An Analysis of Gender Bias studies in Natural Language Processing

Marta R. Costa-jussà

Universitat Politècnica de Catalunya, Barcelona

e-mail: marta.ruiz@upc.edu

Artificial intelligence systems copy and amplify existing societal biases, a problem that by now is widely acknowledged and studied. But is current research of gender bias in natural language processing actually moving towards a resolution, asks Marta R. Costa-jussà.

Demographic biases in artificial intelligence applications are causing alarm both in the scientific community and in the general public. From systems that recognize European or American faces but struggle with Asian ones to those that prioritize less qualified male candidates over more qualified female ones, our automatic tools are just perpetuations of our stereotypes and prejudices. Because gender bias is one of the most prevalent biases in our society, it is relevant to analyze how recent studies are defining an entire new field of research within natural language processing. The main questions to address are whether these studies propose scalable evaluation procedures and whether they have a well-defined scope, and taken together, whether they are moving in the right direction towards a resolution of the problem.

Gender bias in NLP as a research topic

Natural language processing (NLP) is the artificial intelligence branch where machines automatically analyze natural language, for speech recognition, automated translation and other applications that are already in many of our everyday lives. Well-known examples include personal assistants (e.g. Alexa), machine translation systems (e.g. Google translate), but there are many other systems interacting with us that are less visible such as those that monitor and track job applicants. While these applications are designed to enhance our lives, the worry is that they perpetuate and amplify unfairness in our society due to bias in gender, race, age, religion and geographical origins [1].

The topic of bias in NLP is not new[2]. However, deep learning has dramatically improved NLP applications [3] and it is worth focusing on the renaissance of bias as a topic from the start of the deep learning boom around 2012. We focus here on gender bias as even though it is highly relevant among general biases, it still has been understudied [4].

Gender bias can be defined as dominance, in particular contexts (e.g. in occupations or among primary social roles), of one gender over the other. As a consequence, the less dominant gender is underrepresented and stereotypes appear (i.e. *nurses* tend to be *females* and *doctors* tend to be *males*). The interaction of NLP and gender bias is two-fold. On the one hand, NLP can be a tool to detect gender bias in different social contexts, e.g. in online news [5] or advertisements [6]. On the other hand, NLP often produces gender biased systems, thereby perpetuating and amplifying gender bias in society [7, 8]. While gender bias in NLP is mostly attributed to training on large amounts of biased data, the bias amplification is due to the learning algorithms. In this context, there are several studies that propose solutions to de-bias the training data [9] and the learning algorithms [7].

Gender bias as a main topic of research is growing. Table 1 reports the number of papers published in the Association of Computational Linguistics (ACL) anthology and in arXiv that were listed in the results of web search queries on gender bias from 2015 to 2019. In table 1, the papers are ordered by year and the papers that were published in arXiv belong only to the computer science and statistics categories.

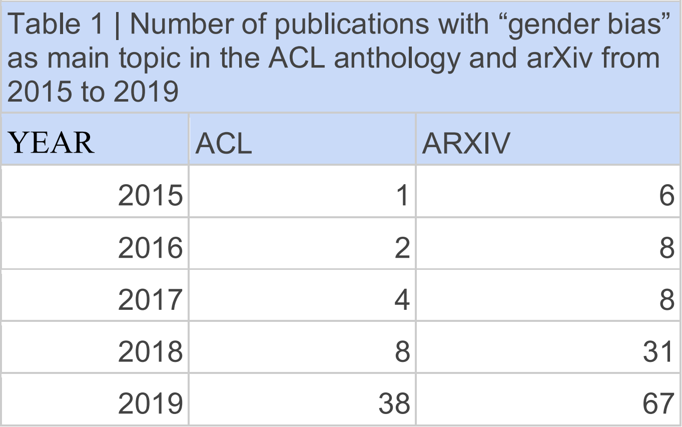

Table 1 | Number of publications with “gender bias” as main topic in the ACL anthology and arXiv from 2015 to 2019

YEAR	ACL	ARXIV
2015	1	6
2016	2	8
2017	4	8
2018	8	31
2019	38	67

While the clear increase in the number of papers on gender bias is a sign of progress and of the rise of the topic as a research field, there are some worrying issues that may undermine the necessary development of research on gender bias and, thus, adversely affect the resolution of the problem.

Some of the main challenges are to find scalable evaluation procedures and to correctly define the scope of gender bias in NLP.

