From: **Sheng Liang** shengliang@cis.lmu.de
Subject: Fwd: Your COLING 2020 Submission (Number 1810)
Date: 30. September 2020 at 13:32
To: Philipp Dufter philipp@cis.lmu.de, Hinrich Schuetze hs2016@cislmu.org

Hi,

I received the email notification from COLING. It's accepted. Thanks for your support.

Best,
Sheng

> **发件人:** Chengqing Zong <noreply@softconf.com>
> **日期:** 2020年9月30日 GMT+2 上午11:50:49
> **收件人:** shengliang@cis.lmu.de
> **主题: Your COLING 2020 Submission (Number 1810)**
> **回复:** "Chengqing Zong" <cqzong@nlpr.ia.ac.cn>
>
> Dear Sheng Liang:
>
> On behalf of the COLING 2020 Program Committee, I am delighted to inform you that the following submission has been accepted to appear at the conference:
>
> Monolingual and Multilingual Reduction of Gender Bias in Contextualized Representations
>
> The reviews with comments and suggestions for the final version of the paper are attached below.
>
> The COLING2020 Program Committee, composed of 51 Area Chairs and 1561 Reviewers, worked very hard to thoroughly review over 1900 submitted papers. Please follow their suggestions when you revise your paper.
>
> **1. Final Submission Instructions**
>
> When you are finished, you can upload your final manuscript at the following site:
>
> https://www.softconf.com/coling2020/papers/
>
> You will be prompted to login to your START account. If you do not see your submission, you can access it with the following passcode:
>
> 1810X-D2C9F6F7A4
>
> Alternatively, you can click on the following URL, which will take you directly to a form to submit your final paper (after logging into your account):
>
> https://www.softconf.com/coling2020/papers/user/scmd.cgi?scmd=aLogin&passcode=1810X-D2C9F6F7A4
>
> The reviews and comments are attached below. Again, try to follow their advice when you revise your paper.
>
> **2. Dual Submission Instructions**
>
> Please remember that if you have also submitted your paper to another conference or archival workshop, you will need to withdraw your paper from the other venues in order to present at COLING2020 and have your paper appear in the proceedings.
>
> **3. COLING2020 Conference**
>
> In response to the COVID-19 pandemic, we will be holding COLING2020 online on December, 8-13. The presentations at the conference will involve a combination of pre-recorded talks, live Q-A sessions, and interactive virtual poster rooms. We are still working on the final details of the virtual setup and will be sending you information about the details of your presentation in the next few weeks, along with details on registration and other practical aspects of the virtual conference. We very much appreciate your patience and understanding regarding the difficulties that the current COVID-19 pandemic is adding to the organization of a conference like ours.
>
> Congratulations on your fine work. If you have any additional questions, please feel free to get in touch.
>
> Best Regards,
> Program Chairs, COLING 2020
> COLING 2020
>
> ```
> ===========================================================================
> COLING 2020 Reviews for Submission #1810
> ===========================================================================
> ```

Title: Monolingual and Multilingual Reduction of Gender Bias in Contextualized Representation
s
Authors: Sheng Liang, Philipp Dufter and Hinrich Schütze

==
REVIEWER #1
==

Reviewer's Scores

Relevance (1-5): 5
Readability/clarity (1-5): 3
Originality (1-5): 4
Technical correctness/soundness (1-5): 4
Reproducibility (1-5): 3
Substance (1-5): 4

Detailed Comments

This paper presents a novel approach and a deeper analysis of different approaches for debias
ing contextual word embedding models such as BERT. The proposed method is based on DensRay an
d can be applied to different heads and layers of BERT. To better evaluate the approach, a ne
w dataset called OCCTMP and based on occupation names is presented. The results show that Den
sRay has the similar performance of previously proposed methods, but it seems to be more robu
st than other approaches and able to better preserve the performance of a downstream task usi
ng BERT. The last experiment shows that the proposed method trained on English data and appli
ed in zero-shot to Chinese can effectively debias the text.

The paper is interesting and full of experiments that analysed the result from different pros
pectives. A new dataset is released for better evaluating the different approaches.

Strengths:

-) The paper addresses an important problem in NLP that consists in reducing the gender bias
of AI tools. This topic has received a lot of attention in the last years in the NLP communit
y showing how deep learning tools are affected by this problem. The proposed approach shows t
hat it is able to mitigate this effect.

