Assessing Gender Bias in Machine Translation

Marcelo Prates Luis Lamb

December 15, 2017

Abstract

1 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" [1], machine translation has only existed as a technological field since the 1950s, with a pioneering memorandum by Warren Weaver [2] 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 [3]. Early systems improved upon the results of the Georgetown-IBM experiment by exploiting 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 ALPAC report [4]. Research was almost completely abandoned in the United States, making a shy re-entrance in the 1970s before the 1980s surge in statistical methods for machine translation [5]. Statistical and example-based machine translation have been on the rise ever since [6], with highly successful applications such as Google Translate (recently ported to a neural translation technology [7]) amounting to over 200 million users daily [8].

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. Chomsky argues that the faith of the MT community in statistical methods is absurd by analogy with a standard scientific field such as physics:

> 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*:

> As another example, consider the Newtonian model of gravitational attraction, which says that the force between two objects of mass m_1 and m_2 a distance r apart is given by
>
> $$F = Gm_1m_2/r^2$$
>
> where G 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 [9] is a microcosmos 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 [10]. Since then, recent worrisome results in machine learning have somewhat supported Schiebinger's view, and perhaps even partially confirmed some of Chomsky's fears. Not only Google photos' statistical image labeling algorithm has been found to classify dark-skinned people as gorillas [11] and purportedly intelligent programs have been suggested to be negatively biased against black prisoners when predicting criminal behavior [12] 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 [13]. *Machine bias*, the phenomenon by which trained statistical models unbeknownst to their creators become vessels of their own prejudices, is growing into a pressing concern for the modern times, and is an invitation for 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.

With this in mind, we propose a quantitative analysis of the phenomenon of gender bias in machine translation. We believe this can be done by simply exploiting Google Translate to map sentences from a gender neutral language to English, as Figure 1 exemplifies.

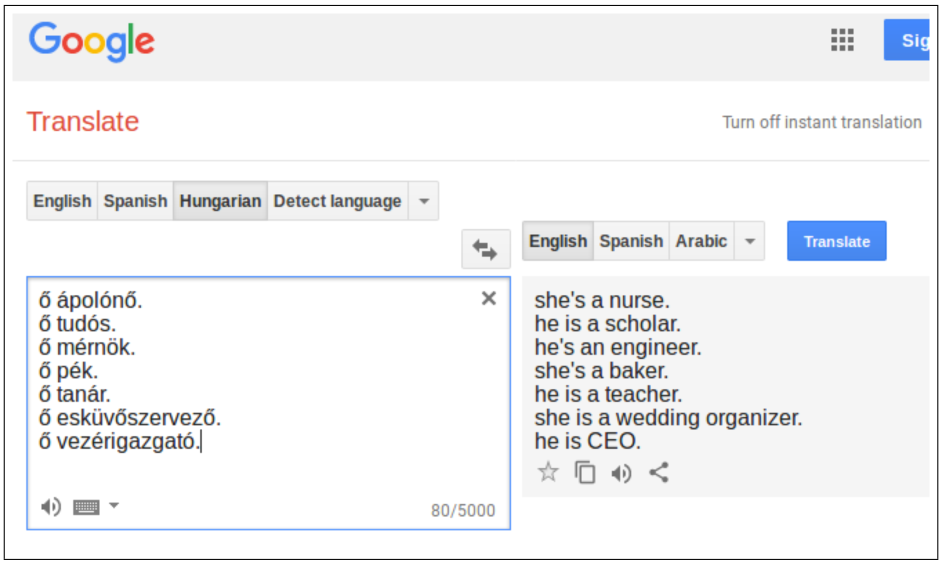

Figure 1: Translating sentences from a gender neutral language such as Hungarian to English provides a glimpse into the phenomenon of gender bias in machine translation. This screenshot from Google Translate shows how occupations from traditionally male-dominated fields such as scholar, engineer and CEO are interpreted as male, while occupations such as nurse, baker and wedding organizer are interpreted as female.

2 Motivation

3 Methods

We believe 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 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" in Google Translate. The same basic template can be ported to all other gender neutral languages, as Table 3 shows. Given the success of Google Translate, which amounts to 200 million users daily, we have decided 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 as many gender neutral languages as possible, obtaining a list of these from the Wikipedia article `https://en.wikipedia.org/wiki/Gender_neutrality_in_genderless_languages`. Table 3 compiles all languages from said article, with additional columns informing whether they 1) exhibit a pronominal gender system and 2) are supported by Google Translate. Because pronominal gender systems defy the purposes of our technique, such languages have been discarded.

