Assessing Gender Bias in Machine Translation -- A Case Study with Google Translate

Abstract

Recently there has been a growing concern about machine bias, where trained statistical models grow to reflect controversial societal asymmetries, such as gender or racial bias. A significant number of AI tools have recently been suggested to be harmfully biased towards some minority, with reports of racist criminal behavior predictors, Iphone X failing to differentiate between two Asian people and Google photos' mistakenly classifying black people as gorillas. Although a systematic study of such biases can be difficult, we believe that automated translation tools can be exploited through gender neutral languages to yield a window into the phenomenon of gender bias in AI. In this paper, we start with a comprehensive list of job positions from the U.S. Bureau of Labor Statistics (BLS) and used it to build sentences in constructions like"He/She is an Engineer"in 12 different gender neutral languages such as Hungarian, Chinese, Yoruba, and several others. We translate these sentences into English using the Google Translate API, and collect statistics about the frequency of female, male and gender-neutral pronouns in the translated output. We show that GT exhibits a strong tendency towards male defaults, in particular for fields linked to unbalanced gender distribution such as STEM jobs. We ran these statistics against BLS' data for the frequency of female participation in each job position, showing that GT fails to reproduce a real-world distribution of female workers. We provide experimental evidence that even if one does not expect in principle a 50:50 pronominal gender distribution, GT yields male defaults much more frequently than what would be expected from demographic data alone. We are hopeful that this work will ignite a debate about the need to augment current statistical translation tools with debiasing techniques which can already be found in the scientific literature.

Introduction

Although the idea of automated translation can in principle be traced back to as long as the 17th century with René Descartes proposal of an “universal language” BIBREF0 , machine translation has only existed as a technological field since the 1950s, with a pioneering memorandum by Warren Weaver BIBREF1 , BIBREF2 discussing the possibility of employing digital computers to perform automated translation. The now famous Georgetown-IBM experiment followed not long after, providing the first experimental demonstration of the prospects of automating translation by the means of successfully converting more than sixty Russian sentences into English BIBREF3 . Early systems improved upon the results of the Georgetown-IBM experiment by exploiting Noam Chomsky's theory of generative linguistics, and the field experienced a sense of optimism about the prospects of fully automating natural language translation. As is customary with artificial intelligence, the initial optimistic stage was followed by an extended period of strong disillusionment with the field, of which the catalyst was the influential 1966 ALPAC (Automatic Language Processing Advisory Committee) report(BIBREF4 . Such research was then disfavoured in the United States, making a re-entrance in the 1970s before the 1980s surge in statistical methods for machine translation BIBREF5 , BIBREF6 . Statistical and example-based machine translation have been on the rise ever since BIBREF7 , BIBREF8 , BIBREF9 , with highly successful applications such as Google Translate (recently ported to a neural translation technology BIBREF10) amounting to over 200 million users daily.

In spite of the recent commercial success of automated translation tools (or perhaps stemming directly from it), machine translation has amounted a significant deal of criticism. Noted philosopher and founding father of generative linguistics Noam Chomsky has argued that the achievements of machine translation, while successes in a particular sense, are not successes in the sense that science has ever been interested in: they merely provide effective ways, according to Chomsky, of approximating unanalyzed data BIBREF11 , BIBREF12 . Chomsky argues that the faith of the MT community in statistical methods is absurd by analogy with a standard scientific field such as physics BIBREF11 :

I mean actually you could do physics this way, instead of studying things like balls rolling down frictionless planes, which can't happen in nature, if you took a ton of video tapes of what's happening outside my office window, let's say, you know, leaves flying and various things, and you did an extensive analysis of them, you would get some kind of prediction of what's likely to happen next, certainly way better than anybody in the physics department could do. Well that's a notion of success which is I think novel, I don't know of anything like it in the history of science.

