March 22, 2018

Dear Editors:

Firstly, we would like to thank the reviewers for their comments. We have thought thoroughly over them and followed all of their insightful suggestions. The reviewers suggestions and criticism contributed in a great extent to improve our manuscript. Next, we present a point-by-point comment of each of the reviewers comments.

Reviewer #1 (Comments to the Author):

> Firstly, I thank for authors to submit their work. This paper presents a quantitative analysis on Google Translation to discover whether gender bias exists when occupation words in other language is translated into English. What they found is interesting. Even though original language is gender-neutral, translation results are more likely to contain male-oriented English words such as "he".

Thank you for your comments.

> This paper well introduces the background of machine bias on MT (machine translation), the methodology used for this paper is sounding, and what they found are interesting; however, I think this research paper is not ready to be published. The weakest point of this paper is insufficient literature survey and hence unclear contribution. This work does not introduce what kinds of previous efforts are paid to understand machine bias (which are well-known as algorithmic bias) and to further correct them. Missing connections to existing works make me difficult to find novelty of this work. Authors should cover existing works on machine bias or algorithmic bias more extensively. For example, authors are recommended to include the following works and to survey more recent studies citing them:
> - Hajian, Sara, Francesco Bonchi, and Carlos Castillo. "Algorithmic bias: From discrimination discovery to fairness-aware data mining." Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. ACM, 2016.
> - Kirkpatrick, Keith. "Battling algorithmic bias: how do we ensure algorithms treat us fairly?." Communications of the ACM 59.10 (2016): 16-17.

We thank the reviewer for the suggestion and have included comments on both references. We have also included other references thus providing improved

related work on technical and ethical aspects of machine bias.

> Moreover, this paper requires a significant amount of re-editing. I really like Abstract and Introduction, yet the remaining sections should be written more clearly and concisely. This manuscript also includes overclaiming sentences and typos.

We thank the reviewer for pointing this out. We believe we have removed overclaiming sentences and hope that no typos remain in the text. We have reorganized the sections which are now presented in a more concise and condensed way, each with its own results and technical contents.

> For the above reasons, I believe this paper should be extensively revised to be acceptable. I also put comments in more detail for future version of this work on Palgrave communication or other venues. I believe the potentials of this work, and thus hope to see improved version of this work.

We thank the reviewer for the positive outlook and encouragement.

> *Detailed comments*
>
> p3. perhaps even partially confirmed some of Chomsky's fears. — I cannot see in what sense the results confirmed Chomsky's fears. Please describe it more specifically.

We thank the reviewer for pointing this out. We have now removed this statement.

> p3. senteces — typo.

We have now corrected this typo.

> p4. Because these tools are trained with real world data they implicitly absorb the biases and stereotypes of our society, which must then be painstakingly removed. — Overclaiming. Can we really say there exist stereotypes in our society by investigating results from a black-box algorithm run by a company? We do not know what kind of data is used for Google MT.

We agree with these comments. We have removed this part and improved on other parts with similar occurrences without overclaiming.

> p4. 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. — I don't see why authors

> focus on such specific cases - translations on occupation. What kinds of other biases could be observed in MT and why does the specific problem (on occupation) matter?

We focus on the specific case of translating occupations because it is an easily measured statistic; one can look at how much participation a job has of each gender and use that to look at the gender role on that position. We have expanded the text further to include comparisons with real world data which, we hope, will shed light on why focus on occupations. Although we focus on occupations, we also have results with adjectives in the manuscript to complement the results on other categories of translations.

> p4. "" - typo

We have now corrected this typo.

> p5. 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 on the URL. — Citing a Wikipedia article is not a good practice for science. I also believe many of Wikipedia articles are trustworthy, yet there could be incorrect or debatable information in some cases. In addition, its content also varies across time. Citing more formal sources is strongly recommended.

Thank you for your suggestions. We have now changed the Wikipedia source for the occupations with the US' Bureau of Labor Statistics, and added a reference to the World Atlas of Language Structures for the gender neutrality of languages. Although, one can argue that Wikipedia article could be used to gather lists we have made it explicit that we also considered and followed other sources for the language list.

> p7. Table 3 — I wonder why some languages only contain in their templates.

This was a problem with UTF8 formatting. We thank the reviewer for spotting this issue and hope we have corrected all occurrences of it.

> p9. (Table 4) 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. — Measuring ratio with ignoring neutral cases cannot fully represent the truth. Alternatively, the fraction of male (or female) pronominal cases could be presented. In addition, how do you say one field is dominated by male or female in our society? Is there external statistics for reference?

We added an external reference for how one can argue when a field is dominated by male or female in society. Further, we removed the measure of the ratio that ignored the neutral cases.

> p10-11. Figure 4-7 - This manuscript does not describe the main observation of each figure. That is, what do authors want to say from each figure?

We hope that we made it clear with the figure captions now what we expect to show the reader with each of the pictures.

> p12. our findings probably generalize to most translations from gender neutral idioms — Overclaiming. This study focuses on a specific case of gender bias on translating occupations, and hence cannot generalize machine biases on gender that can be observed in other MT tasks.

