Measuring gender bias in machine translation beyond sentence level

Natalie Ronco, L6th HAC

Rouse Award: Essay, Computer Science

Supervisor: Dr Denes

Word count: 4211

Acknowledgements

I would like to thank Dr Denes for his support and guidance throughout the research process.

Abstract

Machine translation (MT) is becoming increasingly developed and widely used, however, it can contain gender biases which are harmful to the user and society more generally. Fundamentally, training data is the origin of this bias, however, bias can be amplified by the MT model. I demonstrate qualitatively, using Google Translate from English to Spanish, ways in which the input can be adjusted to cause bias to arise. In particular, I show that the sentence level architecture commonly used in MT can introduce bias in translation, as it leads to a lack of context for the model. I also propose a framework which utilises existing natural language processing techniques in a novel application: to potentially compare gender bias between MT and human translation holistically over a large dataset. This framework, despite providing inconclusive results due to a limited dataset size, may in future be further developed to evaluate gender bias of MT models.

1 Introduction

In November 2022, Open AI released chat GPT [1], a large language model that has brought natural language processing (NLP), a subset of machine learning, to the attention of the general public. This is just one of many advancements in machine learning which have occurred in recent years. NLP is widely used in many applications. For example: spellcheck, sentence completion in search engines, data analytics, chatbots, and translation. However, many NLP systems have been shown to reflect and even exacerbate societal bias, such as gender bias. As they become increasingly utilised, alleviating this bias is a fundamental step in reducing inequality and removing barriers to opportunities for billions of people around the world.

Machine translation (MT), the focus of this paper, is one area in which this gender bias has been shown to arise, leading to negative consequences for the user [2]. Translation can introduce bias due to word choice, ambiguous context, or imbalance in error rates between translating into different genders. In an increasingly globalised society, MT is likely to be increasingly utilised. Therefore, it is essential that we address challenges, as bias in MT can negatively affect people in society, especially those who may rely on translation services.

This research focuses on gender bias in machine translation from English to Spanish. Firstly, I conduct qualitative analysis using Google Translate. I go on to develop the foundations of a novel approach to quantifying this bias on a large scale, by comparing human and machine translation. This work predominantly investigates female/ male bias, due to exploration via grammatical gender, however, I recognise that gender is non-binary.

2 Background and related work

In this section, I explain the principles of NLP and MT as well as providing context on linguistic structures of languages. Then, I explore existing approaches into investigating and removing gender bias in NLP and MT.

2.1 Natural language processing (NLP) overview

NLP is a branch of machine learning which involves algorithms that can analyse text based data. Initial NLP required a complex set of hand-written rules about human language to be coded into the algorithm. This has developed to using statistical analysis and neural networks to allow NLP algorithms to infer rules from a training dataset, and then apply these to unfamiliar situations [3]. Most state-of-the-art techniques use deep learning NLP, which involves neural networks with multiple layers. These can be convolutional neural networks (CNNs), which use a hierarchical function to combine inputs and extract features or recurrent neural networks (RNNs), which process data sequentially, using previous outputs when calculating the next output [4]. More advanced techniques use transformer models, which process multiple inputs in parallel and use self-attention to represent how inputs relate to each other [5]. Google Translate and Chat GPT are both examples of applications of NLP.

2.2 Machine translation (MT) overview

MT architectures vary, however the basic principles remain the same. They require training on many human-translated pairs of sentences in the initial and target language. From this training, each word is assigned a multidimensional vector, known as a word embedding, which captures information from the training data about the likely translations of that word. The model can then be deployed for use in unfamiliar contexts.

The latest models use neural machine translation (NMT). This technique translates sentences holistically by using the vector that corresponds to the target word in combination with output vectors for other vectors in the sentence to generate a result. This allows them to capture useful information such as the gender of nouns and the tone, as well as rejecting unnecessary words and changing the word order [6].

2.3 Linguistic gender structures: is Spanish inherently gender biased?

When examining gender in translation, it is important to consider the varying gender structure between different languages. Stahlberg et al. [7] categorise languages into the following groups based on their gender structure:

Genderless languages, e.g. Hungarian, Turkish, and Persian. These languages contain gender references limited to basic lexical pairs, such as brother/sister. Translations of pronouns, for example he/she, would not be different with respect to gender.

Natural gender languages, e.g. English. These languages contain gender-specific pronouns, such as he/she, in addition to basic lexical gender pairs.

Grammatical gender languages, e.g. Spanish, Hindi and German. As well as gender pronouns and basic lexical pairs, these languages assign a grammatical gender to every noun. For most nouns, such as tree or water, the grammatical gender does not connect to a semantic meaning. However, in the vast majority of personal nouns the grammatical gender corresponds to the meaning. For example, in Spanish “mujer” (woman) is feminine whereas “hombre” (man) is masculine. Adjectives, articles and pronouns can also

"agree" with the gender of the noun they describe. These languages are not limited to just expressing masculine and feminine, for example German also contains the neuter gender.

This paper focuses on analysing gender bias in translation from English into Spanish. In Spanish the grammatical gender determines the ending of nouns and adjectives. The most common example is that "-o" or "-os" is used for masculine, and "-a" or "-as" is used for feminine.