Leading AI researcher and Google's Director of Research Peter Norvig responds to these arguments by suggesting that even standard physical theories such as the Newtonian model of gravitation are, in a sense, trained BIBREF12 :

As another example, consider the Newtonian model of gravitational attraction, which says that the force between two objects of mass INLINEFORM0 and INLINEFORM1 a distance INLINEFORM2 apart is given by INLINEFORM3

where INLINEFORM0 is the universal gravitational constant. This is a trained model because the gravitational constant G is determined by statistical inference over the results of a series of experiments that contain stochastic experimental error. It is also a deterministic (non-probabilistic) model because it states an exact functional relationship. I believe that Chomsky has no objection to this kind of statistical model. Rather, he seems to reserve his criticism for statistical models like Shannon's that have quadrillions of parameters, not just one or two.

Chomsky and Norvig's debate BIBREF12 is a microcosm of the two leading standpoints about the future of science in the face of increasingly sophisticated statistical models. Are we, as Chomsky seems to argue, jeopardizing science by relying on statistical tools to perform predictions instead of perfecting traditional science models, or are these tools, as Norvig argues, components of the scientific standard

since its conception? Currently there are no satisfactory resolutions to this conundrum, but perhaps statistical models pose an even greater and more urgent threat to our society.

On a 2014 article, Londa Schiebinger suggested that scientific research fails to take gender issues into account, arguing that the phenomenon of male defaults on new technologies such as Google Translate provides a window into this asymmetry BIBREF13 . Since then, recent worrisome results in machine learning have somewhat supported Schiebinger's view. Not only Google photos' statistical image labeling algorithm has been found to classify dark-skinned people as gorillas BIBREF14 and purportedly intelligent programs have been suggested to be negatively biased against black prisoners when predicting criminal behavior BIBREF15 but the machine learning revolution has also indirectly revived heated debates about the controversial field of physiognomy, with proposals of AI systems capable of identifying the sexual orientation of an individual through its facial characteristics BIBREF16 . Similar concerns are growing at an unprecedented rate in the media, with reports of Apple's Iphone X face unlock feature failing to differentiate between two different Asian people BIBREF17 and automatic soap dispensers which reportedly do not recognize black hands BIBREF18 . Machine bias, the phenomenon by which trained statistical models unbeknownst to their creators grow to reflect controversial societal asymmetries, is growing into a pressing concern for the modern times, invites us to ask ourselves whether there are limits to our dependence on these techniques – and more importantly, whether some of these limits have already been traversed. In the wave of algorithmic bias, some have argued for the creation of some kind of agency in the likes of the Food and Drug Administration, with the sole purpose of regulating algorithmic discrimination BIBREF19 .

With this in mind, we propose a quantitative analysis of the phenomenon of gender bias in machine translation. We illustrate how this can be done by simply exploiting Google Translate to map sentences from a gender neutral language into English. As Figure FIGREF1 exemplifies, this approach produces results consistent with the hypothesis that sentences about stereotypical gender roles are translated

accordingly with high probability: nurse and baker are translated with female pronouns while engineer and CEO are translated with male ones.

Motivation

As of 2018, Google Translate is one of the largest publicly available machine translation tools in existence, amounting 200 million users daily BIBREF21 . Initially relying on United Nations and European Parliament transcripts to gather data, since 2014 Google Translate has inputed content from its users through the Translate Community initiative BIBREF22 . Recently however there has been a growing concern about gender asymmetries in the translation mechanism, with some heralding it as “sexist” BIBREF23 . This concern has to at least some extent a scientific backup: A recent study has shown that word embeddings are particularly prone to yielding gender stereotypes BIBREF24 . Fortunately, the researchers propose a relatively simple debiasing algorithm with promising results: they were able to cut the proportion of stereotypical analogies from INLINEFORM0 to INLINEFORM1 without any significant compromise in the performance of the word embedding technique. They are not alone: there is a growing effort to systematically discover and resolve issues of algorithmic bias in black-box algorithms BIBREF25 . The success of these results suggest that a similar technique could be used to remove gender bias from Google Translate outputs, should it exist. This paper intends to investigate whether it does. We are optimistic that our research endeavors can be used to argue that there is a positive payoff in redesigning modern statistical translation tools.