-) The proposed method is analysed from different prospectives with a large set of experiment
s. The settings are grounded and the results confirm that the proposed approach is a valuable
alternative to already published techniques.

-) Although DensRay has been previously proposed, its application to BERT is interesting and
useful considering the large popularity of BERT and the large set of applications where it is
used.

-) A new dataset for evaluating debiasing methods is presented and used in the evaluation. Th
is resource can be quite useful in pushing the research on reducing gender bias in work embed
ding.

Weaknesses:

-) The paper is very dense with a lot of experiments and results that, sometimes, are not ful
ly commented. For instance:

a) Section 2.5, a few words should be added to help the reader to understand where it is evid
ent in Figure 1 that the principal component does not correlate with gender.

b) Section 4.1, the differences in managing the gender between BERT base and large are not di
scussed. This would be an interesting contribution for the reader answering the question if m
ore generic training data has an effect on gender bias.

c) Section 4.2, the approaches used to remove the gender bias are not able to achieve the sam
e performance of the gender-biased model and this is considered as a good result. The gender-
biased model will make several mistakes due to the bias, so if the proposed models are able t
o better manage the bias, the overall performance should improve assuming that the generated
text is more balanced between the genders. The fact that the performance does not show this d
eserves a comment.

d) Section 4.5, the paper mentions that the Chinese language does not mak gender, so it is no
t clear the purpose of the experiments. This should be motivated in the paper.

e) The Wikitext-2 task should be better described.

Some experiments can be removed to save space and better comments added in the remaining ones
.

-) It is not clear if the OCCTMP dataset will be made available if the paper is accepted.

-) (Minor comment) the reference "Garg et al" has a different format.

Reviewer's Scores

Overall recommendation (1-5): 4
Confidence (1-5): 3
Presentation Type: Poster
Recommendation for Best Paper Award: No

==========
REVIEWER #2
==========

Reviewer's Scores

Relevance (1-5): 4
Readability/clarity (1-5): 4
Originality (1-5): 3
Technical correctness/soundness (1-5): 4
Reproducibility (1-5): 4
Substance (1-5): 4

Detailed Comments

Monolingual and Multilingual Reduction of Gender Bias in Contextualized Representations

Like prior work on gender bias mitigation removing gender information by eliminating subspace
s, the authors propose to use DensRay for obtaining interpretable dense subspaces. Using Dens
Ray performs on-par with previous approaches but is more robust and preserves the language mo
del performance better. They observe that gender information is spread across all attention h
eads and most of the layers. By using DensRay, they can obtain a gender bias score on the tok
en/sentence level.

The contributions of the paper are summarized as follows:
- adjustment of DensRay to debias contextualized embeddings
- arguments provided to show that DensRay is more robust and preserves better the LM performa
nce
- analysis of how gender info is processed in BERT by applying DensRay to attention heads and
 layers
=> Here I would definitely refer to the work by Bau et al (2018) "Identifying and Cont
- application of debiasing method to multilingual-BERT(use English data to effectively debias
 Chinese)

Overall, the work is good and there are definitely some interesting elements that I believe a
re worht presenting. I do believe the authors need to revise some of the literature and add m
ore work to their bibliography that works on similar topics. Strengths of the paper are an in
-depth analysis (including some interesting visualizations and examples), plus the fact that
they not only look at how the debiasing technique performs but also take into account the ove
rall LM performance. The major weakness is that they do not take into the account the work by
 Gonen and Goldberg (2019) which highlights major issues in how debiasing techniques are being

#Introduction
OK
I'm curious to know whether the work by Gonen & Goldberg was taken into consideration at all.
 It shows how work on debiasing word embeddings tends to indicate significant reductions in ge

#Methodology
in 2.2
"The first principal component of E is the gender direction q" => Which formula is this refer
ing to exactly? (please indicate this)

Currently this method is being appliad to English, Chinese => language that do not require ge
nder agreement. I would be curious to see what happens if you apply this type of gender debia
sing to a language that with gender agreement (e.g. Russian, French, Spanish...). These langu
ages can still be debiased, but the gender information of words needs to be retained (grammat
ical gender, which in some cases coincides with the so-called natural gender, "sister" (femal
e), "brother" (male), "father"...).

