MASTER'S DEGREE THESIS

Master of Science in Advanced Mathematics and Mathematical Engineering

Title: Determining Bias in Machine Translation with Deep Learning Techniques

Author: Joel Escudé Font

Advisor: Marta Ruiz Costa-jussà

Department: Teoria del Senyal i Comunicacions

Academic year: 2018-2019

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Facultat de Matemàtiques i Estadística

Universitat Politècnica de Catalunya
Facultat de Matemàtiques i Estadística

Master in Advanced Mathematics and Mathematical Engineering

Master's Degree Thesis

Determining Bias in Machine Translation with Deep Learning Techniques

Joel Escudé Font

Advisor: Marta R. Costa-jussà

Departament de Teoria del Senyal i Comunicacions

Acknowledgements

I would like to express my gratitude to Marta R. Costa-jussà for advising me during the elaboration of this thesis and for inspiring me throughout the whole process. I would also like to thank the speech processing research group at the TSC department for their insightful comments.

Abstract

Neural machine translation has improved significantly over the last few years for translation tasks between natural languages using deep learning. These deep learning models are trained on available large text corpora which contains biases and stereotypes. As a consequence, models inherit these social biases. An example of this is the fact that "friend" in the English sentence "She works in a hospital, my friend is a nurse" would be correctly translated to "amiga" (feminine of friend) in Spanish, while "She works in a hospital, my friend is a doctor" would be incorrectly translated to "amigo" (masculine of friend) in Spanish. We consider that this translation contains gender bias since it ignores the fact that, for both cases, "friend" is a female and translates by focusing on the occupational stereotypes, i.e. translating doctor as male and nurse as female. Recent state-of-the-art methods have shown results in reducing gender bias in other natural language processing applications such as word embeddings, which is the task of representing words as vectors to extract similarities from them. We take advantage of the fact that word embeddings are used by default in standard neural machine translation systems to propose the first debiased neural machine translation system. Specifically, we propose, experiment and analyze the use of a pair of debiased pre-trained word embeddings in the Transformer translation architecture. We evaluate our proposed system on a generic English-Spanish task, showing that translation performance gains one BLEU point. As for specific evaluation about the bias presence, we generate a new test set of occupations (for the same language pair) and, over this test set, we show that our proposed system learns to neutralize previously existing biases from the baseline system.

Contents

Chapter 1

Introduction

1.1. Motivation and objectives

Language is one of the most interesting and complex skills used in our daily life, and may even be taken for granted on our ability to communicate. However, the understanding of meanings between lines in natural languages is not straightforward for the logic rules of programming languages at first. Natural language processing (NLP) is a sub-field of artificial intelligence (AI) that focuses on making natural languages understandable to computers. Similarly, the translation between different natural languages is a task for machine translation (MT).

Machine learning (ML) has a set of tools for data analysis and building models by learning patterns from data. More specifically, deep learning (DL) uses neural networks to improve the performance of learning tasks over statistical-based models in sequence-to-sequence problems for translation tasks [Sutskever et al., 2014]. Neural machine translation (NMT) is a recent approach in MT which learns patterns between source and target language corpora to produce text translations using deep neural networks.

One downside of models trained with human generated corpora is that social biases present in the data are learned. This is shown when training word embeddings, a vector representation of words, in news sets with crowd-sourcing evaluation to quantify the presence of biases, such as gender bias, in those representation [Bolukbasi et al., 2016]. This can affect downstream applications [Zhao et al., 2018a] and are at risk of being amplified [Zhao et al., 2017].

The objective of this work is to study the presence of gender bias in machine translation and give insight on the impact of debiasing in such systems. While there are some precedent studies in detecting the presence of gender bias in MT [Prates et al., 2018], this is one of the first studies in proposing debiasing techniques for this application. Therefore, we have to define an appropriate framework to evaluate the debiasing techniques.

1.2. Contributions

The main contribution of this thesis is progress on a recent detected problem which is gender bias in machine learning and in the particular application of machine translation. The progress towards reducing gender bias in MT is made in two directions: first, we define a framework to experiment, detect and evaluate gender bias in MT for a particular task; second, we propose to use debiased word embeddings techniques in the MT system to reduce the detected bias.

The work in this thesis has resulted in a paper[1] for an international conference. The arXiv identifier is 1901.03116. The files used in the experiments will be available on a GitHub repository[2] under the same

[1] https://arxiv.org/abs/1901.03116

[2] https://github.com/joelescudefont/genbiasmt

Creative Commons license of this thesis.

1.3. Outline

The rest the thesis is organized as follows. Chapter 2 reports material relevant to the background of the study. Chapter 3 presents previous work on the bias problem. Chapter 4 reports the methodology used for the experiments and chapter 5 details the experimental framework. The results and discussion are included in chapter 6 and chapter 7 presents the main conclusions and ideas for future work.

Chapter 2

Background

In this chapter we introduce the models used in our study. We report a state-of-the-art MT model that is used nowadays by the community which is the Transformer. We also describe word embeddings and two approaches that approach reducing biases present in these word representations. One approach involves removing biases from these embeddings and another approach to learn a gender neutral version of these models.

2.1. Transformer

The Transformer [Vaswani et al., 2017] is a neural networks architecture purely based on a self-attention mechanism that show an improvement in performance on machine translation tasks over previous models. It is more efficient in using computation resources and it trains faster. See Figure 2.1 for a simplification of its architecture.

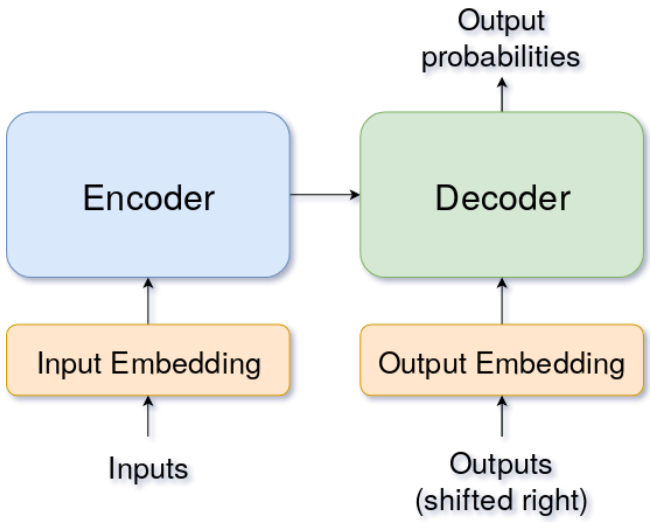

Figure 2.1: Simplification of the Transformer architecture.