There is a prohibitively large class of nouns and adjectives that could in principle be substituted in the templates of Table 3. To simplify our dataset, we have decided to obtain a comprehensive list of professional occupations, which, we believe, are an interesting window into the nature of gender bias. Once again we resorted to a Wikipedia article (`https://en.wikipedia.org/wiki/Lists_of_occupations`) to collect this data, of which the statistics of occupations per category (Artistic, Corporate, Theatre, etc.) is shown in Table 3. Finally, Table 3 shows thirty examples of randomly selected occupations from our dataset.

Language Family	Language	Pronominal Gender System	Supported
Austronesian	Malay	No	Yes
	Tagalog	No	**No**
Finno-Ugric	Estonian	No	Yes
	Finnish	No	Yes
	Hungarian	No	Yes
Indo-European	Armenian	No	Yes
	Bengali	No	Yes
	English	**Yes**	Yes
	Persian	**Yes**	Yes
Indo-Aryan	Maithili	No	**No**
	Nepali	No	Yes
	Oriya	No	**No**
	Japanese	No	Yes
	Korean	**Optional**	Yes
	Turkish	No	Yes
	Yoruba	No	Yes
	Basque	No	Yes
	Swahili	No	Yes
	Chinese	**Optional**	Yes
	Cantonese	No	**No**
	Pipil	No	**No**
	Quechuan	No	**No**
Constructed	Esperanto	**Yes**	Yes
	Ido	**Optional**	**No**
	Lingua Franca Nova	No	**No**
	Interlingua	No	**No**

Table 1: Selected gender neutral languages obtained from the Wikipedia article `https://en.wikipedia.org/wiki/Gender_neutrality_in_genderless_languages`. Languages are grouped according to language families and classified according to whether they exhibit pronominal gender system. For the purposes of this work, we have decided to work only with languages lacking such a system, and as such Persian and Esperanto have been discarded. Because Tagalog, Maithili, Oriya, Cantonese, Pipil, Quechuan, Lingua Franca Nova and Interlingua lack support from Google Translate, these languages have also been ommitted from this work.

Category	# Occupations
Artistic	102
Computer	19
Corporate	50
Dance	9
Film/Television	26
Healthcare	88
Industrial	26
Science	50
Service	10
Theatre	52
Writing	29
Total	436

Table 2: Selected occupations obtained from the Wikipedia article `https://en.wikipedia.org/wiki/Lists_of_occupations`, grouped by category. We have selected a total of 453 occupations from 11 distinct groups (Artistic, Science, Service, etc.).

Language	Sentence template
Malay	dia adalah $\langle$*occupation*$\rangle$
Estonian	ta on $\langle$*occupation*$\rangle$
Finnish	hän on $\langle$*occupation*$\rangle$
Hungarian	ő $\langle$*occupation*$\rangle$
Armenian	$\langle$*occupation*$\rangle$
Japanese	$\langle$*occupation*$\rangle$ です
Turkish	o bir $\langle$*occupation*$\rangle$
Yoruba	o jẹ $\langle$*occupation*$\rangle$
Basque	$\langle$*occupation*$\rangle$ da
Swahili	yeye ni $\langle$*occupation*$\rangle$
Chinese	ta $\langle$*occupation*$\rangle$

Table 3: Templates used to infer gender biases in the translation to the English language.

stagehands	author	neurologist
screenwriter	animator	marketing director
biochemist	endocrinologist	freelancer
neurosurgeon	computer scientist	petrochemical engineer
food stylist	cardiothoracic surgeon	property master
literary editor	video editor	animation director
house manager	chief administrative officer	arts administration
actor	dialysis technician	family nurse practitioner
psychologist	chief creative officer	flash developer
scenic artist	producer	medical laboratory scientist

Table 4: A randomly selected example subset of thirty occupations obtained from our dataset with a total of 436 different occupations.

4 Results

Category	Female	Male	Neutral	Ratio	Total
Artistic	179	504	206	2.816	918
Computer	7	125	39	17.857	171
Corporate	36	340	74	9.444	450
Dance	31	33	17	1.064	81
Film-television	54	125	55	2.315	234
Healthcare	174	475	130	2.730	792
Science	34	357	59	10.500	450
Service	5	63	22	12.600	90
Theatre	75	296	106	3.947	477
Writing	41	148	72	3.610	261
Total	636	2466	780	3.877	

Table 5: Number of female, male and neutral pronominal genders per occupation category in the translated sentences. The corresponding sex ratios (# Male / # Female) show just how much male defaults are prominent in male dominated fields such as computer science, with up to ≈ 18 occurrences of male pronouns for each of a female one.