Assumptions and Preliminaries

In this paper we assume that a statistical translation tool should reflect at most the inequality existent in society – it is only logical that a translation tool will poll from examples that society produced and, as such, will inevitably retain some of that bias. It has been argued that one's language affects one's

knowledge and cognition about the world BIBREF26 , and this leads to the discussion that languages that distinguish between female and male genders grammatically may enforce a bias in the person's perception of the world, with some studies corroborating this, as shown in BIBREF27 , as well some relating this with sexism BIBREF28 and gender inequalities BIBREF29 .

With this in mind, one can argue that a move towards gender neutrality in language and communication should be striven as a means to promote improved gender equality. Thus, in languages where gender neutrality can be achieved – such as English – it would be a valid aim to create translation tools that keep the gender-neutrality of texts translated into such a language, instead of defaulting to male or female variants.

We will thus assume throughout this paper that although the distribution of translated gender pronouns may deviate from 50:50, it should not deviate to the extent of misrepresenting the demographics of job positions. That is to say we shall assume that Google Translate incorporates a negative gender bias if the frequency of male defaults overestimates the (possibly unequal) distribution of male employees per female employee in a given occupation.

Materials and Methods

We shall assume and then show that the phenomenon of gender bias in machine translation can be assessed by mapping sentences constructed in gender neutral languages to English by the means of an automated translation tool. Specifically, we can translate sentences such as the Hungarian “ő egy ápolónő”, where “ápolónő” translates to “nurse” and “ő” is a gender-neutral pronoun meaning either he, she or it, to English, yielding in this example the result “she's a nurse” on Google Translate. As Figure FIGREF1 clearly shows, the same template yields a male pronoun when “nurse” is replaced by “engineer”. The same basic template can be ported to all other gender neutral languages, as depicted in

Table TABREF4 . Given the success of Google Translate, which amounts to 200 million users daily, we have chosen to exploit its API to obtain the desired thermometer of gender bias. Also, in order to solidify our results, we have decided to work with a fair amount of gender neutral languages, forming a list of these with help from the World Atlas of Language Structures (WALS) BIBREF30 and other sources. Table TABREF2 compiles all languages we chose to use, with additional columns informing whether they (1) exhibit a gender markers in the sentence and (2) are supported by Google Translate. However, we stumbled on some difficulties which led to some of those langauges being removed, which will be explained in . There is a prohibitively large class of nouns and adjectives that could in principle be substituted into our templates. To simplify our dataset, we have decided to focus our work on job positions – which, we believe, are an interesting window into the nature of gender bias –, and were able to obtain a comprehensive list of professional occupations from the Bureau of Labor Statistics' detailed occupations table BIBREF31 , from the United States Department of Labor. The values inside, however, had to be expanded since each line contained multiple occupations and sometimes very specific ones. Fortunately this table also provided a percentage of women participation in the jobs shown, for those that had more than 50 thousand workers. We filtered some of these because they were too generic (“Computer occupations, all other”, and others) or because they had gender specific words for the profession (“host/hostess”, “waiter/waitress”). We then separated the curated jobs into broader categories (Artistic, Corporate, Theatre, etc.) as shown in Table TABREF3 . Finally, Table TABREF5 shows thirty examples of randomly selected occupations from our dataset. For the occupations that had less than 50 thousand workers, and thus no data about the participation of women, we assumed that its women participation was that of its upper category. Finally, as complementary evidence we have decided to include a small subset of 21 adjectives in our study. All adjectives were obtained from the top one thousand most frequent words in this category as featured in the Corpus of Contemporary American English (COCA) https://corpus.byu.edu/coca/, but it was necessary to manually curate them because a substantial fraction of these adjectives cannot be applied to human subjects. Also because the sentiment associated with each adjective is not as easily accessible as for example the occupation category of each

job position, we performed a manual selection of a subset of such words which we believe to be meaningful to this study. These words are presented in Table TABREF6 . We made all code and data used to generate and compile the results presented in the following sections publicly available in the following Github repository: https://github.com/marceloprates/Gender-Bias. Note however that because the Google Translate algorithm can change, unfortunately we cannot guarantee full reproducibility of our results. All experiments reported here were conducted on April 2018.

