INFO 4900 Independent Research: A gap in tests and mitigation for bias in language embeddings

Table of Contents

1 Background to the Independent Research

Until recently, language models were simple. Data scientists took a large corpus of words and little bit of context, typically one or two words either side, i.e. 3-gram or 5-gram. Words and small context were encoded into embedding vectors. The idea was that related concepts would be found near each other in embedding space, i.e. small cosine distance between their vectors. How near each other depended on the data the language model was trained on.

Scientists trained language models like GloVe and word2vec on Wikipedia, NYTimes and similar public sources. Other data scientists used GloVe and word2vec as a foundation for their own NLP. Naturally, the historical biases of society were reflected in GloVe and word2vec. Words like "doctor," "nurse" and "engineer" were nearer one gender, while "criminal" and "dishonest" appeared nearer some ethnicity. AI ethics researchers created tests for detecting such biases and developed methods to reduce the effect. One method tried to balance the data on which language models were trained, e.g. more female characters in historical fiction. Another method tried to map the gender bias space inside language models and then added a bias-equalizing signal to any input to the language model.

Typically, language models were a small part of a final application. Most of the intelligence was added in additional layers of supervised learning, e.g. skill classification from resume text. While the ethical implications of bias of language models were important, the real-world effect of bias could be limited by a combination of additional intelligence in the supervised learning layers and adding more balanced data to retrain the language model itself.

The introduction of the BERT and ELMo transformation architectures in 2019 changed the importance and impact of language models. Data scientists at Google and Open AI used supercomputers to create large scale models that captured extensive context for each word, typically 256 words either side. These new types of language models have changed the field of AI radically.

Capture more context and meaning: Now entire sentences and paragraphs can be encoded into massive models with 100 billion parameters. Even more context can be added into the embeddings, e.g. the kind of publication where a sentence appears, the profile of the author, associated reviews, etc. Now, the embeddings carry a much deeper "meaning" of what is typically said when.

Enable more sophisticated AI applications: The semantic richness of the embeddings is now making possible new kinds of AI applications that were hard to build before. Multiple startups are using Open AI's GPT-3 language model. One application automatically writes product descriptions on shopping sites[1] and another writes Facebook and blog posts[2]. Other applications act as companions for people with clinical depression[3] or generate plans for therapists to use with patients[4].

Need only thin last mile layers: The language models have so much knowledge of the world that other scientists use them for "few shot learning". The last mile intelligence is trained on very small amount of data. This means that the bias in language models has a greater effect.

[1] https://copysmith.ai/
[2] https://nichesss.com/
[3] https://www.healtheuropa.eu/chatbot-therapist-to-combat-depression-un/93609/
[4] https://vitafluence.ai/. I will be interning with Vitafluence this summer

2 Focus of Independent Research and changes after initial survey

My idea was to study how AI ethics researchers were testing and mitigating bias in BERT and similar language models. The hunch was that language models were enabling new tasks and therefore I will be able to identify some important task category not adequately covered by ethics researchers.

In the initial proposal for INFO 4900 I said I would focus on Q&A and dialog systems to find a gap. But, after a few weeks of research, it appeared that conversation systems have a lot of last mile intelligence and are less influenced by language models. I switched to investigating language generation because there is relatively low last mile intelligence in these applications and the bias may have greater effect.

Once I did identify an interesting gap area in language generation, I discussed with my advisor Professor Schrader. We concluded that it was more valuable to carefully delineate the gap and propose possible test techniques. The actual creation and proving of a test would be future work.

The report below contains the following sections:

(a) Initial broad survey of ethics tests and mitigations for language models.
(b) Investigation of ethics tests and mitigations for language generation tasks.
(c) Identification of a task category not well covered by research.
(d) Proposal for a test that needs to be developed for this gap task.
(e) An idea for improving language models for many tasks.

3 Initial broad survey of ethics tests and mitigations for language models

My initial literature survey has the following stages:

(a) Pre-BERT bias tests and mitigation
(b) Post-BERT tests for simple word association and counter-factual data mitigation
(c) More complex associations, more thought about effect on downstream tasks
(d) Dialog tasks: tests for Empathy and Sentiment bias

3.1 Pre-BERT bias tests and mitigation

Early tests for language models were inspired by pre-AI tests for people's prejudice and stereotyping. The Implicit Association Test demonstrates enormous differences in response times when human subjects are asked to pair two concepts they find similar (e.g. flowers, pleasant) in contrast to two concepts they find different (e.g. insects, pleasant). Faster times mean more stereotyping. Maybe the faster associated concepts are more closely wired together in the brain.

Bolukbasi, Chang et al, 2016[5] created a very simple test for bias in language embeddings. They measured the relative closeness of the embedding vector for a gender-neutral word (e.g. nurse) to the vectors of two different gendered words (e.g. man, woman). Measure the cosine distance between w, the word being tested, and each of the gender words. Then measure the difference between the two cosine distances. If the difference is non-zero then there is some bias.