It is possible that gender bias in machine translation into Spanish is partly caused by linguistic constraints. Spanish uses the masculine plural when referring to a group of people, even if it were 100 women and one man. In this way, the language itself is biased towards men. It is also strongly biased against non-binary individuals as it does not allow of gender neutral alternatives in language. Informally, the @ symbol or x can be used to represent o and a simultaneously to serve as a neutral ending in written Spanish. In spoken Spanish, the use of e as an alternative ending has been proposed, as well as the gender neutral pronoun "elle", equivalent to the singular use of "they" in English, which would provide of solution to this challenge. However, currently this usage is not recognised in the official Real Academia Espanola Spanish dictionary [8].

2.4 Biased word embeddings from vector geometry

Part of the machine translation process involves assigning multidimensional vectors, known as word embeddings, to each word. These word embeddings aim to capture relative meaning of words: if two words are close together in vector space, then they are likely to appear in similar contexts and therefore are semantically similar. However, these embeddings, trained by algorithms such as word2vec, have been shown to reflect societal gender bias by creating asymmetries in vector geometry. Notably Bolukbasi et al. [9] highlighted this in English word embeddings and Zhou et al. [10] extended these ideas into languages with grammatical gender, including Spanish. It is likely that this bias arises due to unbalanced training data, which reflects historical gender bias, e.g. it may contain more references to male doctors than female doctors. Various ideas have been suggested to reduce this bias, for example changing the relative position of embeddings in vectors space [9]. Other methods also include using constraints to create a more balanced training corpus [11] and using a gender-equalizing loss function upon training [12].

2.5 Bias affected by architecture choice

Although training data is the underlying cause of bias in machine translation, architectural choices can also amplify bias. Costa-Jussà et al. have shown that varying model architecture alone affects bias for NMT [13]. Gender tagging within a sentence is one method which has been shown to improve accuracy in translating gender, and this can be adapted to allow a translation system to work with the use of gender neutral endings, i.e. "buene", if these were to become widely used and recognised [14].

3 Qualitative analysis

Gender bias can appear in MT in several ways: gender choice in ambiguous context, inaccuracy imbalance, and word choice. I investigate these possibilities qualitatively, using Google Translate [15], in order to demonstrate some scenarios in which gender bias occurs in MT from English to Spanish.

Firstly, differences in gender structure between languages introduce some challenges. For example, when translating from English into Spanish there is a requirement to specify the gender of nouns described, since there is no gender neutral form, even if it is ambiguous in the English context. Therefore, by investigating which gender the Google Translate algorithm choses to assign, I can investigate whether the word embeddings it uses are likely to be biased.

Google translate has attempted to combat this issue in their MT algorithm by displaying two alternative translations, allowing the ambiguous entry to be displayed as both feminine and masculine [16] (Fig. 1). However, this approach is currently only a surface level correction, and once the sentence becomes longer it returns to a stereotypical default (Fig. 2).

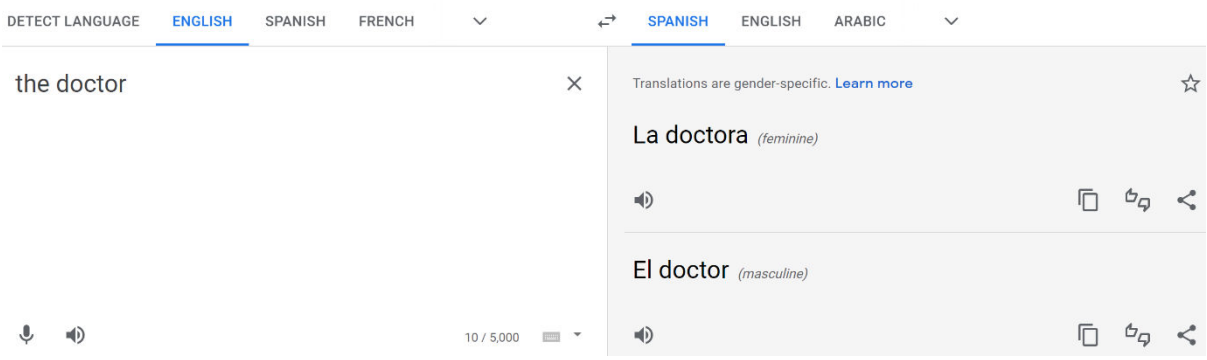

Figure 1: when presented with a short entry Google Translate displays both masculine and feminine endings

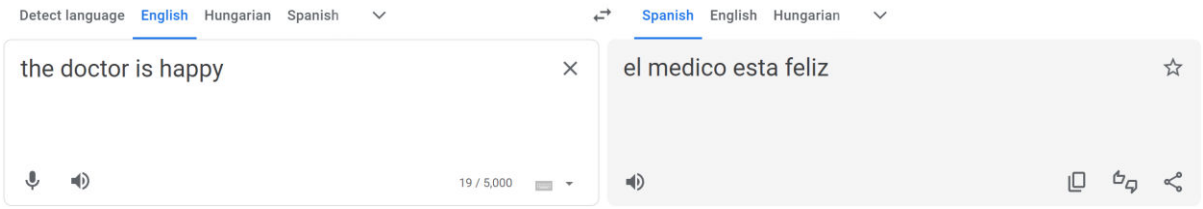

Figure 2: when the entry is longer this correction no longer occurs and Google Translate displays gender bias. It refers to the doctor in masculine. Note: "medico" and "doctor" both translate to doctor in English (it does suggest "el doctor esta feliz" as an alternative translation for this entry)