Lack of coherence in evaluation

Evaluation is a key element in any research area. Several prominent evaluation methods have been explored to measure gender bias in several NLP tasks. Key examples of NLP tasks where gender bias has been studied are words embeddings and machine translation.

Word embeddings cluster words by dimensionality reduction using information of context. In pioneer works [7, 9], gender bias is exemplified by using analogies where arithmetic operations applied to word-embedded vectors show that *man* is a *computer programmer* and a *woman* is a *housekeeper*. Beyond such analogy examples (which have lately been harshly questioned [10]), gender bias is identified in word embeddings by analyzing a gender subspace and direct bias, with classification and clustering [11]. The classification method is used to show that an algorithm can use the representation of words to learn whether the word refers to a female or a male, while the clustering method is used to show that male and female words clearly show two groups. These methods have also been used to evaluate bias in contextual word embeddings [12] . This involves identifying different word embeddings for the same word, depending on the context in which it appears. In this last work, we show that evaluation methods reach different conclusions in comparing contextual and word embeddings, thereby making standardizing such tasks difficult.

Machine translation is the task of automatically translating one language to another. Examples of gender bias reported in machine translation are cases where translations match the most popular stereotypes (e.g., *she is a doctor* when translated to Turkish and then back-translated to English, becomes *he is a doctor*). There are research works [13, 14] that include a standard machine translation evaluation measure to ensure that their methodologies improve the general quality of the system. While the authors of the former study [13] do not propose other specific evaluations for gender bias, we do in ref 14. We propose evaluating a synthetic translation pattern that has the ambiguity of the gender information embedded in the co-reference. And co-reference occurs when two different words within the same text refer to the same entity or person. Then, we compare de-biasing machine translation techniques in terms of accuracy. Another study [15] uses a corpus based on a variety of synthetic patterns of ambiguous gender embedded in co-reference to compare the gender bias encoded in commercial machine translation systems.

While there are indeed interesting ways of detecting and evaluating gender bias in different fields of NLP, there is little consensus about which way is best. Small steps towards this consensus include special attention that has been given to stereotypes (specially, occupations) and co-reference in the tasks of word embeddings and machine translation, respectively. Of course, each NLP task has specific evaluation methods and it seems natural that evaluation of gender bias has been done differently for each NLP task. But coherence should be kept for each task and across languages. In addition to this desired consensus and coherence, another missing ingredient that is necessary to achieve a good evaluation method is the establishment of gold standards that are not artificially created.

Extending research to other languages and cultures

The above discussion shows that the evaluation of gender bias is challenging making it more difficult to find solutions to the problem. Languages and cultures encode gender differently: some ignore gender difference, while others explicitly encode it in articles or in morphemes [16]. Speakers of more gender-biased languages tend to have greater psychological gender bias, thereby suggesting that language statistics could lead to the emergence of individuals' gender stereotypes [17]. The innocent fact that the *sun* is masculine (in Romance languages) or feminine (in German) may possibly lead to many nuances. In fact, like stereotypes based on cultural habits, stereotypes reflected in gender divisions in languages are shown to have a great influence. While a causal path has not clearly been defined in current works[17], obtained correlations make exploring such a path in depth worthwhile to reduce the negative impact of gender stereotypes.

There are several proposals to neutralize gender biases in language [18], which include using paired forms (he/she) and neutral words (participant), especially in education, as young minds tend to have less formed stereotypes. Using strategies like these to neutralize gender bias in our languages, we can obtain the data that is needed to properly train unbiased NLP systems.

Naturally, gender bias is one type of demographic bias among many others (e.g. social, race, origins) and an important question is how to extend the findings of all the gender bias studies to these other dimensions. The scope of gender bias may also be extended to a wider community to include queer and trans people.

While neither of the previously mentioned extensions is straightforward, research on gender bias will hardly impact our everyday lives if it is limited to the study of occupational stereotypes and exclusively focused on the English language, as most research in word embeddings and machine translation seems to be. Reasonable extensions of current research should consider generalizing to multiple languages and cultures. Multilinguality can be addressed through transfer learning approaches[19], which can derive debiased techniques learnt from high-resourced languages to low-resourced ones. However, the impact of gender bias in NLP on different cultures can only be addressed fully by widening the research to social science communities and by involving underrepresented demographic groups. We need to rely on identifying and pursuing such coherent research directions to reducegender bias in NLP applications.

Acknowledgements

This work is supported by the Spanish Ministry of Economy and Competitiveness and the European Regional Development Fund (MINECO/ERDF, EU) through the programme Ramón y Cajal.