Neural networks start by representing individual words as vectors, word embeddings (more on this later), to process language as a vector space representation which can have a fixed or variable length. Words surrounding another particular word determine its meaning and how it is represented in this space, thus context influences in deciding the appropriate meaning for a given task using this representation.

Processing of sequential data is performed with recurrent neural netoworks (RNN) which perform several steps before providing a decision based on distant words. The number of steps needed by a neural network is shown to influence negatively on the decision making process as the steps increase [Hochreiter et al., 2001].

The Transformer computes a reduced constant number of these steps using a self-attention mechanism on each one. This mechanism models the relations between words independently of their position, thus improving the number of steps needed to determine a target word. An attention score is computed for all words in a sentence when comparing the contribution of each word to the next representation. An encoder reads an input sentence to generate a representation which is later used by a decoder to produce a sentence output word by word. New representations are generated at each steps in parallel for all words. The decoder uses self-attention in the generated words and also uses the representations from the last words in the encoder to produce a single word each time.

2.2. Word embeddings

Word embeddings are a low-dimensional representation of words in a vector space. These models are based on the hypothesis that words appearing in the same context share semantic meaning, thus semantically similar words are encoded with real values in a continuous vector space representation. They are used in many NLP applications for being a less sparse and more expressive representation than discrete atomic symbols and one-hot vectors,

Arithmetic operations can be computed with these word vectors, and analogies between pairs of words are found with the pattern "A is to B what C is to D" [Mikolov et al., 2013]. As an instance, analogies such as countries and their respective capitals, verb tenses and royal gendered pairs. See Figure 2.2.

While there are many techniques for extracting word embeddings, in this work we are using Global Vectors, or GloVe [Pennington et al., 2014]. GloVe is an unsupervised method for learning word embeddings. This count-based method, uses statistical information of word occurrences from a given corpus to train a vector space for which each vector is related to a word and their values describes their semantic relations.

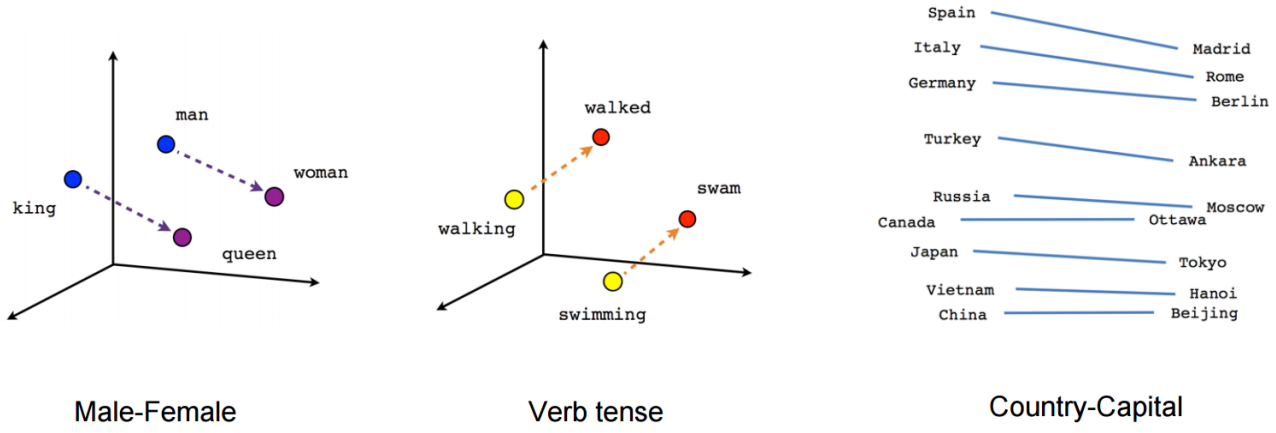

Figure 2.2: Examples of analogies in word embeddings. Source: `https://www.tensorflow.org/images/linear-relationships.png`.

2.3. Debiasing embeddings

The presence of biases in word embeddings has aroused as a topic of discussion about fairness in ML. More specifically, gender stereotypes are learned from human generated corpora as shown by [Bolukbasi et al., 2016] which led to a debiasing method for word embeddings, we will refer to this approach as Debiaswe. Another method is to train gender neutral word embeddings, this is the approach of GN-GloVe [Zhao et al., 2018b]. The main ideas behind these algorithms are described next.

2.3.1 Debiaswe

Debiaswe [Bolukbasi et al., 2016] is a post-process method for debiasing word embeddings. It consists of two main parts: First, identifying the direction in the embeddings for a type of bias. Second, gender neutral words in this direction are neutralized to zero and also equalized by making the neutral word equidistant to the remaining ones in the set. The disadvantage of the first part of the process is that it removes valuable semantic information. For example, words with several meanings not related to the particular bias being treated.

2.3.2 GN-GloVe

GN-GloVe [Zhao et al., 2018b] is an algorithm for learning gender netural word embeddings. It is based on GloVe [Pennington et al., 2014] and it is modified to have protected attributes in the embeddings so to capture particular information in a specific dimension. For a protected attribute like gender, the minimization objective captures word proximity like GloVe and restricts gender information in a specific dimension so that the remaining are neutral. A set of seed male and female words are used to define metrics for computing the optimization and a set of gender neutral words is used for restricting neutral words in a gender direction.

Chapter 3

Related work

Currently, there is few work on biases in the specific task of machine translation. However, there are studies on the presence of biases in many other topics such as in word embedding models, abusive language detection and other related applications. There are also papers that propose methods for reducing the presence of bias. This chapter reports some of these works.