Rationale for language exceptions

While it is possible to construct gender neutral sentences in two of the languages omitted in our experiments (namely Korean and Nepali), we have chosen to omit them for the following reasons:

We faced technical difficulties to form templates and automatically translate sentences with the right-to-left, top-to-bottom nature of the script and, as such, we have decided not to include it in our experiments.

Due to Nepali having a rather complex grammar, with possible male/female gender demarcations on the phrases and due to none of the authors being fluent or able to reach someone fluent in the language, we were not confident enough in our ability to produce the required templates. Bengali was almost discarded under the same rationale, but we have decided to keep it because of our sentence template for Bengali has a simple grammatical structure which does not require any kind of inflection.

One can construct gender neutral phrases in Korean by omitting the gender pronoun; in fact, this is the default procedure. However, the expressiveness of this omission depends on the context of the sentence being clear, which is not possible in the way we frame phrases.

Distribution of translated gender pronouns per occupation category

A sensible way to group translation data is to coalesce occupations in the same category and collect statistics among languages about how prominent male defaults are in each field. What we have found is that Google Translate does indeed translate sentences with male pronouns with greater probability than it does either with female or gender-neutral pronouns, in general. Furthermore, this bias is seemingly aggravated for fields suggested to be troubled by male stereotypes, such as life and physical sciences, architecture, engineering, computer science and mathematics BIBREF32 . Table TABREF11 summarizes these data, and Table TABREF12 summarizes it even further by coalescing occupation categories into broader groups to ease interpretation. For instance, STEM (Science, Technology, Engineering and Mathematics) fields are grouped into a single category, which helps us compare the large asymmetry between gender pronouns in these fields (INLINEFORM0 of male defaults) to that of more evenly distributed fields such as healthcare (INLINEFORM1).

Plotting histograms for the number of gender pronouns per occupation category sheds further light on how female, male and gender-neutral pronouns are differently distributed. The histogram in Figure FIGREF13 suggests that the number of female pronouns is inversely distributed – which is mirrored in the data for gender-neutral pronouns in Figure FIGREF15 –, while the same data for male pronouns (shown in Figure FIGREF14) suggests a skew normal distribution. Furthermore we can see both on Figures FIGREF13 and FIGREF14 how STEM fields (labeled in beigeexhibit predominantly male defaults – amounting predominantly near INLINEFORM0 in the female histogram although much to the right in the male histogram.

These values contrast with BLS' report of gender participation, which will be discussed in more detail in Section SECREF8 .

We can also visualize male, female, and gender neutral histograms side by side, in which context is useful to compare the dissimilar distributions of translated STEM and Healthcare occupations (Figures FIGREF16 and FIGREF17 respectively). The number of translated female pronouns among languages is not normally distributed for any of the individual categories in Table TABREF3 , but Healthcare is in many ways the most balanced category, which can be seen in comparison with STEM – in which male defaults are second to most prominent.

The bar plots in Figure FIGREF18 help us visualize how much of the distribution of each occupation category is composed of female, male and gender-neutral pronouns. In this context, STEM fields, which show a predominance of male defaults, are contrasted with Healthcare and educations, which show a larger proportion of female pronouns.

Although computing our statistics over the set of all languages has practical value, this may erase subtleties characteristic to each individual idiom. In this context, it is also important to visualize how each language translates job occupations in each category. The heatmaps in Figures FIGREF19 , FIGREF20 and FIGREF21 show the translation probabilities into female, male and neutral pronouns, respectively, for each pair of language and category (blue is INLINEFORM0 and red is INLINEFORM1). Both axes are sorted in these Figures, which helps us visualize both languages and categories in an spectrum of increasing male/female/neutral translation tendencies. In agreement with suggested stereotypes, BIBREF32 STEM fields are second only to Legal ones in the prominence of male defaults. These two are followed by Arts & Entertainment and Corporate, in this order, while Healthcare, Production and Education lie on the opposite end of the spectrum.