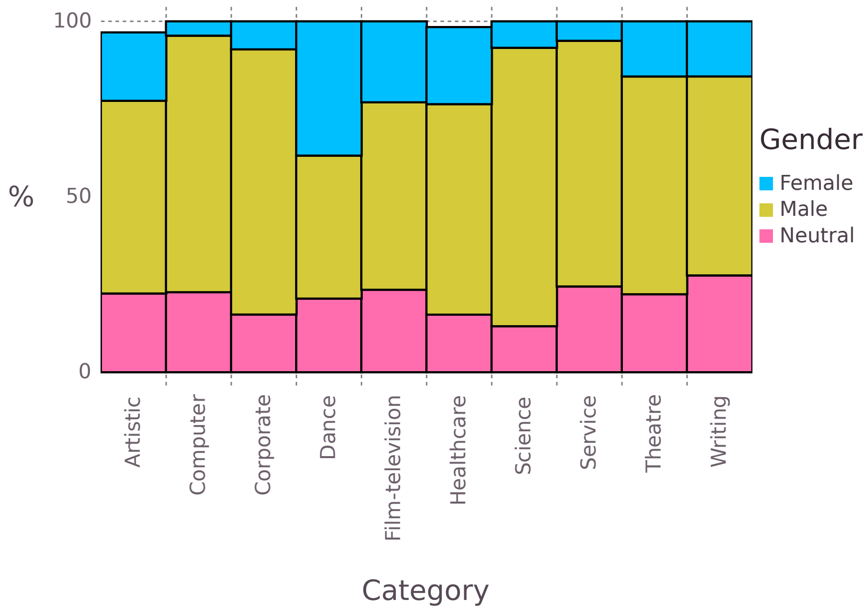

Figure 2: The distribution of pronominal genders in the translated sentences not only suggests a tendency towards male defaults but also reflects the structure of male dominated fields, with the proportion of male pronouns amounting to 73% in computer related jobs and 76% in corporate jobs respectively. Because Google Translate occasionally fails to translate a sentence, the bars for some categories fail to add up to 100%.

Language	Female	Male	Neutral	Ratio	Total
Malay	43	392	0	9.116	436
Estonian	121	309	0	2.554	436
Finnish	167	263	0	1.575	436
Hungarian	174	255	2	1.465	436
Armenian	94	337	1	3.585	436
Japanese	2	207	222	103.500	436
Turkish	19	368	44	19.368	436
Yoruba	11	290	131	26.364	436
Basque	5	45	380	9.000	436
Swahili	68	363	0	5.338	436
Chinese	13	359	58	27.615	436
Total	717	3188	838	4.446	

Table 6: Number of female, male and neutral pronominal genders per language in the translated sentences. The corresponding sex ratios (# Male / # Female) show just how much male defaults are prominent in some languages such as Chinese, with almost 30 male pronouns for each female one.