Word embeddings can learn biases from human generated corpora. [Bolukbasi et al., 2016] showed that stereotypical analogies are present in word embeddings both for gender and race and proposed a method for reducing such biases in these representations. [Caliskan et al., 2017] found also a strong gender and racial bias presence is pre-trained embeddings and proposed a method for measuring bias in word embeddigns.

[Zhao et al., 2018a] showed that sexism is present in coreference resolution systems due to the word embeddings components. Applications that use these embeddings, such as curriculum filtering, may discriminate candidates because of their gender. The amplification of biases in downstream applications is a concerning problem also that can enlarge the gap between genders, for example, in search engines, for professions where the name of the candidates may be discriminated by the algorithm because of their bias towards a specific gender. Thus, broadening even further gender inequality for a given field. [Zhao et al., 2017] shows that gender bias is learned and amplified in models trained from data sets containing web images used in language modelling tasks. As an example, the word "cooking" is more probable to be related to females than males and it can also be further amplified.

There are debiasing methods for reducing the presence of biases in word embeddings. [Bolukbasi et al., 2016] first quantified the presence of bias in this representation and proposed an algorithm for removing gender stereotypes from these embeddings. On the other hand, [Zhao et al., 2018b] proposed GN-GloVe, an algorithm to generate gender neutral word embeddings. The approach is to restrict gender information attributes in certain dimensions to keep the remaining free of this attributes.

Gender biases are also found in detection algorithms. [Park et al., 2018] studies the reduction of such biases in abusive language detection. These models have a strong bias towards words that identify gender because of the data sets in which they are trained. Sentences that do not necessarily show sexism are detected as false positives compromising the robustness of the models. Debiased word embeddings combined with augmenting and swapping gender data was found the most effective method for reducing gender bias in this task.

[Prates et al., 2018] performs a case study on gender bias in machine translation. They built a test set consisting of a list of jobs and gender-specific sentences. Using English as a target language and a variety of gender neutral languages as a source, i.e. languages that do not explicitly give gender information about the subject, they test these sentences on the translating service Google Translate. They find that occupations related to science, engineering and mathematics present a strong stereotype toward male subjects.

Chapter 4

Methods

In this chapter, we describe the methodology used for the study of the presence of bias in a translating system. We make use of the modularity of the Transformer to train models with gender neutral and debiased pre-trained embeddings to study their impact on the translations compared to a baseline model.

First, we learn word embeddings with GloVe [Pennington et al., 2014] and gender neutral embeddings with GN-GloVe [Zhao et al., 2018b] using the same corpus for the training of the models. We also debias the first GloVe embeddings with the Debiaswe algorithm from [Bolukbasi et al., 2016]. Second, the models are trained using the Transformer. The Transformer learns word embeddings in its first layers. For this module where the embeddings are learned, we use pre-trained embeddings for the training of all models with the exception of a baseline model. In this baseline the embeddings are learned within the system. Another baseline model is trained with pre-trained embeddings which are not debiased. Also, the gender neutral and debiased embeddings are used for training the other models.

Moreover, the pre-trained embeddings can be used in the encoder, the decoder or both sides. For the gender neutral and debiased pre-trained embeddings we train two models, one with pre-trained embeddings in the encoder and decoder sides, the other with pre-trained embeddings only in the encoder side. See Figure 4.1 and Figure 4.2, respectively.

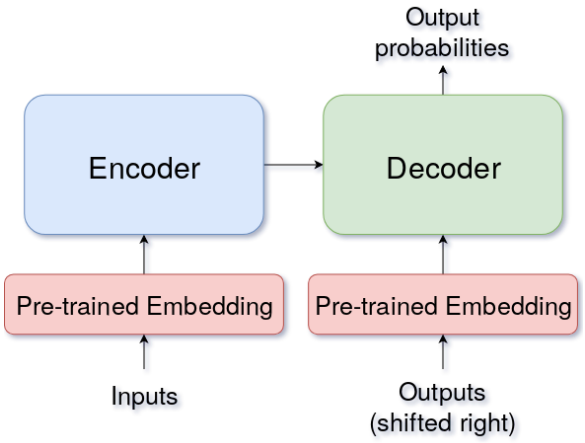

Figure 4.1: Pre-trained word embeddings in both the encoder and decoder sides.

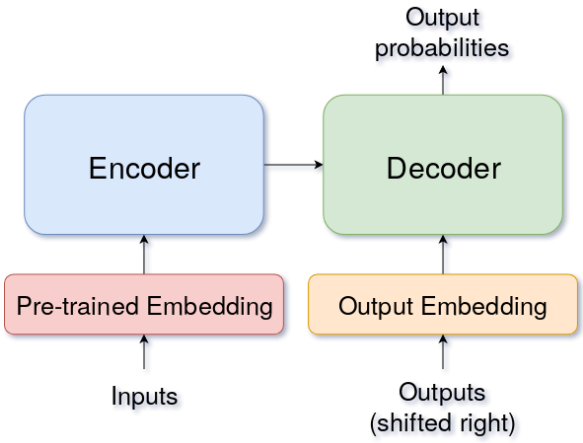

Figure 4.2: Pre-trained word embeddings only in the encoder side.

A summary of the models trained for our study is given below:

- A baseline model where the embeddings are learned within the system, i.e. no pre-trained embeddings used. Another baseline model for which we use pre-trained embeddings learned with GloVe (no debiasing) using the same corpus for training the model.

- The other models are trained with pre-trained embeddings for the encoder and decoder sides, also for only the encoder side. One with the gender neutral embeddings learned with GN-GloVe, also using the same corpus for training the model. Another with debiased GloVe embeddings using Debiaswe.

The implementation of the Transformer used in the experiments is OpenNMT [Klein et al., 2017], an open-source toolkit for NMT which is actively maintained. To evaluate the performance of the models we use the BLEU metric [Papineni et al., 2002]. This metric gives a score for a predicted translation set compared to its expected outputs, a reference set.