Our analysis is not truly complete without tests for statistical significant differences in the translation tendencies among female, male and gender neutral pronouns. We want to know for which languages and categories does Google Translate translate sentences with significantly more male than female, or male

than neutral, or neutral than female, pronouns. We ran one-sided t-tests to assess this question for each pair of language and category and also totaled among either languages or categories. The corresponding p-values are presented in Tables TABREF22 , TABREF23 , TABREF24 respectively. Language-Category pairs for which the null hypothesis was not rejected for a confidence level of INLINEFORM0 are highlighted in blue. It is important to note that when the null hypothesis is accepted, we cannot discard the possibility of the complementary null hypothesis being rejected. For example, neither male nor female pronouns are significantly more common for Healthcare positions in the Estonian language, but female pronouns are significantly more common for the same category in Finnish and Hungarian. Because of this, Language-Category pairs for which the complementary null hypothesis is rejected are painted in a darker shade of blue (see Table TABREF22 for the three examples cited above.

Although there is a noticeable level of variation among languages and categories, the null hypothesis that male pronouns are not significantly more frequent than female ones was consistently rejected for all languages and all categories examined. The same is true for the null hypothesis that male pronouns are not significantly more frequent than gender neutral pronouns, with the one exception of the Basque language (which exhibits a rather strong tendency towards neutral pronouns). The null hypothesis that neutral pronouns are not significantly more frequent than female ones is accepted with much more frequency, namely for the languages Malay, Estonian, Finnish, Hungarian, Armenian and for the categories Farming & Fishing & Forestry, Healthcare, Legal, Arts & Entertainment, Education. In all three cases, the null hypothesis corresponding to the aggregate for all languages and categories is rejected. We can learn from this, in summary, that Google Translate translates male pronouns more frequently than both female and gender neutral ones, either in general for Language-Category pairs or consistently among languages and among categories (with the notable exception of the Basque idiom).

Distribution of translated gender pronouns per language

We have taken the care of experimenting with a fair amount of different gender neutral languages. Because of that, another sensible way of coalescing our data is by language groups, as shown in Table TABREF25 . This can help us visualize the effect of different cultures in the genesis – or lack thereof – of gender bias. Nevertheless, the barplots in Figure FIGREF26 are perhaps most useful to identifying the difficulty of extracting a gender pronoun when translating from certain languages. Basque is a good example of this difficulty, although the quality of Bengali, Yoruba, Chinese and Turkish translations are also compromised.

Distribution of translated gender pronouns for varied adjectives

We queried the 1000 most frequently used adjectives in English, as classified in the COCA corpus [https://corpus.byu.edu/coca/], but since not all of them were readily applicable to the sentence template we used, we filtered the N adjectives that would fit the templates and made sense for describing a human being. The list of adjectives extracted from the corpus is available on the Github repository: https://github.com/marceloprates/Gender-Bias.

Apart from occupations, which we have exhaustively examined by collecting labor data from the U.S. Bureau of Labor Statistics, we have also selected a small subset of adjectives from the Corpus of Contemporary American English (COCA) https://corpus.byu.edu/coca/, in an attempt to provide preliminary evidence that the phenomenon of gender bias may extend beyond the professional context examined in this paper. Because a large number of adjectives are not applicable to human subjects, we manually curated a reasonable subset of such words. The template used for adjectives is similar to that used for occupations, and is provided again for reference in Table TABREF4 .

Once again the data points towards male defaults, but some variation can be observed throughout different adjectives. Sentences containing the words Shy, Attractive, Happy, Kind and Ashamed are

predominantly female translated (Attractive is translated as female and gender-neutral in equal parts), while Arrogant, Cruel and Guilty are disproportionately translated with male pronouns (Guilty is in fact never translated with female or neutral pronouns).

