DeBiasByUs: Raising Awareness and Creating a Database of MT Bias

Joke Daems
Postdoctoral Researcher @ LT3 UGent
joke.daems@ugent.be

Janiça Hackenbuchner
Research Associate @ TH Köln
janica.hackenbuchner@th-koeln.de

Abstract

This paper presents the project proposed by the DeBiasByUs[1] team resulting from the Artificially Correct Hackathon. We briefly explain the hackathon challenge on 'Database and detection of gender bias in A.I. translations', highlight the importance of gender bias in Machine Translation (MT), describe our solution, the current status of the project, and our future plans.

1 Introduction to DeBiasByUs

The DeBiasByUs project was a winning solution to a challenge on 'Creating Datasets and Resources against Societal Biases in AI[2]' at the Artificially Correct Hackathon organised by the Goethe-Institut in October 2021. The initial Hackathon team consisted of five participants[3], of which the authors of the present paper will continue to develop the project. The goal of the challenge was to define and analyse gender bias from MT systems and either create a dataset or a platform for users to gather, describe, and discuss cases of bias.

2 The Problem of Gender Bias in MT

MT systems are trained with data that contain biases present in our society and in our language. As such, these systems will reproduce or even heighten these biases, potentially leading to discrimination and harm. For example, translation datasets have a dominance of white male representation (Saunders and Byrne, 2020), and word embeddings (used to train MT systems) have been shown to reinforce gender stereotypes (Bolukbasi et al., 2016).

Different factors can contribute to bias in MT. There are linguistic factors, socio-cultural factors, reinforcement of historical gender stereotypes, especially in professions, and a lack of an explicit linguistic representation of nonbinary gender. While less apparent for genderless languages (e.g. Finnish) and notional gender languages (e.g. Danish), MT most often exhibits gender bias or opts for the generic masculine for grammatical gender languages (e.g. Spanish), where nouns, verbs, adjectives etc. carry gender inflections (Savoldi et al., 2021). Technical factors include MT sampling methods favoring masculine forms due to asymmetrical gender distributions in the training datasets, leading to reinforcement of gender stereotypes as the most common form is subsequently being chosen as a most-likely translation by the MT system (Shah et al., 2020).

3 Solution: Raising Awareness and Database Creation

As a solution to the hackathon challenge, we created a website[4] that serves a dual purpose: 1) raise public awareness about the issue of gender bias in MT by providing information and research findings, and 2) create a community-driven database of occurrences of gender bias in MT. The collected inputs can be moderated and reviewed by experts. Through such collaborative and community-driven action, we aim to create a database representing different language combinations that can then be used as biased test datasets for further research. The moderated datasets will be made freely available for download. The data will consist of a source sentence (e.g., 'The Professor is an expert on machine translation') and a biased MT output

[1] The original Hackathon project was "BiasByUs" but has now been changed to "DeBiasByUs"

[2] https://www.goethe.de/prj/one/en/aco/ver/hac/cha.html#i7094314

[3] Joke Daems, Janiça Hackenbuchner, Bettina Koch, Bhargavi Mahesh, Shrishti Mohabey

[4] Hackathon proof of concept (to be updated): https://artificiallycorrec.wixsite.com/biasbyus

(e.g., 'Der Professor ist Experte für maschinelle Übersetzung' as an example of stereotyping, where a 'professor' is assumed to be male in German). The availability of datasets with biased MT output will support research in gender-bias by focusing on datasets with specific occurrences of gender-bias instead of using large noisy datasets. We further envision it being used for crosslinguistic and diachronic analyses of gender bias in MT (as new data will continuously be added).

With millions of online MT users noticing “how commercial systems entrench social gender expectations” (Savoldi et al., 2021), we believe that community awareness and involvement is key in tackling the challenge of bias in society and MT. Since society and gender roles are constantly evolving, the only way to ensure our technologies evolve alongside with it is to observe that evolution in real time.

Our current website is a proof of concept. There are numerous sections on concepts aiming to raise awareness (impact of gender bias, gender bias in language, gender bias in MT, and categories of bias), and users can submit occurrences of bias to our database by copy/pasting a source sentence, the MT output containing bias. Optionally, they can provide their reference suggestion for an unbiased translation, highlight the specific type of bias they encountered, offer clarifications, and name the source of the MT output, as well as their own familiarity with gender bias.

4 Further Steps

The Goethe-Institut has agreed to continue to fund our project. By October 2022, we aim to professionalise our website, expand our theoretical information on gender bias in MT, develop a browser plug-in, and secure a server[5] to host our database. The plug-in would become active when users consult MT resources online and so enable users to conveniently add instances of bias to our website.

The following aim will be to collect as much data as possible by marketing our initiative to interested users, and by collaborating with organisations, such as the Goethe-Institut, supporters of gender equality, experts in the field of both gender bias and MT, and research universities. As the database grows, it will become a rich resource for researchers working on gender-fair language and MT development.

A potential area of collaboration is with the other winning team of the Artificially Correct Hackathon Word2Vec[6], whose developed tool highlights words in a text that have a high probability of containing bias in translation. Once fully developed, this tool could be integrated on the DeBiasByUs website.

5 Conclusion

Bias awareness needs to be continuously raised as it is impossible to tell what will be the next arising bias in society (like Chinese discrimination due to the in Wuhan originated COVID-19 virus). The platform created by DeBiasByUs is an effort to help prevent bias representation in MT, by focusing on raising awareness of gender bias in MT as well as creating a community-driven database of gender-bias occurrences in MT outputs for research purposes to support collaborative work.

References

Bolukbasi, Tolga, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam T. Kalai. 2016. Man is to computer programmer as woman is to homemaker? *NIPS'16: Proceedings of the 30th International Conference on Neural Information Processing Systems*: 4356-4364.

Saunders, Danielle and Bill Byrne. 2020. Reducing Gender Bias in Neural Machine Translation as a Domain Adaptation Problem. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics: 7724-7736.

Savoldi, Beatrice, Marco Gaido, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2021. Gender bias in machine translation. *Transactions of the Association for Computational Linguistics, 9*: 845-874

Shah, Deven, Andrew H. Schwartz, and Dirk Hovy. 2020. Predictive Biases in Natural Language Processing Models: A Conceptual Framework and Overview. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics*: 5248-5284.

[5] We will most likely be able to host our database on servers at Ghent University. Upcoming project proposals related to MT and bias by the authors will also include funding requests to ensure the sustainability of the platform and database.

[6] https://www.goethe.de/prj/one/en/aco.html