Chapter 5

Experimental framework

In this chapter, we present the experimental framework. We describe the train, validation and test data sets. We also introduce a custom test set used for studying debiasing in the models. We give details on the word embedding models and the parameters values used for their training. We also give details on the parameters for training the models and the implementation code we used for this task. Finally, we briefly comment on the usage of computational resources.

5.1. Corpora

The language pair used for the experiments is English-Spanish. The training set consists of sentences from proceedings of the European Parliament [Koehn, 2005]. The validation and test sets are the *newstest2012* and *newstest2013* data sets, respectively. They are provided by the Workshop on Machine Translation 2013[1] (WMT). See Table A.1 for sample sentences from these data sets. For details on their sizes, see Table 5.1.

Language	Data set	Num. of sentences	Num. of words	Vocab. size
English	Train	16.6M	427.6M	1.32M
	Dev	3k	73k	10k
	Test	3k	65k	9k
	Occupations test	1k	17k	0.8k
Spanish	Train	16.6M	477.3M	1.37M
	Dev	3k	79k	12k
	Test	3k	71k	11k
	Occupations test	1k	17k	0.8k

Table 5.1: English-Spanish data sets sizes.

To study gender bias, we have developed an additional test set with custom sentences to evaluate the quality of the translation in the models. We built this test set using a sentence pattern "*I've known* {*her, him,* <*proper noun*>} *for a long time, my friend works as* {*a, an*} <*occupation*>" for a list of occupations from different professional areas. We refer to this test as *Occupations test*. Sample sentences from this set are in Table 5.2. Sizes related to this test are also listed in Table 5.1. We use Spanish proper names to reduce ambiguity in this particular test. These sentences are used in their tokenized version.

With these test sentences we see how "friend" is translated into its Spanish equivalent "amiga" or "amigo" which has a gender relation for each word, female and male, respectively. Note that we are formulating

[1] http://www.statmt.org/wmt13/

sentences with an ambiguous word "friend" that can be translated into any of the two words and we are adding context in the same sentence so that the system has enough information to translate them correctly. The list of occupations is from the U.S. Bureau of Labor Statistics[2], which also includes statistical data for gender and race for most professions. We use a pre-processed version of this list from [Prates et al., 2018].

(En) *I've known her for a long time, my friend works as an accounting clerk.* (Es) *La conozco desde hace mucho tiempo, mi amiga trabaja como contable.*
(En) *I've known him for a long time, my friend works as an accounting clerk.* (Es) *Lo conozco desde hace mucho tiempo, mi amigo trabaja como contable.*
(En) *I've known María for a long time, my friend works as an accounting clerk.* (Es) *Conozco a María desde hace mucho tiempo, mi amiga trabaja como contable.*
(En) I've known Juan for a long time, my friend works as an accounting clerk. (Es) *Conozco a Juan desde hace mucho tiempo, mi amigo trabaja como contable.*

Table 5.2: Sample sentences from the *Occupations test* set. English (En) and Spanish (Es).

5.2. Models

5.2.1 Word embeddings models

The word embeddings are trained from the same corpus, using GloVe [Pennington et al., 2014] and GN-GloVe [Zhao et al., 2018b]. The dimension of the vectors is settled to 512 as standard and kept through all the experiments in this study. The parameter values for training the word embedding models with GloVe and GN-GloVe methods are listed in Table 5.3.

Parameter	Value
Vector size	512
Memory	4.0
Vocab. min. count	5
Max. iter.	15
Window size	15
Num. threads	8
X max.	10
Binary	2
Verbose	2

Table 5.3: Parameter values used for learning GloVe and GN-GloVe embeddings.

Debiaswe [Bolukbasi et al., 2016] is a debiasing post-process performed on trained embeddings. Instead of having parameters for learning the representation it uses a set of words to define the gender direction and

[2] `https://www.bls.gov/cps/tables.htm#empstat`

to neutralize and equalize the bias from the word vectors.

Three set of words are used in the Debiaswe algorithm. One set of ten pairs of words such as woman-man, girl-boy, she-he... are used to define the gender direction. Another set of 218 gender-specific words such as aunt, uncle, wife, husband... are used for learning a larger set of gender-specific words. Finally, a set of crowd-sourced male-female equalization pairs such as dad-mom, boy-girl, granpa-grandma... that represent gender direction are equalized in the algorithm. For the Spanish side, these sets of words have been adapted for the task. The sets are translated to Spanish and slightly modified to avoid unnecessary repetitions or unclear translations from specific English words. The seed words used in GN-GloVe has been similarly adapted to the Spanish language.

These pre-trained embeddings have been implemented with the codes from the GitHub repositories of GloVe[3], Debiaswe[4] and GN-GloVe[5].

5.2.2 Transformer models

The architecture to train the models for the translation task is the Transformer [Vaswani et al., 2017]. The evaluation of the performance of the model is obtained by its BLEU score [Papineni et al., 2002]. The parameter values used in the Transformer are the same as proposed in the baseline provided by the toolkit OpenNMT[6] and listed in Table 5.4. OpenNMT has built-in tools for training with pre-trained embeddings.

Parameter	Value	Parameter	Value
Layers	6	RNN size	512
Word vec. size	512	Transformer ff	2048
Heads	8	Encoder type	transformer
Decoder type	transformer	Position encoding	(used)
Train steps	200000	Max. generator batches	2
Dropout	0.1	Batch size	4096
Batch type	tokens	Normalization	tokens
Accum. count	2	Optim.	adam
Adam beta2	0.998	Decay method	noam
Warmup steps	8000	Learning rate	2
Max. grad. norm.	0	Param. init.	0
Param. init. glorot	(used)	Label smoothing	0.1
Valid. steps	10000	GPUs	4

Table 5.4: Parameter values used in the implementation of the Transformer.

5.3. Hardware

About the computational resources used for training the models, a cluster of four NVIDIA TITAN Xp GPUs and another cluster of four NVIDIA GeForce GTX TITAN GPUs were used separately and independently

[3] https://github.com/stanfordnlp/GloVe
[4] https://github.com/tolga-b/debiaswe
[5] https://github.com/uclanlp/gn_glove
[6] http://opennmt.net/

of the model. The duration of the training time was approximately 3 and 5 days, respectively. For the implementation, the model is set to accumulate the gradient two times before updating the parameters which simulates 4 more GPUs, i.e. giving a total of 8 GPUs, however the training takes longer.