We have now expressed our ideas more clearly. We do not claim that our results generalize to other machine translation tasks apart from the context of occupations; what we claim instead is that our results possibly generalize to other combinations of translated languages. In our study, we translated sentences from gender neutral languages to a single non-gender neutral language, namely English. A comprehensive study would require selecting a sample of multiple gender neutral languages and performing translations among all of them. We did not go into these lengths because, as we explain in the paper, Google Translate typically uses English as a *lingua franca* when translating from idiom A to idiom B (that is to say GT translates from A to English and then from English to B, see `https://cloud.google.com/translate/docs/languages#languages-nmt`). We also do not claim that our results generalize beyond Google Translate itself. We believe we have now made both points clearer in the current version of the paper.

> p13. We think this work can shed further light on some of the technical and ethical difficulties that arise from statistical machine translation, and hope that it will lead to discussions about the role of AI engineers on minimizing potentially harmful effects of the pressing modern concerns of machine bias. — The authors did not describe "technical and ethical difficulties that arise from statistical machine translation" in the manuscript. Please describe contributions of this work more specifically.

We have now included references to studies discussing both technical and ethical aspects of machine bias and machine translation. In particular, we have included references showing at what lengths developers must go to *debias* models trained on real-world data. We have also included a small section in the current version of the paper describing our assumptions about what kind of behaviour should be

expected from a machine translation tool if we are to consider it not negatively biased towards female users.

Other comments

> What would it happen if one translates gender-neutral version of English sentences into other languages?

Thank you for your suggestion. We are also curious about this point, but we leave such an analysis (and others) for future studies.

> For figures: Resolution should be improved.

We thank the reviewer for noticing this, and we have worked to improve the resolution in the needed places.

Reviewer #2 (Comments to the Author):

> This paper tests a hypothesis that Google Translate (GT) are biased against female computer scientists. The results of analysis support the author(s)'s surmise, and help the author(s) to raise the issue of gender inequality in the computer assisted algorithms which are supposed to be neutral.

Thank you for your comments. However, we would like to make our intentions clearer. We aim at showing that there is an algorithmic bias towards women in general, not only female computer scientists. We also do not suppose these algorithms would be gender-neutral, since their statistical nature means they may carry some of the gender inequalities existent in their training data. We have tried to make our statements and assumptions more clear in this current version of the manuscript.

> While the problematic sounds legit, this research is in need of a more clear theory, which should be epitomized in the null hypothesis. The author(s) seems to assume that GT should produce 50:50 gender pronouns in its translation. But there is little discussion about why and how it should be accepted positively as well as normatively.

We assumed that GT must reproduce a gender distribution that is similar to the one in the real world (and thus, statistically presented in the training data). We have created a section purely to specify our assumptions and preliminaries over the tool we are assessing. We have conducted further studies in the current version of the paper, the results of which are summarized on Section 8, to compare the demographics of women employment with the frequency of male defaults in Google Translate. Our results show that Google Translate translates sentences using male defaults with a greater frequency than what employment statistics would suggest. In this context, our null hypothesis is indeed rejected.

> Let H0: When translating gender a neutral pronoun, the chance for GT to produce a female pronoun would be same as to produce a male pronoun.
>
> If it were the H0 the author(s) aims to build up, following theoretical and empirical issues should be addressed.
>
> No matter what gender ratio we get from GT, the solution for the observed gender bias is very simple. GT is mechanically keep 50:50 female/male odds when translating gender neutral pronouns (tertiary genders are not considered here). GT does not have to follow a logistic distribution with given contexts and factors. That seems why the author(s) starts with Chomsky's criticism of statistical reasoning. However, it is still an open question what is a good translation: Keep the norm of gender equality or represent the empirical reality?

Thank you for your comments. We have now added additional references and make use of them to suggest why gender equality in translation should be a goal, and we also analysed some references and take into account referees suggestions to improve this issue.

> To clarify this issue, the author(s) should have explained what algorithm/logic is behind GT. If GT is designed to follow the distribution of the particular area, GT is exonerated from the critical assessment. Then, the real question becomes whether the "statistical discrimination" exists in GT, and H0 should be that the odds ratio of female/male in GT results to the female/male in the empirical area would be 1.

Unfortunately, one can have no full access to the algorithms and datasets behind GT and thus one cannot explain it in detail. We have added a session on our assumptions and preliminaries about it, though, and hope that it helps showing what was our hypothesis. We have also referenced a recent study which proposes a simple *debiasing* algorithm. Although it does not provide a full description of the inner workings of MT tools – or Google Translate, for that matter –, it suggests that gender bias is entrenched in the behaviour of word embedding algorithms. This suggests that although GT is probably not designed to follow distributions of particular areas, machine bias is still present. Also, we conducted further studies in the current version of the paper which show that GT translates sentences using male defaults with a larger frequency than job occupation statistics would suggest. In this context, GT is definitely not exonerated from critical assessments.

> With more of data analysis, the author(s) would be able to address these points of improvements.

Thank you. We have now improved the datasets and made more data analysis,

which, we hope, enriched our discussion.

Finally, we would like to thank the reviewers for their constructive comments.

Yours sincerely,

The authors