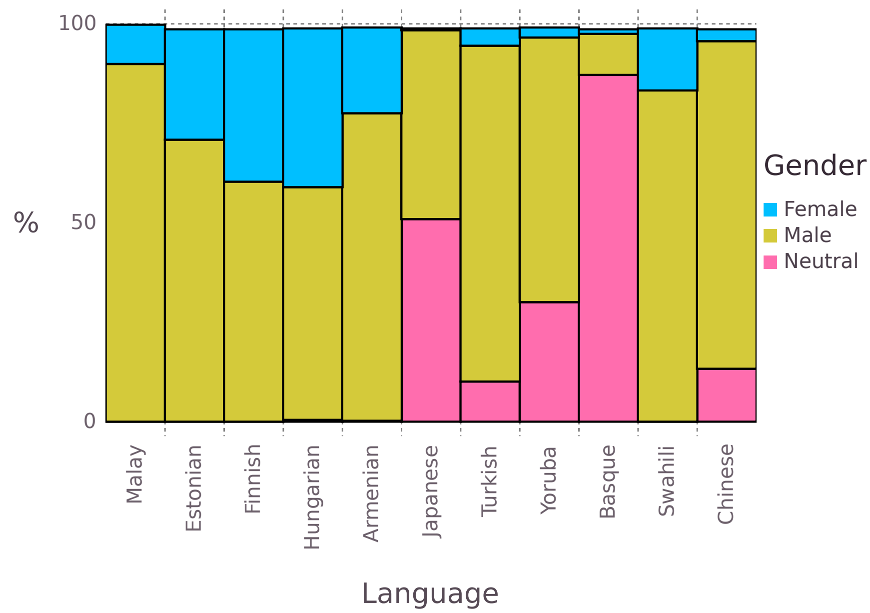

Figure 3: The distribution of pronominal genders per language also suggests a tendency towards male defaults, with female pronouns reaching as low as 0.46% and 2.98% for Japanese and Chinese respectively. Some languages such as Japanese (and particularly Basque) were observed to yield a high number of neutral pronouns, but that is the exception rather than the rule among the tested idioms. Once again not all bars add up to 100% as Google Translate occasionally fails to translate sentences.

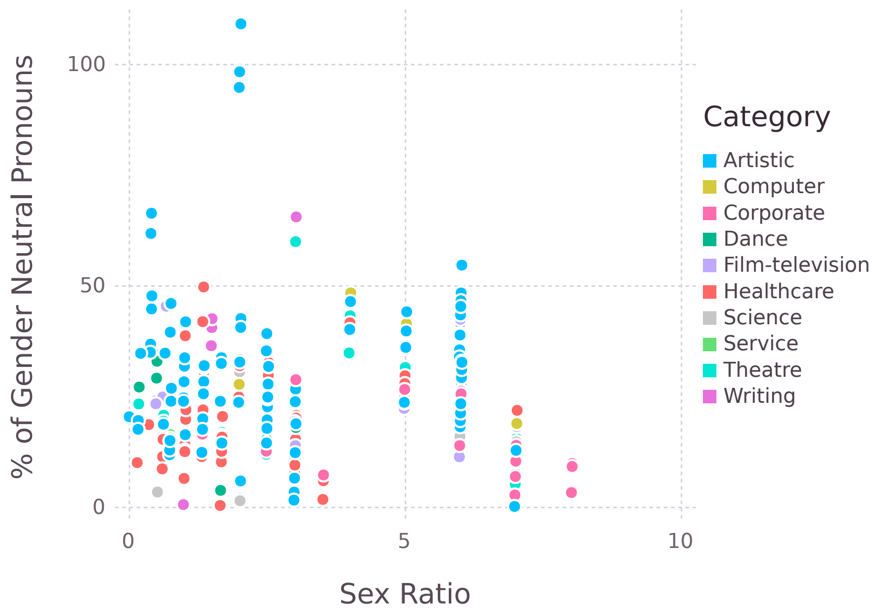

Figure 4: Scatter plot of translated sentences' statistics. Each point (color coded according to its category) corresponds to a single occupation, of which the sex ratio and the percentage of gender neutral pronouns are averaged over all tested languages (Malay, Estonian, Finnish, Hungarian, Armenian, Japanese, Turkish, Yoruba, Basque, Swahili and Chinese).

5 Discussion

References

[1] Marcelo Dascal. Universal language schemes in england and france, 1600-1800 comments on james knowlson. *Studia leibnitiana*, pages 98–109, 1982.

[2] Warren Weaver. Translation. *Machine translation of languages*, 14:15–23, 1955.

[3] Michael D Gordin. *Scientific Babel: How science was done before and after global English*. University of Chicago Press, 2015.

[4] William John Hutchins. *Machine translation: past, present, future*. Ellis Horwood Chichester, 1986.

[5] Philipp Koehn. *Statistical machine translation.* Cambridge University Press, 2009.

[6] Michael Carl and Andy Way. *Recent advances in example-based machine translation*, volume 21. Springer Science & Business Media, 2003.

[7] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*, 2016.

[8] Yoshua Bengio, Ian J Goodfellow, and Aaron Courville. Deep learning. *Nature*, 521:436–444, 2015.

[9] Peter Norvig. On chomsky and the two cultures of statistical learning. In *Berechenbarkeit der Welt?*, pages 61–83. Springer, 2017.

[10] Londa Schiebinger. Scientific research must take gender into account. *Nature*, 507(7490):9, 2014.

[11] Megan Garcia. Racist in the machine: The disturbing implications of algorithmic bias. *World Policy Journal*, 33(4):111–117, 2016.

[12] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias: There's software used across the country to predict future criminals. and it ' s biased against blacks. *https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing*, 2016.

[13] Yilun Wang and Michal Kosinski. Deep neural networks are more accurate than humans at detecting sexual orientation from facial images. *Journal of Personality and Social Psychology (in press)*, 2017.