Chapter 6

Results

In this chapter we give the performance of the models in terms of the BLEU score and compare them to the baseline. We also present the results from experiments on the use of debiased embedddings in the translation system by reporting the BLEU score for the *Occupations test* and doing a qualitative analysis on gender bias in some of the translations from this test.

6.1. Evaluation

The BLEU scores for the test set *newstest2013* are listed in Table 6.1. Note that we report the scores for the case of using pre-trained word embeddings on both the encoder and decoder and only in the encoder.

The scores for all models are around 30 BLEU points and the difference between the lowest and highest score is one point. The experiments with fixed embeddings in the models, i.e. preventing its values from being further updated during training, show a decrease in performance and are not further evaluated for the gender bias experiments. We focus our study on the models with updated pre-trained embeddings which are later used for a qualitative analysis on the impact of using debiased word embeddings in the translation system.

	BLEU	
Pre-trained embeddings	Updated	Fixed
None	29.78	-
GloVe	30.62	30.16
GloVe and Debiaswe	29.95	29.74
GloVe and Debiaswe (only encoder)	30.16	-
GN-GloVe	**30.74**	29.78
GN-GloVe (only encoder)	29.12	-

Table 6.1: BLEU scores for the *newstest2013* test set. English-Spanish. Fixed and updated word embeddings for training. Pre-trained embeddings used both in the encoder and decoder unless stated otherwise. Best results in bold.

Note that the best score is for the case of GN-GloVe embeddings, we achieve an improvement of almost one BLEU point. Moreover, the BLEU scores do not decrease for the other models embeddings, except for the case of GN-GloVe with pre-trained embeddings only in the encoder side. In the next subsection, we show how each of the models performs on gender debiasing, but we want to underline that these models do not decrease the quality of translation in terms of BLEU.

6.2. Discussion

A qualitative analysis is performed on the *Occupations test* set, examples of which are shown in Table 5.2. This test set has context information to translate the ambiguous word "friend" to its gender-specific translation in Spanish, either "amigo" or "amiga". The lowest the bias in the system, the better the system will be able to translate the correct gender. See Table 6.2 and Table 6.3 for percentages on how "friend" is translated in each model, for female-male pronouns and proper names, respectively. Note that we are using "updated" embeddings in all cases, since they are the best quality systems from Table 6.1.

The system translates the masculine pronoun with an accuracy of almost 100% for all models, despite not all occupations being fully translated well into Spanish. For the feminine pronoun, the accuracy of this task is not as precise and shows a decrease for all models, specially for GN-GloVe. See Table 6.2. Note that the percentages do not need sum exactly 100% since some predictions may be incomplete.

Pre-trained embeddings	her		him	
	amiga	amigo	amiga	amigo
None	99.8	0.2	0.1	99.9
GloVe (enc. and dec.)	96.4	0.0	0.0	99.7
GloVe and Debiaswe (enc. and dec.)	99.9	0.1	0.0	100.0
GloVe and Debiaswe (only encoder)	100.0	**0.0**	0.0	100.0
GN-GloVe (enc. and dec.)	99.6	0.4	0.0	100.0
GN-GloVe (only encoder)	100.0	**0.0**	0.0	100.0

Table 6.2: Percentage of "friend" being translated as "amiga" or "amigo" for female-male pronouns in test sentences from the *Occupations test*. Best results in bold.

For the case using a proper names like, "María" or "Juan", the accuracy on this task is similar between these names. Compared to the values for the female-male pronouns, the test with proper names show a decrease in accuracy for some models. See Table 6.3. Also, for the later models, the proper name "María" is not identified properly, thus giving a higher rate of "amigo" presence in the predictions with the exception of the model trained with debiased GloVe pre-trained embeddings.

Pre-trained embeddings	María		Juan	
	amiga	amigo	amiga	amigo
None	90.4	9.6	0.1	99.9
GloVe (enc. and dec.)	99.9	0.1	0.0	100.0
GloVe and Debiaswe (enc. and dec.)	98.8	**1.2**	0.7	99.3
GloVe and Debiaswe (only encoder)	8.8	91.2	0.0	100.0
GN-GloVe (enc. and dec.)	66.0	34.0	0.0	100.0
GN-GloVe (only encoder)	4.9	95.1	0.0	100.0

Table 6.3: Percentage of "friend" being translated as "amiga" or "amigo" for proper names in test sentences from the *Occupations test*. Best results in bold.

Note that gender debiasing is performed while keeping the quality of translation as shown in Table 6.1.

The most neutral model for this specific task is achieved with GloVe and Debiaswe pre-trained embeddings, updated during training.

However, the quality of the translations for each model not only depends on "friend" but also for the prediction of the occupations. Therefore, we proceed with a qualitative analysis on the predictions for each model. The predictions for the occupations we present show two qualities, the quality of the translation being closer or similar to a reference sentence and the gender of the occupation with respect to the subject.

For the male pronoun "him", most translations do not have trouble getting the gender of the profession right if the quality of the overall prediction is good. Therefore, in this analysis we focus on the female pronoun "her". There are exceptions, specially for occupations related to the medical field. See Table 6.8. Particularly, "nurse midwife", or even for "male nurse midwife", which got no acceptable translation for any model in the context of a male subject.

For the female pronoun "her", we present translations for sentences in the *Occupations test* that show more accuracy in predicting the gender of a given occupation than the baseline models. We comment on each instance.

For the sentence with the occupation "real estate assessor", the prediction with best quality is produced by the model with GN-GloVe pre-trained embeddings in the encoder and decoder sides. Together with the same embeddings only in the encoder side, these models are the ones giving the correct gender for the pronoun "her" for this test sentence. See Table 6.4.