Gender bias can also occur due to different levels of inaccuracy. For example, a MT algorithm would be biased if it translated sentences involving masculine nouns more accurately than those involving feminine. Google Translate (Fig. 2) displays biased results when provided with ambiguous context in a short input. However, this can also occur in longer texts depending on how much of the surrounding text the algorithm takes into account upon translation of a given word. Google translate uses a hybrid RNN and transformer model [17], in which it uses the current sentence as context when translating a given word [18]. This is different to a human translator, who would use surrounding sentences or paragraphs if faced with ambiguity. This means the MT algorithm can still encounter cases of ambiguous gender, as in the google translate sentences above, within a much larger document, as it ignores everything outside of the current sentence.

For example, if our first sentence is "this is a story about a programmer", both a human and a machine translator into Spanish would be faced with ambiguity as to whether "programmer" should be translated into masculine or feminine. However, if the second sentence is "she is disappointed because she has a bug in her code", then a human translator would realise the gender used in the first sentence should be feminine whereas a machine translator might treat the second sentence separately, meaning any assumption in the first sentence would remain the same. As shown in Fig. 3, sometimes just combining two sentences into one can improve accuracy on the machine translator.

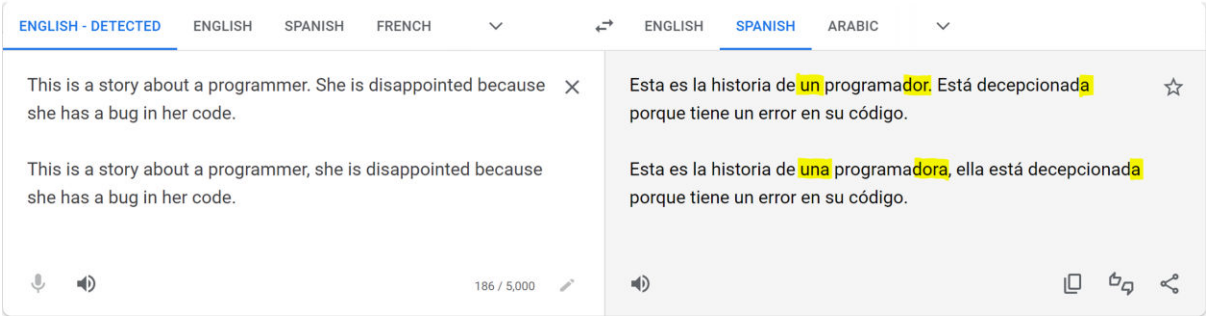

Figure 3: in the top entry Google Translate incorrectly translates the first sentence by using masculine to describe the programmer, despite the second sentence referring to her in the feminine. In the bottom entry, the phrases have been separated by a comma so they are now treated as the same sentence. This leads to "programmer" being correctly translated using feminine. I have added highlighting on the endings.

Here, the initial assumption that the programmer would be male is likely caused by bias word embeddings due to biased training data. However, the inability to correct it when more context is provided is due to the architecture, which focuses on sentence level translation. There is existing research which could begin to solve this problem. For example, Wang et al. [19] and Miculicich et al. [20] have developed models for carrying over context between sentences within the attention layer in NMT software. With further development, similar models could perhaps be applied to the Google Translate software, which I suggest would likely reduce gender bias caused by inaccuracies in translation.

These findings also lead me to suggest a limitation of current translation accuracy evaluators. Bilingual evaluation understudy (BLEU) is a common algorithm used to calculate accuracy of MT, however this only looks at accuracy of individual sentences [21]. When Google released their NMT model in 2016, they measured its accuracy based off BLEU scores in combination with human evaluation, also done on a sentence level [18]. In addition, a common method of assessing gender bias in MT is WinoMT [22]. This is a challenge set which uses coreference resolution to determine accuracy, however this also focuses on accuracy of individual sentences. These methods may result in inaccuracies, such as in Fig. 3, being missed. Therefore, the algorithm may seem more accurate, and less biased, than it actually is. It is important to find ways of assessing gender bias, or accuracy more generally, across the entire output instead of at a sentence level. I aim to do this with my methodology in the quantitative analysis below.

Aside from inaccuracies, machine translation could also be gender biased in its word choice. For example, does the translation selected for a given word change between when it is spoken by a woman or spoken by a man? This is likely to be the case due to bias within the original vector embeddings, however this effect would most likely be much more subtle when compared against the biased translation errors mentioned above. This area requires further investigation, however my methodology below could act as a framework for doing this.

