Reproducibility study: Identifying and reducing gender bias in word-level language models (Bordia & Bowman, 2019)

Angelo Groot, Eui Yeon Jang, Reitze Jansen, Tom Kersten

University of Amsterdam

January 31, 2020

Overview

1. Bordia & Bowman (2019): Original Study
2. Methods: Reproducing the Study
3. Experiment
4. Results
5. Conclusion & Discussion

Bordia & Bowman (2019): Identifying and reducing gender bias in word-level language models

- Gender bias in text corpora
  - Leads to gender bias in language models
  - Reinforces future gender biases
- Bordia & Bowman (2019):
  1. Train a word-level language model with proposed method
  2. Generate text using the model
  3. Measure bias in data set and generated text

Overview

1. Bordia & Bowman (2019): Original Study
2. Methods: Reproducing the Study
3. Experiment
4. Results
5. Conclusion & Discussion

Bordia & Bowman (2019): Language Model

ASGD Weight-Dropped LSTM (AWD-LSTM) as described in Merity et al. (2018):

- Supplied model did not run
- Code was unclear and not properly documented
- Rebuilt from the ground up

Bordia & Bowman (2019): Debiasing Method

During training, the following regularisation loss is used to minimise the gender bias:

$$\mathcal{L}_B = \lambda ||NB||_F^2$$

- N: Matrix of embeddings of the words we wish to debias
- B: Gender subspace based on the gender pairs
- λ: Importance of the debiasing loss term during training.

Bordia & Bowman (2019): Bias Metric

The gender bias of word w within a corpus T:

$$bias_T(w) = log\left(\frac{P(w|f)}{P(w|m)}\right)$$

with f being a female context and m being a male context

- **Fixed Context**:
  - Window size $k = 10$
  - Weight of all words uniformly 0.05.
- **Infinite Context**:
  - Infinite window size
  - The weight of a word is $0.05 \cdot 0.95^{d(word,w)}$

Bordia & Bowman (2019): Problems with the Bias Metric

The gender bias of word w within a corpus T:

$$bias_T(w) = log\left(\frac{P(w|f)}{P(w|m)}\right)$$

However...

- Discrepancies between paper and implementation
- No specifications for handling zero-probabilities

Bordia & Bowman (2019): Evaluating bias reduction

- Measure bias for each word w
- Calculate mean absolute μ_λ and standard deviation σ_λ over bias
- Calculate the amplification measure β: $bias_\lambda(w) = \beta \cdot bias_{train}(w) + c$

Overview

Reproducing the paper: Data sets used and preprocessing

We use the same data sets as Bordia & Bowman:

- Penn Treebank (Marcus et al., 1993)
  - Wall Street journal stories
  - As preprocessed in Mikolov et al. (2010)
- WikiText-2 (Merity et al., 2016).
  - Wikipedia articles
- CNN/Daily mail (Hermann et al. (2015))
  - Combined data set of CNN and Daily Mail stories

Reproducing the paper: Experimental setup

- Train, generate, measure
- λ: 0, 0.001, 0.01, 0.1, 0.5, 0.8, 1
- Generate 2000 files, 500 words
- Measure using proposed bias metric

Overview

Results: Penn Treebank data set

	Original				**Reproduced**			
λ	μ	σ	β	*Ppl.*	μ	σ	β	*Ppl.*
train	0.83	1.00	-	-	**2.06**	0.98	-	-
0.0	0.74	0.91	0.40	62.56	2.55	0.89	0.33	111.05
0.001	0.69	0.88	0.34	62.69	2.09	**0.81**	**0.27**	125.21
0.01	**0.63**	**0.81**	**0.31**	62.83	2.52	0.87	0.32	127.02
0.1	0.64	0.82	0.33	**62.48**	3.28	0.95	0.32	**109.90**
0.5	0.70	0.91	0.39	62.50	2.63	0.87	0.32	116.25
0.8	0.76	0.96	0.45	63.36	2.74	0.87	0.32	121.52
1.0	0.84	0.94	0.38	62.63	3.84	1.02	0.36	111.90

Table: Reproduced and original results for the Penn Treebank data set and generated text for different λ values with fixed context.

Results: Wikitext-2 data set

	Original				**Reproduced**			
λ	μ	σ	β	*Ppl.*	μ	σ	β	*Ppl.*
train	0.80	1.00	-	-	**1.39**	1.00	-	-
0.0	0.70	0.84	0.29	**67.67**	1.56	**0.73**	0.14	
0.001	0.69	0.84	0.27	67.84	1.65	0.75	0.13	
0.01	**0.61**	**0.79**	**0.20**	67.78	1.86	0.77	0.12	
0.1	0.65	0.82	0.24	67.89	2.05	0.80	0.15	
0.5	0.70	0.88	0.31	69.07	2.02	0.80	0.16	
0.8	0.65	0.84	0.28	69.36	2.11	0.81	0.14	
1.0	0.74	0.92	0.27	69.56	1.63	0.76	**0.10**	

Table: Reproduced and original results for the Wikitext-2 data set and generated text for different λ values with fixed context.