Pre-trained word embeddings	Prediction *la conozco desde hace mucho tiempo,*
None	*mi amiga trabaja como evaluador de bienes inmuebles.*
GloVe (enc. and dec.)	*mi amiga actúa como assessor de bienes raíces.*
GloVe with Debiaswe (enc. and dec.)	*mi amiga trabaja como asesor inmobiliario.*
GloVe with Debiaswe (only encoder)	*mi amiga trabaja como assessor de bienes raíces.*
GN-GloVe (enc. and dec.)	*mi amiga trabaja como* **asesora inmobiliaria**.
GN-GloVe (only encoder)	*mi amiga trabaja como especialista en bienes raíces.*
Reference	*mi amiga trabaja como asesora inmobiliaria.*

Table 6.4: Spanish predictions for the test sentence "*I've known her for a long time, my friend works as a real estate assessor.*". Best results in bold.

In Spanish, both "gerente" or "gerenta" is accepted for the translation of "manager". However, GN-GloVe pre-trained embeddings give a more explicit translation for the gender of "transportation manager". See Table 6.5. Debiased GloVe pre-trained embeddings in only the encoder side also give a good translation on this test sentence.

Pre-trained word embeddings	Prediction *la conozco desde hace mucho tiempo,*
None	*mi amiga trabaja como gestor de transportes.*
GloVe (enc. and dec.)	*mi amiga trabaja como gestor de transporte.*
GloVe with Debiaswe (enc. and dec.)	*mi amiga trabaja como gerente de transporte.*
GloVe with Debiaswe (only encoder)	*mi amiga trabaja como* **gerente de transportes**.
GN-GloVe (enc. and dec.)	*mi amiga trabaja como* **administradora de transportes**.
GN-GloVe (only encoder)	*mi amiga trabaja como gerente de transporte.*
Reference	*mi amiga trabaja como gerente de transportes.*

Table 6.5: Spanish predictions for the test sentence "*I've known her for a long time, my friend works as a transportation manager.*". Best result in bold.

The sentence for "auditing clerk" has better quality in the models with debiased GloVe embeddings, and also for GN-GloVe pre-trained embeddings than the baseline models. The gender prediction of these models is correct for the Spanish translation of this profession. See Table 6.6.

Pre-trained word embeddings	Prediction *la conozco desde hace mucho tiempo,*
None	*mi amiga trabaja como empleado de auditoría.*
GloVe (enc. and dec.)	*mi amiga trabaja como clerk.*
GloVe with Debiaswe (enc. and dec.)	*mi amiga trabaja como* **secretaria de auditoría.**
GloVe with Debiaswe (only encoder)	*mi amiga trabaja como* **secretaria de auditoría.**
GN-GloVe (enc. and dec.)	*mi amiga trabaja como* **empleada de auditoría**.
GN-GloVe (only encoder)	*mi amiga trabaja como asistente de auditoría.*
Reference	*mi amiga trabaja como trabaja de auditora.*

Table 6.6: Spanish predictions for the test sentence "*I've known her for a long time, my friend works as an auditing clerk.*". Best results in bold.

The gender prediction for "hazardous materials removal worker" is accurate for all the models with pre-trained embeddings. The quality of translations is similar, and the models other than the baseline show good quality. See Table 6.7.

Pre-trained word embeddings	Prediction *la conozco desde hace mucho tiempo, mi amiga trabaja como*
None	*trabajador de remoción de materiales peligrosos.*
GloVe (enc. and dec.)	*trabajadora para la eliminación de materiales peligrosos.*
GloVe with Debiaswe (enc. and dec.)	**trabajadora de remoción de materiales peligrosos**.
GloVe with Debiaswe (only encoder)	**trabajadora de remoción de materiales peligrosos**.
GN-GloVe (enc. and dec.)	**trabajadora de remoción de materiales peligrosos**.
GN-GloVe (only encoder)	**trabajadora de remoción de materiales peligrosos**.
Reference	*trabajadora de remoción de materiales peligrosos.*

Table 6.7: Spanish predictions for the test sentence "*I've known her for a long time, my friend works as a hazardous materials removal worker.*". Best results in bold.

The last instance we present, "health diagnosing practitioner", gives no prediction for the baseline with pre-trained embeddings. The translations for the debiased GloVe and GN-GloVe pre-trained embeddings show the best quality. However, the Spanish words in the professions are gender neutral in this case.

Pre-trained word embeddings	Prediction *la conozco desde hace mucho tiempo, mi amiga*
None	*trabaja como médico tratante.*
GloVe	. (no prediction)
GloVe with Debiaswe	*trabaja como médico.*
GloVe with Debiaswe (only encoder)	*trabaja como* **especialista en diagnóstico de salud**.
GN-GloVe	*trabaja como* **profesional de diagnóstico sanitario**.
GN-GloVe (only encoder)	*trabaja como médico tratante de salud.*
Reference	*trabaja como médica especialista en diagnóstico de la salud.*

Table 6.8: Spanish predictions for the test sentence "*I've known her for a long time, my friend works as a health diagnosing practitioner.*". Best results in bold.

Chapter 7

Conclusion and future work

Biases learned from human generated corpora in NLP applications is a topic that has been gaining relevance over the last few years. Tools from machine learning provide approaches suited for this specific task. Specifically, for machine translation, studies quantifying gender bias present in news corpora and proposing debiasing approaches for word embedding models have shown improvements on this matter.

We proposed a framework for studying the impact of debiased word embeddings on a neural network architecture. We trained sets of word embeddings with the standard GloVe algorithm. Then, we debiased these embeddings using Debiaswe. We also trained a gender neutral version with GN-GloVe. We used all these different models on the Transformer. Experiments were reported on using these embeddings on both the encoder and decoder sides, or only the encoder side. In all cases, we trained our models using an open-source implementation OpenNMT.

The models are evaluated using the BLEU metric for the WMT *newstest2013* test set. The performance is around 30 BLEU points with a difference of one BLEU point for the lowest and highest BLEU score.