4 Quantitative analysis

In light of these qualitative findings using Google Translate, I also conducted a quantitative analysis of gender bias in MT from English to Spanish. My methodology uses a word2vec algorithm [23] alongside vector processing techniques to investigate gender bias [9]. By performing this analysis on a dataset of MT output, I demonstrate the framework for this novel end-to-end approach. The measurement of bias across a dataset allows me to assess gender bias holistically instead of sentence level, to further investigate my qualitative findings. Additionally, applying these techniques to MT allows direct comparison between human and machine translation, overcoming challenges relating to where in the process the bias arises (later explained in Fig. 7)

Briefly, my methodology can be summarised as:

4.1) Collect 2 datasets. English to Spanish MT, and English to Spanish human translation using the same original texts.

4.2) Process these datasets using a word2vec algorithm [23] to create 2 sets of multidimensional vectors of the most common words from the datasets.

4.3) Analyse these vector sets using techniques similar to Bolukbasi et al. [9] to quantify levels of bias in certain word pairings.

4.4) Calculate the difference in levels of bias between human and machine datasets, to investigate whether there is a consistent and statistically significant difference.

Below, I explain these steps in more detail and provide my results. The code I used can be found here: https://github.com/natalie1247/Rouse

4.1) Dataset collection:

I constructed a dataset of 10 books, all originally written into English, by using the PDF downloads from InfoBooks [24] (see appendix for list of titles). These were chosen somewhat arbitrarily with my requirements being that they were in an easily usable PDF format, had Spanish translations and were no longer under copyright. I then collected two datasets of translations into Spanish, one using Google's machine translation tool [15] and the other from the human-translated Spanish versions from InfoLibros [25].

4.2) Dataset processing with word2vec:

Then, in order to generate sets of vectors for both human and machine translation, which I could compare to investigate varying levels of gender bias, I ran a separate analysis on each of the Spanish datasets. For this, I used a word2vec-style machine learning algorithm to represent the 4095 most common words in the dataset.

The code I used was adapted from a TensorFlow tutorial on word2vec [26]. This code is an implementation of the continuous skip gram model which was developed by Mikolov et al. in 2013 [23] [27]. I generated 128 dimensional vectors to represent the most common words in the dataset. The algorithm generates training examples of pairs of words that appear close to each other, referred to as skip grams, as well as negative skip grams that do not. It then uses these training examples to adjust the vectors so they can be used to indicate the probability of context words appearing within a context window of the target word. The relationships between the resulting vectors aim to convey relationships between words. For example, synonyms should appear close together in vector space and vector subtraction can also reveal useful analogies.

I adapted the existing code by changing the epochs and number of negative samples. I adjusted the values manually to reduce loss within a reasonable time frame. In addition, I carried out some pre and post processing in order to convert the dataset into a format so that it could be processed by the model. As part of this, I removed all Spanish accents, as these came up as unrecognised characters. I was able to do this with Spanish since in the majority of cases removing an accent does not make the meaning ambiguous. Therefore, when I refer to Spanish words below, I sometimes write them without accents even though this is not standard Spanish.

Since machine learning algorithms have a random element, due to the random initialisation of the vectors, and my datasets were limited in size, I ran the word2vec algorithm 10 times for both human and machine translation sets in order to capture the variation in the output vectors. My resultant sets of

vectors shows some useful relationships geometrically. For example, in my first vector set from using the machine translation dataset the nearest neighbour to “hermana” (sister) is “prima” (female cousin) and the nearest neighbour to “noche” (night) is “oscuridad” (darkness), so this shows some useful semantic relationships.

However, it is important to emphasise that due to the limited size of my dataset and limited number of training epochs, there is a lot of variation between these relationships across different vector sets from the same dataset. There are also more seemingly random close matches, for example “noche” (night) is also close to “playa” (beach) which likely arises due to the specific dataset so is unlikely to reflect general usage (Fig. 4)