2.5
both directions are okay (depending on how you look at it?). It isn't very clear to me what t
his Figure 1 is supposed to show the reader (or what it adds to the paper) => Add some more c
larification as to why this is here?

2.6
each sentence contains 'one' of the gendered words (is this a limitation?)

each sentence contains "one" of the gendered words (is this a limitation?)

3. Experiments
- If you limit the gendered words to words from the family category, wouldn't this maybe pres
ent different information compared to gendered words such as 'waiter' vs 'waitress' or 'host'
 vs 'hostess' since the first ones focus to some extend on relations between words that are ba

3.2
What does OCCTMP stand for?

4. Results
4.1 Table 2, I find it a bit misleading to say "with Densray and Conceptor having the best pe
rformance three times, and hard-debiasing 5 tunes". Why would you suddenly group Densray and
Conceptor as one? Also, I don't see immediately in the table why hard-debiasing has the best
performance 5 times?

4.4 analysis
First conclusions similar to work Bau et al (2019) (see below for title)

I like the visualizations (Figure 2) and the example for quantifying bias (model:bert-base) i
n Table 5. Very interesting.

5. Related Work

I know not everybody agrees on this, but I would prefer to have the related work somewhere in
 the beginning, to provide some context with respect to the work that has been done. Please ac

Related Work
- Bau et al (2019) "Identifying and Controling Impotant Neurons in Neural Machine Translation
"
- Gonen and Goldberg 2019 "Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Bi
ases in Word Embeddings But do not Remove Them"
- Prost et al. (2019): extends the post-processing procedure to all words in the corpus in or
der to reduce the remaining bias
- Vanmassenhove et al (2018) "Getting Gender Right in NMT"
- Moryosseft et al (2019) "Filling Gender & Number Gaps in Neural Machine Translation with Bl
ack-box Context Injection"

References:
- please change the older arxiv papers with the proper reference (if the work has been publis
hed by now)

Reviewer's Scores

            Overall recommendation (1-5): 4
                        Confidence (1-5): 4
                      Presentation Type: Oral
     Recommendation for Best Paper Award: No

==========
                          REVIEWER #3
==========

Reviewer's Scores

                         Relevance (1-5): 5
               Readability/clarity (1-5): 4
                       Originality (1-5): 4
   Technical correctness/soundness (1-5): 4
                  Reproducibility (1-5): 4
                        Substance (1-5): 3

Detailed Comments

The paper addresses the problem of debiasing contextualized representations. To achieve this
goal, it proposes Densray (Dufter and Schutze, 2019) to learn orthogonal transformation matri
x Q for (each) layer in pretrained LM (BERT, multilingual BERT). The matrix Q is learned by c
ollecting sentences that contain words in given wordlists. Corresponding contextualized vecto
rs are extracted and the analytical solution of Q can be found. The contextualized vector (at
 a given layer) is then subtracted by the projection vector onto Q. Evaluation shows that on t

Strength:
- simple and straightforward application of Densray to contextualized model
- evaluation shows promising results
- good analysis of debiasing per attention head and per layer

Weakness:

- only show experiments for gender. Maybe the authors could show the results for other type o
f biases (e.g., racial, religious)
- while the zero-shot experiment seems interesting, I think it lacks the nuance in debiasing.
 It assumes that there is an universal bias amongst all the languages (and people who speak th

Questions:
- Could the authors explain the debiasing step in 2.4 a bit more in detail. My understanding
of the paper is that the matrix Qi is computed based on the vectors at layer i in the origina
l BERT. After debiasing using the equation in 2.4, the computation from the word embeddings u
pto layer i is the same, does the new debiased values x_i is used to compute the top layers f
rom (i+1) to the last layer?
- I also wonder when debiasing is applied to `all` layers, how principled is that? The transf
ormation matrix Qi are computed from the original model. When debiasing, let say, layer 1, th
en the dynamic of the network changes. Layer 2 or 3, for instance, will be different because
the input is now debiased. Then does it make sense to use q2 and q3 (learned from the origina
l model) to debiased further?

Reviewer's Scores

            Overall recommendation (1-5): 4
                        Confidence (1-5): 3
                      Presentation Type: Poster
     Recommendation for Best Paper Award: No

--
COLING 2020 - https://www.softconf.com/coling2020/papers