For the study of the bias for the translations of the models, we developed a specific test set named *Occupations test*. This set consists of sentences that includes context of the gender of the ambiguous "friend" in the English-to-Spanish translation. This word can be translated to feminine or masculine and the proper translation has to be derived from context. Our hypothesis is that if the system is gender biased, the context will be disregarded, while if the system is neutral, the translation will be correct. Results show that the male pronoun is always identified, despite not all occupations being well translated, while the female pronoun has different ratio of appearance for the models. Successfully, we achieve almost 100% accuracy of translation on this pronouns when using the debiased word embeddings on the encoder and decoder sides. The qualitative analysis performed on several predictions shows that the quality of the translations can increase when these pre-trained embeddings are used in the Transformer, and that debiased and gender-neutral embeddings aid the prediction of the gender in the occupations for the Spanish translation. Also, this system does not decrease the BLEU performance from the baseline translation system. Therefore, we are "equalizing" the translation, while keeping its quality.

As already mentioned, this is the first work on proposing gender debiased translation systems. However, further work on debiasing translation algorithms is required.

In this thesis we studied gender as a bias in machine translation, however other social constructs and stereotypes may be present in corpora, whether individually or combined, such as race, religious beliefs or age; this being just a small subset of possible biases which will present new challenges for fairness both in ML and MT. For the task in hand, the type of corpora used for the study is commonly related to news articles. Human corpora has a broad spectrum of categories, as an instance: industrial, medical, legal... and biases particular to each area may present other interesting problems. Also, other language pairs with different degree in specifying gender information in their written or spoken communication could be studied for the evaluation of debiasing in machine translation.

Bibliography

[Bolukbasi et al., 2016] Bolukbasi, T., Chang, K., Zou, J. Y., Saligrama, V., and Kalai, A. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *CoRR*, abs/1607.06520.

[Caliskan et al., 2017] Caliskan, A., Bryson, J. J., and Narayanan, A. (2017). Semantics derived automatically from language corpora necessarily contain human biases. *CoRR*, abs/1608.07187.

[Hochreiter et al., 2001] Hochreiter, S., Bengio, Y., and Frasconi, P. (2001). Gradient flow in recurrent nets: the difficulty of learning long-term dependencies. In Kolen, J. and Kremer, S., editors, *Field Guide to Dynamical Recurrent Networks*. IEEE Press.

[Klein et al., 2017] Klein, G., Kim, Y., Deng, Y., Senellart, J., and Rush, A. M. (2017). OpenNMT: Open-source toolkit for neural machine translation. In *Proc. ACL*.

[Koehn, 2005] Koehn, P. (2005). Europarl: A Parallel Corpus for Statistical Machine Translation. In *Conference Proceedings: the tenth Machine Translation Summit*, pages 79–86, Phuket, Thailand. AAMT, AAMT.

[Mikolov et al., 2013] Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

[Papineni et al., 2002] Papineni, K., Roukos, S., Ward, T., and jing Zhu, W. (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.

[Park et al., 2018] Park, J. H., Shin, J., and Fung, P. (2018). Reducing gender bias in abusive language detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 2799–2804.

[Pennington et al., 2014] Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

[Prates et al., 2018] Prates, M. O. R., Avelar, P. H. C., and Lamb, L. C. (2018). Assessing gender bias in machine translation - A case study with google translate. *CoRR*, abs/1809.02208.

[Sutskever et al., 2014] Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to sequence learning with neural networks. *CoRR*, abs/1409.3215.

[Vaswani et al., 2017] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. (2017). Attention is all you need. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

[Zhao et al., 2017] Zhao, J., Wang, T., Yatskar, M., Ordonez, V., and Chang, K. (2017). Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *EMNLP*, pages 2979–2989. Association for Computational Linguistics.

[Zhao et al., 2018a] Zhao, J., Wang, T., Yatskar, M., Ordonez, V., and Chang, K. (2018a). Gender bias in coreference resolution: Evaluation and debiasing methods. In *NAACL-HLT (2)*, pages 15–20. Association for Computational Linguistics.

[Zhao et al., 2018b] Zhao, J., Zhou, Y., Li, Z., Wang, W., and Chang, K. (2018b). Learning gender-neutral word embeddings. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 4847–4853.

Appendix A

Sample sentences from data sets

Sample sentences for the train, validation and test data sets are given in Table A.1.

(En) with regard to the situation of children with disabilities, the Committee remains concerned at the lack of adequate infrastructure, at the limited qualified staff and specialized institutions for these children, and at the absence of adequate resources, both financial and human. (Es) en cuanto a la situación de los niños discapacitados, el Comité sigue preocupado por la falta de una infraestructura suficiente, la limitada cantidad de personal calificado y de instituciones especializadas para estos niños y la ausencia de recursos adecuados, tanto financieros como humanos.
(En) Mr. Peter Kenmore was appointed Co-Executive Secretary of the Rotterdam Convention in February 2007. (Es) el Sr. Peter Kenmore fue nombrado co-secretario Ejecutivo del Convenio de Rotterdam en febrero de 2007.

(a) Sample sentences from training data set.

(En) Buffett claims to have paid 170 dollars on average per share and now has a 5.5 percent stake. (Es) tras algunas indicaciones, Buffett pagó 170 dólares de promedio por cada unidad y mantuvo el 5,5 por ciento.
(En) after a brief rest at the end of last week, when it became clear that Berlusconi would resign, the Italian debt costs have now reached critical levels between uncertainties about whether the new prime minister will succeed. (Es) tras un breve respiro a finales de la semana pasada, cuando quedó claro que Berlusconi dimitiría, los costes de la deuda italiana han vuelto ahora a niveles críticos entre la incertidumbre sobre si el nuevo primer ministro tendrá éxito.

(b) Sample sentences from validation set *newstest2012*.

(En) however, with time, it will become reality, when earthlings will go to a faraway planet in their ships, and it will become the home for their offspring, who were born in space. (Es) sin embargo, con el tiempo esto será una realidad, cuando los seres humanos en sus naves se dirijan a un planeta lejano que será el hogar para sus hijos nacidos en el espacio.
(En) one demonstrator at the Tahrir warned: "you are letting loose a monster that you can no longer control." (Es) un manifestante en la plaza Tahrir advierte: "estáis concibiendo un monstruo que no podréis controlar".