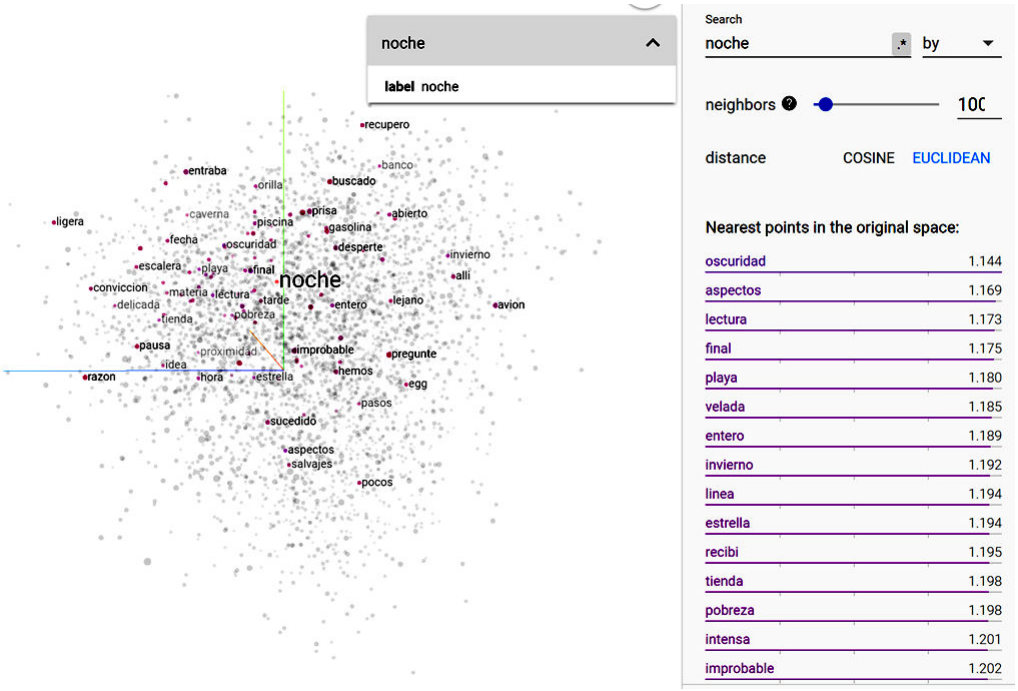

Figure 4: using the TensorFlow embedding projector software [28] to visualise one of my vector sets from the machine translate dataset in 3D. Each dot represents a word embedding and those in purple, with the word marked, are the closest by Euclidean distance to the target word “noche” (night). “oscuridad” (darkness) is close to night, reflecting sematic similarity, whereas “playa” (beach) is also close which seems random.

4.3) Measuring bias in these vectors:

The proximity of words, measured using the Euclidean distance, can be used to indicate semantic similarity. Therefore, an idea introduced by Bolukbasi et al. [9] and used by Qian et al. [12] alongside many others, is that we can create gender pairings, e.g. mother/ father or sister/ brother, which have the same meaning apart from referring to a different gender. We would expect most words to be gender neutral, and therefore they should be equidistant to each word from the gender pairing. If a word appears much closer to all of the female words from the pairings than the male ones, it is considered to be gender biased towards female, and vice versa.

Following Bolukbasi et al. [9] I created my list of gender pairings. For my experiment these were: senor/ senora, hombre/ mujer, hermano/ hermana, hijo/ hija, and padre/ madre. These translate in English to: Mr/ Mrs, man/ woman, brother/ sister, son/ daughter and father/ mother.

In other papers, such as Bolukbasi et al. [9], the authors use occupation words for comparison against the gender pairings. However, in Spanish this technique is less useful because different words are used when referring the men and women due to gender agreement. Also, I was limited to what was included in both datasets. For example, doctora or medica (female doctor), ingeniero/ ingeniera (engineer) and enfermero/ enfermera (nurse), all did not appear. Additionally, I aimed to select words which were

relatively common in the dataset to reduce noise in my results. The list I compiled was: poder (power/ to be able to), matrimonio (marriage), fuerte (strong), llorar (to cry), exito (success), trabajar (to work), decidio (he/she decided), gana (he/she wins).

In retrospect, this word list has various limitations. Some of the words are nouns (matrimonio, exito, poder) and therefore have grammatical gender so it would not necessarily be biased if these were closer to words with the same grammatical gender. The Mr/ Mrs word pairing is also less useful since Mrs contains information about marital status whereas Mr does not (this makes the comparison with "marriage" particularly unbalanced). Some words also have other possible translations depending on the context. For example, "de buena gana" can mean "enthusiastically" and "fuerte" can mean "healthy" or "fort". I believe "llorar", "decidio" and "trabajar" remain useful, however, with a larger dataset a larger and more robust selection of pairings could be made.

When calculating distances between the test word and word from the gender pair, I used Euclidean distance. Cosine distance could also be used as a potentially useful extension of this work.

This initial stage showed that in both human and machine translated text we can see gender bias in some of the words chosen. On average, the vectors for "poder" and "exito" are biased towards male and "llorar" and "gana" are biased towards female. As mentioned above, "llorar" is probably the most useful of these. This is shown in Spanish (Fig. 5) and with the English translations (Fig. 6)

HUMAN TRANSLATION	poder	matrimonio	fuerte	llorar	exito	trabajar	decidio	gana
senor/ senora	m	m	f	f	f	f	f	f
mujer/ hombre	f	f	m	f	m	f	f	f
padre/ madre	m	m	m	f	m	m	m	f
hermana/ hermano	m	f	f	f	m	m	m	f
hija/ hijo	m	m	m	f	m	m	m	f
MACHINE TRANSLATION	poder	matrimonio	fuerte	llorar	exito	trabajar	decidio	gana
senor/ senora	m	m	f	f	m	f	f	f
mujer/ hombre	m	f	m	f	m	m	f	f
padre/ madre	m	f	f	f	m	f	m	f
hermana/ hermano	f	m	m	f	f	f	f	f
hija/ hijo	m	m	f	f	m	f	f	f

Figure 5: results table showing whether given words are, on average across the 10 datasets, closer to the male (m) or female (f) version of the word pairings. "Llorar" and "gana" display a trend of bias towards females and "poder" and "exito" display a trend of bias towards male.