References

1. O’Neil, C. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy* (Crown Publishing Group, New York, NY, USA, 2016).

2. Calders, T. & Verwer, S. Three naive bayes approaches for discrimination- free classification. *Data Min. Knowl. Discov.* **21**, 277–292 (2010). URL http://dx.doi.org/10.1007/s10618-010-0190-x.

3. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).

4. Cislak, A., Formanowicz, M. & Saguy, T. Bias against research on gender bias. *Scientometrics* **115**, 189–200 (2018). URL https://doi.org/10.1007/s11192-018-2667-0.

5. Ross, K., Boyle, K., Carter, C. & Ging, D. Women, men and news. *Journalism Studies* **19**, 824–845 (2018). URL https://doi.org/10.1080/1461670X.2016.1222884. https://doi.org/10.1080/1461670X.2016.1222884.

6. Sweeney, L. Discrimination in online ad delivery. *Queue* **11**, 10:10–10:29 (2013). URL http://doi.acm.org/10.1145/2460276.2460278.

7. Bolukbasi, T., Chang, K.-W., Zou, J., Saligrama, V. & Kalai, A. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Proceedings*

of the 30th International Conference on Neural Information Process- ing Systems, NIPS'16, 4356–4364 (Curran Associates Inc., USA, 2016). URL http://dl.acm.org/citation.cfm?id=3157382.3157584.

8. Caliskan, A., Bryson, J. J. & Narayanan, A. Semantics derived automati- cally from language corpora contain human-like biases. *Science* **356**, 183–186 (2017). URL https://science.sciencemag.org/content/356/6334/183. https://science.sciencemag.org/content/356/6334/183.full.pdf.

9. Zhao, J., Zhou, Y., Li, Z., Wang, W. & Chang, K. Learning gender-neutral word embeddings. *CoRR* **abs/1809.01496** (2018). URL http://arxiv.org/abs/1809.01496. 1809.01496.

10. Nissim, M., van Noord, R. & van der Goot, R. Fair is better than sensational: Man is to doctor as woman is to doctor. *CoRR* **abs/1905.09866** (2019). URL http://arxiv.org/abs/1905.09866. 1905.09866.

11. Gonen, H. & Goldberg, Y. Lipstick on a pig: Debiasing methods cover up system- atic gender biases in word embeddings but do not remove them. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 609– 614 (Association for Computational Linguistics, Minneapolis, Minnesota, 2019). URL https://www.aclweb.org/anthology/N19-1061.

12. Basta, C., Costa-jussà, M. R. & Casas, N. Evaluating the underlying gender bias in contextualized word embeddings. In *Proceedings of the First Workshop on Gender Bias in Natu- ral*

Language Processing, 33–39 (Association for Computational Linguistics, Florence, Italy, 2019). URL https://www.aclweb.org/anthology/W19-3805.

13. Vanmassenhove, E., Hardmeier, C. & Way, A. Getting gender right in neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 3003–3008 (Association for Computational Linguistics, Brussels, Belgium, 2018). URL https://www.aclweb.org/anthology/D18-1334.

14. Escudé Font, J. & Costa-jussà, M. R. Equalizing gender bias in neural machine translation with word embeddings techniques. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, 147–154 (Association for Computational Linguistics, Florence, Italy, 2019). URL https://www.aclweb.org/anthology/W19-3821.

15. Stanovsky, G., Smith, N. A. & Zettlemoyer, L. Evaluating gender bias in machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguis- tics*, 1679–1684 (Association for Computational Linguistics, Florence, Italy, 2019). URL https://www.aclweb.org/anthology/P19-1164.

16. Menegatti, M. & Rubini, M. Gender bias and sexism in language (2017). URL https://oxfordre.com/communication/view/10.1093/acrefore/9780190228613.0

17. Lewis, M. & Lupyan, G. What are we learning from language? associations between gender biases and distributional semantics in 25 languages (2019). URL psyarxiv.com/7qd3g.

18. Lindqvist, A., Renström, E. A. & Gustafsson Sendén, M. Reducing a male bias in language? establishing the efficiency of three different gender-fair language strategies. *Sex Roles* **81**,

109–117 (2019). URL https://doi.org/10.1007/s11199-018-0974-9.

19. Aharoni, R., Johnson, M. & Firat, O. Massively multilingual neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 3874–3884 (Association for Computational Linguistics, Minneapolis, Minnesota, 2019). URL https://www.aclweb.org/anthology/N19-1388.