Overview

Conclusion and Discussion

- Implementation deviates from paper
- Experimental setup not specified nor motivated
- Hindered ability to reproduce the results
- Unable to conclude on their method nor metric
- Broader implications: awareness of issues posed in the domains of **FA**C**T**
- At most, 'Artifact Available' badge

Awarding an ACM Badge

References

1. Shikha Bordia and Samuel R Bowman. 2019. Identifying and reducing gender bias in word-level language models.arXivpreprintarXiv:1904.03035 (2019).
2. Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of english: The penn treebank. Computational Linguistics, 19(2):313–330.
3. Tomas Mikolov, Martin Karafiat, Luk 'as Burget, Jan' Cernocky, and Sanjeev Khudanpur. 2010. Recur- ' rent neural network based language model. In INTERSPEECH, pages 1045–1048. ISCA
4. Stephen Merity, Nitish Shirish Keskar, and Richard Socher. 2018. Regularizing and Optimizing LSTM Language Models. In International Conference on Learning Representations. https://openreview.net/forum?id=SyyGPP0TZ
5. Karl Moritz Hermann, Tomás Kociský, Edward Grefenstette, Lasse Espeholt,Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching Machines toRead and Comprehend.CoRRabs/1506.03340 (2015). arXiv:1506.03340 http://arxiv.org/abs/1506.03340

Gender pairs per dataset: Penn Treebank

- Penn Treebank
  - **male**: “actor” “boy” “father” “he” “him” “his” “male” “man” “men” “son” “sons” “spokesman” “wife” “king” “brother”
  - **female**: “actress” “girl ” “mother” “she” “her ” “her” “female” “woman” “women” “daughter” “daughters” “spokeswoman” “husband” “queen” “sister”

Gender pairs per dataset: WikiText-2

- WikiText-2
  - **male**: “actor” “Actor” “boy” “Boy” “boyfriend” “Boys” “boys” “father” “Father” “Fathers” “fathers” “Gentleman” “gentleman” “gentlemen” “Gentlemen” “grandson” “he” “He” “hero” “him” “Him” “his” “His” “Husband” “husbands” “King” “kings” “Kings” “male” “Male” “males” “Males” “man” “Man” “men” “Men” “Mr.” “Prince” “prince” “son” “sons” “spokesman” “stepfather” “uncle” “wife” “king”
  - **female**: “actress” “Actress” “girl” “Girl” “girlfriend” “Girls” “girls” “mother” “Mother” “Mothers” “mothers” “Lady” “lady” “ladies” “Ladies” “granddaughter” “she” “She” “heroine” “her” “Her” “her” “Her” “Wife” “wives” “Queen” “queens” “Queens” “female” “Female” “females” “Females” “woman” “Woman” “women” “Women” “Mrs.” “Princess” “princess” “daughter” “daughters” “spokeswoman” “stepmother” “aunt” “husband” “queen”

Gender pairs per dataset: CNN/Daily Mail

- CNN/Daily Mail
  - **male**: “actor” “boy” “boyfriend” “boys” “father” “fathers” “gentleman” “gentlemen” “grandson” “he” “him” “his” “husbands” “kings” “male” “males” “man” “men” “prince” “son” “sons” “spokesman” “stepfather” “uncle” “wife” “king” “brother” “brothers”
  - **female**: “actress” “girl” “girlfriend” “girls” “mother” “mothers” “lady” “ladies” “granddaughter” “she” “her” “her” “wives” “queens” “female” “females” “woman” “women” “princess” “daughter” “daughters” “spokeswoman” “stepmother” “aunt” “husband” “queen” “sister” “sisters”

Dataset specifics

- Penn Tree Bank (Marcus et al., 1993)
  - Wall Street journal stories
  - As preprocessed in Mikolov et al. (2010)
  - Contains a higher count of male words than female words
  - 200 epochs with batch size 40
- WikiText-2 (Merity et al., 2016).
  - Wikipedia articles
  - More diverse. Contains a more balanced ratio of male to female words
  - 200 epochs with batch size 80
- CNN/Daily mail (Hermann et al. (2015))
  - Combined dataset of CNN and Daily Mail stories
  - Contains yet more diverse content in terms of topics and has an even more balanced ratio of male to female words
  - Bordia & Bowman do not specify exactly how this is preprocessed
  - Subsampled stories by a factor of 100, resulting in a dataset comparable in size to WikiText-2
  - for the train:validation:test split, we used a ratio of 12:1:1, which is comparable to WikiText-2
  - 150 epochs with batch size 80

Bordia & Bowman (2019): Debiasing Method

Let $\boldsymbol{u}_i$ and $\boldsymbol{v}_i$ be the embeddings of gender pair i, then:

$$C = \begin{bmatrix} \left(\frac{\boldsymbol{u}_1 - \boldsymbol{v}_1}{2}\right) \\ \left(\frac{\boldsymbol{u}_2 - \boldsymbol{v}_2}{2}\right) \\ \vdots \\ \left(\frac{\boldsymbol{u}_n - \boldsymbol{v}_n}{2}\right) \end{bmatrix} = U\Sigma V$$

the gender subspace will then be defined by

$$B = V_{1:\kappa}$$

where κ is chosen to capture 50% of the variation

Results: CNN / Daily Mail data set

	Original				**Reproduced**			
λ	μ	σ	β	*Ppl.*	μ	σ	β	*Ppl.*
train	0.72	0.94	-	-	**0.91**	0.91	-	-
0.0	0.51	0.68	0.22	118.01	1.79	0.81	0.1	261.69
0.001	-	-	-	-	1.13	0.75	0.14	260.41
0.01	-	-	-	-	1.93	0.81	0.20	260.72
0.1	0.38	0.52	0.19	116.49	1.66	0.85	0.19	**259.48**
0.5	**0.34**	**0.48**	**0.14**	**116.19**	1.11	**0.69**	**0.07**	445.91
0.8	0.40	0.56	0.19	121.00	1.14	0.77	0.17	290.22
1.0	0.62	0.83	0.21	120.55	1.25	0.75	0.15	309.66

Table: Reproduced and original results for the CNN/Daily Mail data set and generated text for different λ values with fixed context.