Comparison with women participation data across job positions

A sensible objection to the conclusions we draw from our study is that the perceived gender bias in Google Translate results stems from the fact that possibly female participation in some job positions is itself low. We must account for the possibility that the statistics of gender pronouns in Google Translate outputs merely reflects the demographics of male-dominated fields (male-dominated fields can be considered those that have less than 25% of women participation BIBREF20 , according to the U.S. Department of Labor Women's Bureau). In this context, the argument in favor of a critical revision of statistic translation algorithms weakens considerably, and possibly shifts the blame away from these tools.

The U.S. Bureau of Labor Statistics data summarized in Table TABREF3 contains statistics about the percentage of women participation in each occupation category. This data is also available for each individual occupation, which allows us to compute the frequency of women participation for each 12-quantile. We carried the same computation in the context of frequencies of translated female pronouns, and the resulting histograms are plotted side-by-side in Figure FIGREF29 . The data shows us that Google Translate outputs fail to follow the real-world distribution of female workers across a comprehensive set of job positions. The distribution of translated female pronouns is consistently inversely distributed, with female pronouns accumulating in the first 12-quantile. By contrast, BLS data shows that female participation peaks in the fourth 12-quantile and remains significant throughout the next ones.

Averaged over occupations and languages, sentences are translated with female pronouns INLINEFORM0 of the time. In contrast, the gender participation frequency for female workers averaged over all occupations in the BLS report yields a consistently larger figure of INLINEFORM1 . The variance reported for the translation results is also lower, at INLINEFORM2 in contrast with the report's INLINEFORM3 . We ran an one-sided t-test to evaluate the null hypothesis that the female participation frequency is not significantly greater then the GT female pronoun frequency for the same job positions, obtaining a p-value INLINEFORM4 vastly inferior to our confidence level of INLINEFORM5 and thus rejecting H0 and concluding that Google Translate's female translation frequencies sub-estimates female participation frequencies in US job positions. As a result, it is not possible to understand this asymmetry as a reflection of workplace demographics, and the prominence of male defaults in Google Translate is, we believe, yet lacking a clear justification.

Discussion

At the time of the writing up this paper, Google Translate offered only one official translation for each input word, along with a list of synonyms. In this context, all experiments reported here offer an analysis of a “screenshot” of that tool as of August 2018, the moment they were carried out. A preprint version of this paper was posted the in well-known Cornell University-based arXiv.org open repository on September 6, 2018. The manuscript soon enjoyed a significant amount of media coverage, featuring on The Register BIBREF33 , Datanews BIBREF34 , t3n BIBREF35 , among others, and more recently on Slator BIBREF36 and Jornal do Comercio BIBREF37 . On December 6, 2018 the company's policy changed, and a statement was released detailing their efforts to reduce gender bias on Google Translate, which included a new feature presenting the user with a feminine as well as a masculine official translation (Figure FIGREF30). According to the company, this decision is part of a broader goal of promoting fairness and reducing biases in machine learning. They also acknowledged the technical reasons behind gender bias in their model, stating that:

Google Translate learns from hundreds of millions of already-translated examples from the web. Historically, it has provided only one translation for a query, even if the translation could have either a feminine or masculine form. So when the model produced one translation, it inadvertently replicated gender biases that already existed. For example: it would skew masculine for words like “strong” or “doctor,” and feminine for other words, like “nurse” or “beautiful.”

Their statement is very similar to the conclusions drawn on this paper, as is their motivation for redesigning the tool. As authors, we are incredibly happy to see our vision and beliefs align with those of Google in such a short timespan from the initial publishing of our work, although the company's statement does not cite any study or report in particular and thus we cannot know for sure whether this paper had an effect on their decision or not. Regardless of whether their decision was monocratic, guided by public opinion or based on published research, we understand it as an important first step on an ongoing fight against algorithmic bias, and we praise the Google Translate team for their efforts.

Google Translate's new feminine and masculine forms for translated sentences exemplifies how, as this paper also suggests, machine learning translation tools can be debiased, dropping the need for resorting to a balanced training set. However, it should be noted that important as it is, GT's new feature is still a first step. It does not address all of the shortcomings described in this paper, and the limited language coverage means that many users will still experience gender biased translation results. Furthermore, the system does not yet have support for non-binary results, which may exclude part of their user base.

