[Tatman, 2017] Tatman, R. (2017). Gender and dialect bias in YouTube's automatic captions. In *Proc. 1st ACL Workshop on Ethics in NLP*, pages 53–59.

[Tiedemann and Thottingal, 2020] Tiedemann, J. and Thottingal, S. (2020). OPUS-MT — Building open translation services for the World. In *Proceedings of the 22nd An-*

nual Conferenec of the European Association for Machine Translation (EAMT), Lisbon, Portugal.

[Tomalin et al., 2021] Tomalin, M., Byrne, B., Concannon, S., Saunders, D., and Ullmann, S. (2021). The practical ethics of bias reduction in machine translation: why domain adaptation is better than data debiasing. *Ethics and Information Technology*, pages 1–15.

[Vanmassenhove et al., 2018] Vanmassenhove, E., Hardmeier, C., and Way, A. (2018). Getting gender right in neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.

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

[Vaswani et al., 2017] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

[Vila et al., 2018] Vila, L., Escolano, C., Fonollosa, J. A. R., and Costa-jussà, M. R. (2018). End-to-End Speech Translation with the Transformer. In *IberSPEECH*, pages 60–63.

[Webster et al., 2018] Webster, K., Recasens, M., Axelrod, V., and Baldridge, J. (2018). Mind the gap: A balanced corpus of gendered ambiguous pronouns. *Transactions of the Association for Computational Linguistics*, 6:605–617.

[Webster et al., 2020] Webster, K., Wang, X., Tenney, I., Beutel, A., Pitler, E., Pavlick, E., Chen, J., and Petrov, S. (2020). Measuring and reducing gendered correlations in pre-trained models. *ArXiv*, abs/2010.06032.

[Weiss et al., 2017] Weiss, R. J., Chorowski, J., Jaitly, N., Wu, Y., and Chen, Z. (2017). Sequence-to-sequence models can directly translate foreign speech. In *Interspeech*, pages 2625–2629.

[Zhang et al., 2018] Zhang, B. H., Lemoine, B., and Mitchell, M. (2018). Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 335–340. ACM.

[Zhao et al., 2019] Zhao, J., Wang, T., Yatskar, M., Cotterell, R., Ordonez, V., and Chang, K.-W. (2019). Gender bias in contextualized word embeddings. In *Proc. of the Conference of the NAACL.*

[Zhao et al., 2018a] Zhao, J., Wang, T., Yatskar, M., Ordonez, V., and Chang, K.-W. (2018a). 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)*, pages 15–20, New Orleans, Louisiana. Association for Computational Linguistics.

[Zhao et al., 2018b] Zhao, J., Zhou, Y., Li, Z., Wang, W., and Chang, K.-W. (2018b). Learning gender-neutral word embeddings. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4847–4853, Brussels, Belgium. Association for Computational Linguistics.

[Zhou et al., 2019] Zhou, P., Shi, W., Zhao, J., Huang, K.-H., Chen, M., Cotterell, R., and Chang, K.-W. (2019). Examining gender bias in languages with grammatical gender. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5276–5284, Hong Kong, China. Association for Computational Linguistics.

[Zmigrod et al., 2019] Zmigrod, R., Mielke, S. J., Wallach, H., and Cotterell, R. (2019). Counterfactual data augmentation for mitigating gender stereotypes in languages with rich morphology. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1651–1661, Florence, Italy. Association for Computational Linguistics.