HUMAN TRANSLATION	power/ to be able to	marriage	strong	to cry	success	to work	he/ she decided	he/ she wins
mr/mrs	m	m	f	f	f	f	f	f
woman/ man	f	f	m	f	m	f	f	f
father/ mother	m	m	m	f	m	m	m	f
sister/ brother	m	f	f	f	m	m	m	f
daughter/ son	m	m	m	f	m	m	m	f
MACHINE TRANSLATION	power/ to be able to	marriage	strong	to cry	success	to work	he/ she decided	he/ she wins
mr/ mrs	m	m	f	f	m	f	f	f
woman/ man	m	f	m	f	m	m	f	f
father/ mother	m	f	f	f	m	f	m	f
sister/ brother	f	m	m	f	f	f	f	f
daughter/ son	m	m	f	f	m	f	f	f

Figure 6: same results table as in Fig.5 but translated into English, showing whether given words are, on average across the 10 datasets, closer to the male (m) or female (f) version of the word pairings.

This result is unsurprising, because although it could be due in part to the translation process, it is also likely to have occurred due to unbalanced representation in the initial text. This imbalance has led to biased word embeddings, as found in other papers such as Bolukbasi et al [9].

4.4) Do machine and human translation display different levels of bias?

Next, I calculated the difference between the levels of bias for human and machine translation on the same word pairing, and analysed whether this was statistically significant. Since I know these texts were originally the same (in English), an increased difference between male and female distances in the machine translated version would indicate that the machine translation algorithms are more biased than human translators (Fig. 7)

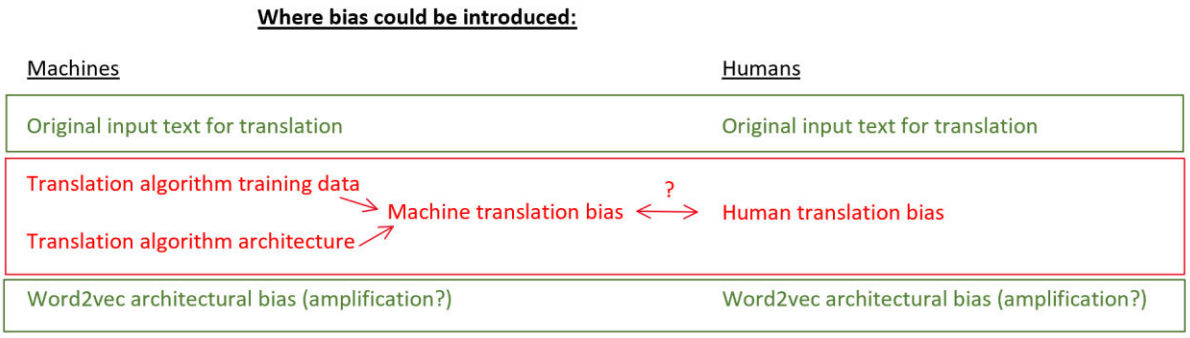

Figure 7: bias can be introduced in various places, but by measuring the difference in bias I am only investigating differences between machine and human translation (in red), since other variables remain constant throughout the experiment (in green).

For given pairings of a target word with a gender pairing I produced a set of results, including the 10 different vector sets, of the difference between the distance to the male and female version from the pairing. The equivalent sets for human and machine translation did not show a significant difference when I performed a t-test ($p > 0.05$ in all cases) for "llorar" and "poder". Also, it varies between humans and machines being more biased. This could indicate that machine translation is no more biased than human translation. For example, Fig. 8 displays results from comparing human and machine biases for "llorar".

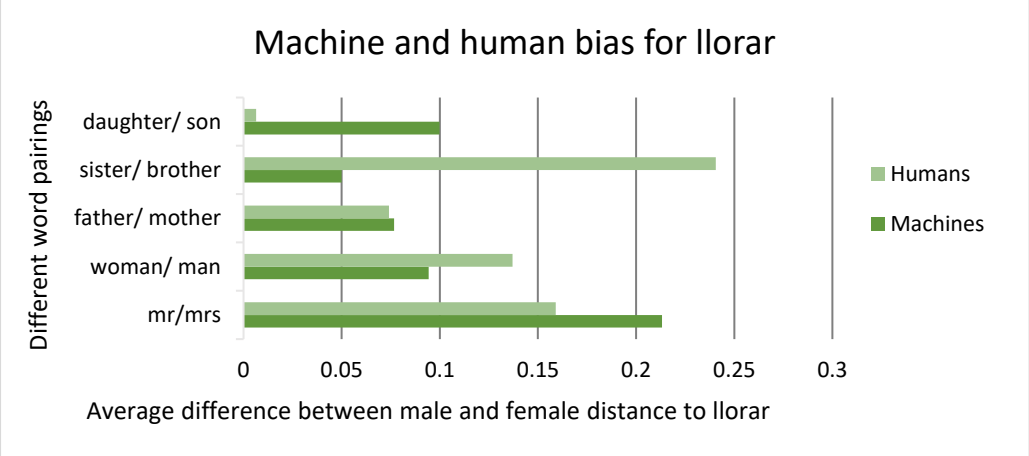

Figure 8: comparing magnitude of female bias for human and machine translation for "llorar" (to cry) finds no clear trend when different gender pairings are used. Each pair of bars is the human and machine translation result for a given gender pairing.

However, there is also significant variation within multiple runs of the same dataset (machine or human). Therefore, a more likely conclusion is that I had insufficient training data to get consistent vector geometry. This means I was unable to draw a reliable comparison between the datasets in this case. However, I propose that this methodology could be utilised on a larger scale, which may reveal more subtle distinctions.