In addition, one should note that further evidence is mounting about the kind of bias examined in this paper: it is becoming clear that this is a statistical phenomenon independent from any proprietary tool. In this context, the research carried out in BIBREF24 presents a very convincing argument for the sensitivity of word embeddings to gender bias in the training dataset. This suggests that machine translation engineers should be especially aware of their training data when designing a system. It is not feasible to

train these models on unbiased texts, as they are probably scarce. What must be done instead is to engineer solutions to remove bias from the system after an initial training, which seems to be the goal of Google Translate's recent efforts. Fortunately, as BIBREF24 also show, debiasing can be implemented with relatively low effort and modest resources. The technology to promote social justice on machine translation in particular and machine learning in general is often already available. The most significant effort which must be taken in this context is to promote social awareness on these issues so that society can be invited into the conversation.

Conclusions

In this paper, we have provided evidence that statistical translation tools such as Google Translate can exhibit gender biases and a strong tendency towards male defaults. Although implicit, these biases possibly stem from the real world data which is used to train them, and in this context possibly provide a window into the way our society talks (and writes) about women in the workplace. In this paper, we suggest that and test the hypothesis that statistical translation tools can be probed to yield insights about stereotypical gender roles in our society – or at least in their training data. By translating professional-related sentences such as “He/She is an engineer” from gender neutral languages such as Hungarian and Chinese into English, we were able to collect statistics about the asymmetry between female and male pronominal genders in the translation outputs. Our results show that male defaults are not only prominent, but exaggerated in fields suggested to be troubled with gender stereotypes, such as STEM (Science, Technology, Engineering and Mathematics) occupations. And because Google Translate typically uses English as a lingua franca to translate between other languages (e.g. Chinese INLINEFORM0 English INLINEFORM1 Portuguese) BIBREF38 , BIBREF39 , our findings possibly extend to translations between gender neutral languages and non-gender neutral languages (apart from English) in general, although we have not tested this hypothesis.

Our results seem to suggest that this phenomenon extends beyond the scope of the workplace, with the proportion of female pronouns varying significantly according to adjectives used to describe a person. Adjectives such as Shy and Desirable are translated with a larger proportion of female pronouns, while Guilty and Cruel are almost exclusively translated with male ones. Different languages also seemingly have a significant impact in machine gender bias, with Hungarian exhibiting a better equilibrium between male and female pronouns than, for instance, Chinese. Some languages such as Yoruba and Basque were found to translate sentences with gender neutral pronouns very often, although this is the exception rather than the rule and Basque also exhibits a high frequency of phrases for which we could not automatically extract a gender pronoun.

In order to strengthen our results, we ran pronominal gender translation statistics against the U.S. Bureau of Labor Statistics data on the frequency of women participation for each job position. Although Google Translate exhibits male defaults, this phenomenon may merely reflect the unequal distribution of male and female workers in some job positions. To test this hypothesis, we compared the distribution of female workers with the frequency of female translations, finding no correlation between said variables. Our data shows that Google Translate outputs fail to reflect the real-world distribution of female workers, under-estimating the expected frequency. That is to say that even if we do not expect a 50:50 distribution of translated gender pronouns, Google Translate exhibits male defaults in a greater frequency that job occupation data alone would suggest. The prominence of male defaults in Google Translate is therefore to the best of our knowledge yet lacking a clear justification.

We think this work sheds new light on a pressing ethical difficulty arising from modern statistical machine translation, and hope that it will lead to discussions about the role of AI engineers on minimizing potential harmful effects of the current concerns about machine bias. We are optimistic that unbiased results can be obtained with relatively little effort and marginal cost to the performance of current methods, to which current debiasing algorithms in the scientific literature are a testament.

Acknowledgments

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001 and the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq).

This is a pre-print of an article published in Neural Computing and Applications.