(c) Sample sentences from test set *newstest2013*.

Table A.1: Sample sentences from train, validation, and test sets. English (En) and Spanish (Es).

Appendix B

Data related to predicting the gender for some occupations

Here we gather tables related to the translations of the professions in the *Occupations test*.

Some of the professions used to built this test set have a main word upon which the translation is based. As an instance, the word "engineer" appears 26 times in specific professions for sentences in the *Occupations test*. Some examples are "aerospace engineer", "agricultural engineer", "biomedical engineer"... For this particular word, all professions are translated with the masculine word "ingeniero" in Spanish, "ingeniero aerospacial", "ingeniero agrónomo", "ingeniero biomédico"... See Table B.3.

However, the distribution of these appearances for different professions and for each of our trained models gives some information on their performance. We include tables for core words such as "professor", "scientist", "technician", "director", "worker", "clerk" which built more specific professions. Their respective tables includes the number of times their translation to Spanish is female (normally with "-a") or male (normally with "-o") for the subjects "her", "him", "María" and "Juan" which are used in the analysis of gender bias in the translation task. The number of occurrences of each words are given in the caption of the tables, together with an example of an specific profession for each word. Notice that some of the translations may be incomplete despite these more common words being translated well.

Pre-trained embeddings	profesora				profesor			
	her	him	María	Juan	her	him	María	Juan
None	2	0	2	0	5	7	5	10
GloVe	10	0	10	0	0	9	0	8
GloVe and Debiaswe	7	0	7	0	0	13	0	13
GloVe and Debiaswe (only encoder)	7	0	2	0	0	9	6	8
GN-GloVe	16	0	12	0	0	16	4	16
GN-GloVe (only encoder)	13	0	0	0	0	14	14	15

Table B.1: Number of times the occupations including the word "teacher" or "professor" are translated as "profesora" or "profesor" for 16 occurrences in the sentences from the *Occupations test*. E.g. "university professor".

Pre-trained embeddings	científica				científico			
	her	him	María	Juan	her	him	María	Juan
None	8	0	2	0	6	14	12	14
GloVe	11	1	2	1	1	14	11	13
GloVe and Debiaswe	9	0	12	0	3	12	0	12
GloVe and Debiaswe (only encoder)	1	0	0	0	13	14	14	14
GN-GloVe	9	0	2	0	3	14	10	14
GN-GloVe (only encoder)	14	0	3	0	0	14	11	14

Table B.2: Number of times the occupations including the word "scientist" are translated as "científica" or "científico" for 15 occurrences in the sentences from the *Occupations test*. E.g. "computer research scientist".

Pre-trained embeddings	ingeniera				ingeniero			
	her	him	María	Juan	her	him	María	Juan
None	0	0	0	0	26	26	26	26
GloVe	0	0	0	0	26	26	26	26
GloVe and Debiaswe	0	0	0	0	26	26	26	26
GloVe and Debiaswe (only encoder)	0	0	0	0	26	26	26	26
GN-GloVe	0	0	0	0	26	26	26	26
GN-GloVe (only encoder)	0	0	0	0	25	26	26	26

Table B.3: Number of times the occupations including the word "engineer" are translated as "ingeniera" or "ingeniero" for 26 occurrences in the sentences from the *Occupations test*. E.g. "biomedical engineer".

Pre-trained embeddings	técnica				técnico			
	her	him	María	Juan	her	him	María	Juan
None	0	0	0	0	37	20	35	11
GloVe	20	0	9	0	14	9	26	5
GloVe and Debiaswe	0	0	0	0	34	20	34	14
GloVe and Debiaswe (only encoder)	0	0	0	0	34	11	2	5
GN-GloVe	0	0	0	0	20	25	10	20
GN-GloVe (only encoder)	0	0	0	0	34	12	11	9

Table B.4: Number of times the occupations including the word "technician" are translated as "técnica" or "técnico" for 37 occurrences in the sentences from the *Occupations test*. E.g. "wind turbine service technician".

Pre-trained embeddings	directora				director			
	her	him	María	Juan	her	him	María	Juan
None	0	0	3	0	6	9	3	5
GloVe	22	0	18	0	0	7	0	5
GloVe and Debiaswe	14	0	8	0	1	14	1	8
GloVe and Debiaswe (only encoder)	5	0	0	0	1	6	6	6
GN-GloVe	6	0	4	0	1	13	3	10
GN-GloVe (only encoder)	6	0	0	0	0	8	4	8

Table B.5: Number of times the occupations including the word "manager" or "director" are translated as "directora" or "director" for 22 occurrences in the sentences from the *Occupations test*. E.g. "industrial production manager".

Pre-trained embeddings	trabajadora				trabajador			
	her	him	María	Juan	her	him	María	Juan
None	59	0	12	0	16	82	61	84
GloVe	78	0	80	1	0	70	0	58
GloVe and Debiaswe	76	0	82	0	0	80	0	75
GloVe and Debiaswe (only encoder)	79	0	7	0	2	58	56	59
GN-GloVe	77	0	60	0	1	76	13	75
GN-GloVe (only encoder)	82	0	0	0	1	82	81	81

Table B.6: Number of times the occupations including the word "worker" are translated as "trabajadora" or "trabajador" for 84 occurrences in the sentences from the *Occupations test*. E.g. "machinery maintenance worker".

Pre-trained embeddings	secretaria				secretario			
	her	him	María	Juan	her	him	María	Juan
None	2	0	4	0	0	1	1	1
GloVe	1	0	2	0	0	1	0	2
GloVe and Debiaswe	37	0	10	0	0	3	0	3
GloVe and Debiaswe (only encoder)	15	0	0	0	0	1	2	2
GN-GloVe	1	0	0	0	0	1	1	1
GN-GloVe (only encoder)	1	0	0	0	0	1	1	1

Table B.7: Number of times the occupations including the word "clerk" or "secretary" are translated as "secretaria" or "secretario" for 37 occurrences in the sentences from the *Occupations test*. E.g. "judicial law clerk".