5 Limitations

The main limitation to this research is the requirement for a larger corpus of data. Additionally, the target words set could have perhaps been larger and better selected, something a larger corpus of data would make easier.

Alongside these challenges, another limitation to this approach the difference in publishing time between human and machine translation. Gender attitudes and societal biases can shift over time and the human translations used were carried out in the past, whereas the machine translations were carried out at the time of research. It would be interesting to explore whether more recent human translation data exhibits less bias than historical translation via this method.

The bias metrics are also limited and could be further developed. Primarily, they would be improved by the addition of more words to the challenge set. In this case this was unnecessary as I would have remained limited by the dataset size. Caution must be taken in attempts to correct bias since these metrics do not holistically reflect the entire bias of the dataset. As shown by Gohen et Goldberg [29], if bias reduction techniques simply focus on reducing the measured bias metrics then they will only be surface level corrections.

This work would also benefit from extension into other languages. I used Spanish since I have some knowledge of the language and the use of the roman alphabet alongside the ability to remove accents meant there was accessible software for me to use. Chen et al. [30] have highlighted how the NLP pipeline is much less developed in less widely spoken languages. It is therefore important to extend research to include these languages otherwise we may produce unfair systems which exhibit increased levels of bias towards speakers of less common languages. Within the context of this research this is especially important due to the variation of gender structure between languages. This means mitigations techniques may not be able to be used across languages so perhaps machine translation is already more biased in less common languages. This work could be extended into genderless languages, or languages with a third gender such as neuter in German.

Finally, this methodology is not limited to use in the context of gender bias. If different evaluation metrics were developed, this could also be used to investigate other types of bias such as racial bias. It may also act as a framework for exploring other ideas, such as how style and tone is affected by translation.

6 Conclusion

Overall, this research explores the ongoing challenge of gender bias in machine translation. Although bias fundamentally derives from training data, I demonstrate that sentence level translation leads to lack of context, which can introduce gender bias in translation. This shows translation architecture which carries information forward between sentences is fundamental in reducing gender bias as well as improving accuracy. I develop a framework that, despite providing inconclusive results due to a limited dataset size, may in future be used to evaluate gender bias. I test it on a small scale for English to Spanish translation, and find it to display some bias. Although my quantitative comparison results are not statistically significant, I suggest ways in which it could be extended in the future that may lead to improved results.

7 Appendix

The books in my dataset from InfoBooks, which I translated using Google Translate, are as follows:

- The Strange Case of Dr Jekyll and Mr Hyde, Robert Louis Stevenson
- Wuthering Heights, Emily Bronte
- Pride and Prejudice, Jane Austen

- Frankenstein, Mary Shelley
- The Great Gatsby, F Scott Fitzgerald
- The Scarlett Letter, Nathaniel Hawthorne
- The Portrait of Dorian Grey, Oscar Wilde
- The House on the Borderland, William Hope Hodgson
- At the Mountains of Madness, H P Lovecraft
- The War of the Worlds, H G Wells

Their Spanish translations from InfoLibros are:

- El Extraño Caso del Doctor Jekyll y el Senor Hyde, Robert Louis Stevenson
- Cumbres Borrascosas, Emily Bronte
- Orgullo y Prejuicio, Jane Austen
- Frankenstein, Mary Shelley
- El Gran Gatsby, F Scott Fitzgerald
- La Letra Escarlata, Nathaniel Hawthorne
- El Retrato de Dorian Grey, Oscar Wilde
- La Casa en el Confín de la Tierra, William Hope Hodgson
- En las Montañas de la Locura, H P Lovecraft
- La Guerra de los Mundos, H G Wells

8 References

[1] Open AI, 'Introducing ChatGPT', Nov. 30, 2022. https://openai.com/blog/chatgpt (accessed Mar. 31, 2023).

[2] 'Gender Bias in Machine Translation | Transactions of the Association for Computational Linguistics | MIT Press'. https://direct.mit.edu/tacl/article/doi/10.1162/tacl_a_00401/106991/Gender-Bias-in-Machine-Translation (accessed Mar. 31, 2023).

[3] E. Cambria and B. White, 'Jumping NLP Curves: A Review of Natural Language Processing Research [Review Article]', *IEEE Computational Intelligence Magazine*, vol. 9, no. 2, pp. 48–57, May 2014, doi: 10.1109/MCI.2014.2307227.

[4] W. Yin, K. Kann, M. Yu, and H. Schütze, 'Comparative Study of CNN and RNN for Natural Language Processing'. arXiv, Feb. 07, 2017. Accessed: May 12, 2023. [Online]. Available: http://arxiv.org/abs/1702.01923

[5] A. Vaswani *et al.*, 'Attention is All you Need', in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2017. Accessed: May 12, 2023. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

[6] Microsoft, 'Machine translation'. https://www.microsoft.com/en-us/translator/business/machine-translation/ (accessed Apr. 01, 2023).

