GeBNLP 2020

The Second Workshop on Gender Bias in Natural Language Processing

Proceedings of the Workshop

December 13, 2020
Barcelona, Spain (Online)

The organizers gratefully acknowledge the support they received: Marta R. Costa-jussà from the Spanish Ministerio de Ciencia e Innovación and the Agencia Estatal de Investigación, through the postdoctoral senior grant Ramón y Cajal, and from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 947657), and Christian Hardmeier from the Swedish Research Council under grant 2017-930.

ISBN 978-1-952148-43-9

Preface

This volume contains the proceedings of the Second Workshop on Gender Bias in Natural Language Processing held in conjunction with the 28th International Conference on Computational Linguistics in Barcelona. The workshop received 19 submissions of technical papers (11 long papers, 8 short papers), of which 12 were accepted (8 long, 4 short), for an acceptance rate of 63%. We thank the Program Committee members, who provided extremely valuable reviews to help us compile an exciting programme of high-quality research works.

This year we are especially grateful to the new programme committee members from the social sciences and humanities who provided feedback on the bias statements, a new feature that we asked authors to include in their research papers. The idea behind this requirement is to encourage a common format for discussing the assumptions and normative stances inherent in any research on bias, and to make them explicit so they can be discussed. This is inspired by the recommendations by Blodgett et al. (2020)[1], and we borrow from them in our definition of the bias statement. We provided a blog post available from the workshop webpage, which explicitly provided some guidance to help authors in writing a bias statement. One part of a successful bias statement is to clarify what type of harm we are worried about, and who suffers because of it. Doing so explicitly serves two purposes. On the one hand, by describing certain behaviours as harmful, we make a judgement based on the values we hold. It's a normative judgement, because we declare that one thing is right (for instance, treating all humans equally), and another thing wrong (for instance, exploiting humans for profit). On the other hand, being explicit about our normative assumptions also makes it easier to evaluate, for ourselves, our readers and reviewers, whether the methods we propose are in fact effective at reducing the harmful effects we fear, and that will help us make progress more quickly.

The accepted papers cover a wide range of applications in natural language processing, including words embeddings, topic modelling, poetry composition, sentiment analysis, conversational assistants and neural machine translation. Within these applications, these papers cover a variety of gender (and intersectional) bias approaches, including dataset generation, mitigation algorithms, evaluation and bias-aware research methodology.

Finally, the workshop counts on two impressive keynote speakers: Natalie Schluter, who in addition to being a Senior Research Scientist at Google Brain and an Associate Professor at the IT University of Copenhagen is also the first Equity Director of the Association for Computational Linguistics, and Dirk Hovy, an Associate Professor at Bocconi University with a distinguished publication record on bias and social aspects of NLP.

We are very excited about the interest that this workshop has generated and we look forward to a lively discussion about how to tackle bias problems in NLP applications when we meet virtually on the 13th December 2020!

November 2020
Marta R. Costa-jussà, Christian Hardmeier, Will Radford, Kellie Webster

[1] Blodgett, Su Lin et al. "Language (Technology) is Power: A Critical Survey of 'Bias' in NLP." *ACL* (2020).

Organizers:

Marta R. Costa-jussà, Universitat Politècnica de Catalunya (Spain)
Christian Hardmeier, Uppsala University (Sweden)
Will Radford, Canva (Australia)
Kellie Webster, Google AI (USA)

Programme Committee:

Dorna Behdadi, University of Gothenburg (Sweden)
Jenny Björklund, Uppsala University (Sweden)
Su-Lin Blodgett, University of Massachussetts Amherst (USA)
Matthias Gallé, NAVER LABS Europe (France)
Mercedes García-Martínez, Pangeanic (Spain)
Zhengxian Gong, Soochow University (China)
Ben Hachey, Harrison.ai (Australia)
Dirk Hovy, Bocconi University (Italy)
Svetlana Kiritchenko, National Research Council (Canada)
Sharid Loáiciga, University of Potsdam (Germany)
Carla Perez Almendros, Cardiff University (UK)
Vinodkumar Prabhakaran, Stanford SPARQ, Google Research (USA)
Marta Recasens, Google (USA)
Sonja Schmer-Galunder, Smart Information Flow Technologies (USA)
Sverker Sikström, Lund University (Sweden)
Kathleen Siminyu, Artificial Intelligence for Development – Africa Network
Bonnie Webber, University of Edinburgh (UK)
Steven Wilson, University of Edinburgh (UK)

Invited Speakers:

Natalie Schluter, IT University of Copenhagen/Google Brain (Denmark)
Dirk Hovy, Bocconi University (Italy)

Table of Contents

Conference Program

Sunday, December 13, 2020

09:00–09:10 ***Introductory Remarks***

09:10–10:00 ***Keynote: Natalie Schluter***

The Impact of a Gender in NLP

Intersectionality

10:00–10:15 *Unmasking Contextual Stereotypes: Measuring and Mitigating BERT's Gender Bias*
Marion Bartl, Malvina Nissim and Albert Gatt

10:15–10:25 *Interdependencies of Gender and Race in Contextualized Word Embeddings*
May Jiang and Christiane Fellbaum

10:25–10:35 ***Shared Q&A***

10:35–11:10 ***Break***

Machine Translation and Multilinguality

11:10–11:20 *Fine-tuning Neural Machine Translation on Gender-Balanced Datasets*
Marta R. Costa-jussà and Adrià de Jorge

11:20–11:30 *Neural Machine Translation Doesn't Translate Gender Coreference Right Unless You Make It*
Danielle Saunders, Rosie Sallis and Bill Byrne

11:30–11:45 *Can Existing Methods Debias Languages Other than English? First Attempt to Analyze and Mitigate Japanese Word Embeddings*
Masashi Takeshita, Yuki Katsumata, Rafal Rzepka and Kenji Araki

11:45–12:00 *Evaluating Bias In Dutch Word Embeddings*
Rodrigo Alejandro Chávez Mulsa and Gerasimos Spanakis

12:00–12:30 ***Shared Q&A***

12:30–14:00 ***Break***

Sunday, December 13, 2020 (continued)

14:00–14:50 ***Keynote: Dirk Hovy***

Sampling, Syntax, and Sentence Completions – The (Overlooked?) Impact of Gender on NLP Tools

NLP Applications

14:50–15:00 *Conversational Assistants and Gender Stereotypes: Public Perceptions and Desiderata for Voice Personas*
Amanda Cercas Curry, Judy Robertson and Verena Rieser

15:00–15:15 *Semi-Supervised Topic Modeling for Gender Bias Discovery in English and Swedish*
Hannah Devinney, Jenny Björklund and Henrik Björklund

15:15–15:30 *Investigating Societal Biases in a Poetry Composition System*
Emily Sheng and David Uthus

15:30–15:45 ***Shared Q&A***

15:45–16:15 ***Break***

Data and Methodology

16:15–16:30 *Situated Data, Situated Systems: A Methodology to Engage with Power Relations in Natural Language Processing Research*
Lucy Havens, Melissa Terras, Benjamin Bach and Beatrice Alex

16:30–16:45 *Gender and sentiment, critics and authors: a dataset of Norwegian book reviews*
Samia Touileb, Lilja Øvrelid and Erik Velldal

16:45–17:00 *Gender-Aware Reinflection using Linguistically Enhanced Neural Models*
Bashar Alhafni, Nizar Habash and Houda Bouamor

17:00–17:15 ***Shared Q&A***

17:15–17:30 ***Closing Remarks***