[7] D. Stahlberg, F. Braun, L. Irmen, and S. Sczesny, 'Representation of the Sexes in Language', in *Social communication*, Pyschology Press, 2007, pp. 163–166. Accessed: Mar. 31, 2023. [Online]. Available: https://books.google.co.uk/books?hl=en&lr=&id=zzW5dqN8NiUC&oi=fnd&pg=PA163&ots=fzTfarFF6l&sig=y3H21aBpQlV_Te1Au3wM7lm1oSA&redir_esc=y#v=onepage&q&f=false

[8] Real Academia Española, 'Elle', *Diccionario de la lengua Española*. Accessed: Apr. 18, 2023. [Online]. Available: https://dle.rae.es/elle?m=form

[9] T. Bolukbasi, K.-W. Chang, J. Y. Zou, V. Saligrama, and A. T. Kalai, 'Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings', in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2016. Accessed: Dec. 05, 2022. [Online]. Available:

https://proceedings.neurips.cc/paper/2016/hash/a486cd07e4ac3d270571622f4f316ec5-Abstract.html
[10] P. Zhou *et al.*, 'Examining Gender Bias in Languages with Grammatical Gender'. arXiv, Sep. 09, 2019. doi: 10.48550/arXiv.1909.02224.
[11] J. Zhao, T. Wang, M. Yatskar, V. Ordonez, and K.-W. Chang, 'Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints'. arXiv, Jul. 28, 2017. doi: 10.48550/arXiv.1707.09457.
[12] Y. Qian, U. Muaz, B. Zhang, and J. W. Hyun, 'Reducing Gender Bias in Word-Level Language Models with a Gender-Equalizing Loss Function'. arXiv, Jun. 03, 2019. doi: 10.48550/arXiv.1905.12801.
[13] M. R. Costa-jussà, C. Escolano, C. Basta, J. Ferrando, R. Batlle, and K. Kharitonova, 'Gender Bias in Multilingual Neural Machine Translation: The Architecture Matters'. arXiv, Dec. 24, 2020. doi: 10.48550/arXiv.2012.13176.
[14] D. Saunders, R. Sallis, and B. Byrne, 'Neural Machine Translation Doesn't Translate Gender Coreference Right Unless You Make It'. arXiv, Dec. 10, 2020. doi: 10.48550/arXiv.2010.05332.
[15] 'Google translate'. Google. Accessed: Apr. 01, 2023. [Online]. Available: https://translate.google.com/
[16] M. Johnson, 'A Scalable Approach to Reducing Gender Bias in Google Translate', Apr. 22, 2020. A Scalable Approach to Reducing Gender Bias in Google Translate – Google AI Blog (googleblog.com) (accessed Mar. 31, 2023).
[17] I. Caswell and B. Liang, 'Recent advances in Google Translate', Jun. 08, 2020. https://ai.googleblog.com/2020/06/recent-advances-in-google-translate.html (accessed May 08, 2023).
[18] Y. Wu *et al.*, 'Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation'. arXiv, Oct. 08, 2016. doi: 10.48550/arXiv.1609.08144.
[19] L. Wang, Z. Tu, A. Way, and Q. Liu, 'Exploiting Cross-Sentence Context for Neural Machine Translation'. arXiv, Jul. 23, 2017. doi: 10.48550/arXiv.1704.04347.
[20] L. Miculicich, D. Ram, N. Pappas, and J. Henderson, 'Document-Level Neural Machine Translation with Hierarchical Attention Networks'. arXiv, Oct. 01, 2018. doi: 10.48550/arXiv.1809.01576.
[21] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, 'Bleu: a Method for Automatic Evaluation of Machine Translation', in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, Jul. 2002, pp. 311–318. doi: 10.3115/1073083.1073135.
[22] G. Stanovsky, N. A. Smith, and L. Zettlemoyer, 'Evaluating Gender Bias in Machine Translation'. arXiv, Jun. 03, 2019. doi: 10.48550/arXiv.1906.00591.
[23] T. Mikolov, K. Chen, G. Corrado, and J. Dean, 'Efficient Estimation of Word Representations in Vector Space'. arXiv, Sep. 06, 2013. Accessed: Apr. 18, 2023. [Online]. Available: http://arxiv.org/abs/1301.3781
[24] 'Info Books'. https://www.infobooks.org/ (accessed May 05, 2023).
[25] 'Info libros'. https://infolibros.org/ (accessed May 05, 2023).
[26] The TensorFlow Authors, 'word2vec'. Accessed: Apr. 18, 2023. [Online]. Available: https://github.com/tensorflow/docs/blob/master/site/en/tutorials/text/word2vec.ipynb
[27] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, 'Distributed Representations of Words and Phrases and their Compositionality', in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2013. Accessed: Apr. 18, 2023. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2013/hash/9aa42b31882ec039965f3c4923ce901b-Abstract.html
[28] 'TensorFlow embedding projector'. [Online]. Available: https://projector.tensorflow.org/
[29] H. Gonen and Y. Goldberg, 'Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them'. arXiv, Sep. 24, 2019. doi: 10.48550/arXiv.1903.03862.
[30] Y. Chen *et al.*, 'Gender Bias and Under-Representation in Natural Language Processing Across Human Languages', in *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and*

Society, in AIES ’21. New York, NY, USA: Association for Computing Machinery, Jul. 2021, pp. 24–34. doi: 10.1145/3461702.3462530.