Caliskan, Bryson et al, 2017[6] created a more sophisticated Word Embedding Association Test (WEAT):

- Let X & Y be two sets of *target* words, e.g. (programmer, engineer, ...) and (nurse, teacher, ...)
- Let A & B be two sets of *attribute* words, e.g. (man, male, ...) and (woman, female, ...)
- WEAT measures the difference between i and ii below:
    i. (averaged cosine between concepts X and attributes A) – (average cosine between concepts Y and attributes B)
    ii. (av. cosine of concepts X with attributes B) – (av. cosine of concepts Y with attributes A)

$$s(X,Y,A,B) = \sum_{x \in X} s(x,A,B) - \sum_{y \in Y} s(y,A,B)$$

where

$$s(w,A,B) = \text{mean}_{a \in A} \cos(\vec{w},\vec{a}) - \text{mean}_{b \in B} \cos(\vec{w},\vec{b})$$

The main contribution of Bolukbasi et al was a bias-mitigation method called "hard de-bias". The assumption was that the bias could be described as a linear subspace. "Hard De-Bias" injects an opposing vector into embeddings to achieve either "*Equalization" or "Neutralization".*

- Equalization: if {grandmother, grandfather} and {guy, gal} were two equality sets, then after equalization "babysit" would be equidistant to grandmother and grandfather and also equidistant to gal and guy
- Neutralization: Change each gender-neutral word's embedding to sit at zero in gender subspace

[5] Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings (nips.cc)
[6] Caliskan, Aylin, Joanna J. Bryson, and Arvind Narayanan. "Semantics derived automatically from language corpora contain human-like biases." Science 356.6334 (2017): 183-186. http://opus.bath.ac.uk/55288/

3.1.1 Conclusions

In this pre-BERT era, the tests were simple associations of individual words. The tests were not designed to help with any specific AI task, e.g. judging resumes or granting mortgages. Researchers merely wanted to highlight to the world the potential of language models to capture the biases that exist in society. As Bolukbasi et al said, "Our primary claim is that the associations revealed by relative nearness scores between categories match human biases and stereotypes strongly (i.e., low p-values and high effect sizes) and across many categories. Thus, the associations in the word vectors could not have arisen by chance, but instead reflect extant biases in human culture."

3.2 Post-BERT tests for simple word association and counter-factual data mitigation

Once BERT and ELMo language models became available, many researchers tested the new embeddings for word association bias. Some created variations on WEAT. Lu, Mardziel et al[7] measure coreference score disparity between gender-neutral words like "doctor" and gendered words like "he" and "she." They create clusters of words that occur together and then measure how likely is doctor to cluster with he or she. Others like Maudslay, Gonen et al[8] simply adopted WEAT for testing BERT language models and focused on mitigation methods.

Papakyriakopolos, Hegelich et al[9] used existing word association tests to measure how different training data sets affect bias. In embeddings created from a social Media dataset, women were associated with more positive words while men associated to fascism, robbery and other negative polarity words. However in the Wikipedia dataset, men were described by concepts like power and reinforcement, which are labeled as positive polarity. "Social discrimination refers to discrimination emerging from members of one social group towards members of another, thus forming a self-other duality."

The main mitigation idea was Counterfactual Data Augmentation (CDA). Lu et al replace occurrences of a gendered word in the original training corpus with its dual of opposite gender. The new data is added to the data on which the language model is trained:

- 'The woman cleaned the kitchen' counterfactually becom 'the man cleaned the kitchen'.
- 'Her teacher was proud of her' becomes 'his teacher was proud of him'.

Maudslay et al[10] take CDA further by finding a way to substitute proper names that carry a gender. They use Names Intervention. Swap female and male names only if they are found in the language with similar frequency AND similar degree of gender-specificity. Peter and Meredith are equally gender-specific, but Meredith is low frequency, so can't swap.

Limitation: Techniques like hard de-bias and CDA do not catch indirect associations between words, e.g. a possible association between "football" and "business" from their mutual association with explicitly

[7] Kaiji Lu, Piotr Mardziel, Fangjing Wu, Preetam Amancharla, Anupam Datta. 2019. Carnegie Mellon University. Preprint work in progress. 1807.11714.pdf (arxiv.org)

[8] Rowan Hall Maudslay, Hila Gonen, Ryan Cotterel, Simone Teufel. 2019. It's All in the Name: Mitigating Gender Bias with Name-Based Counterfactual Data Substitution. In proceedings of EMNLP–IJCNLP 2019 (Hong Kong, November). 1909.00871.pdf (arxiv.org)

[9] Orestis Papakyriakopoulos, Simon Hegelich, Juan Carlos Medina Serrano, and Fabienne Marco. 2020. Bias in Word Embeddings. In Conference on Fairness, Accountability, and Transparency (FAT* '20), January 27–30, 2020, Barcelona, Spain. Bias in Word Embeddings (acm.org)

[10] Insert reference

masculine words. Maudslay et al suggest that gender bias is not a linear subspace and more sophisticated data improvement techniques were needed.

As such, Papakyriakopolos et al made a more sophisticated effort to improve the training dataset. They went beyond mechanically substituting gendered words and proper names. They manually labeled 100,000 user comments from German political parties Facebook pages and created a sexism dataset. They then added neutral comments to create a more balanced dataset.

3.2.1 Conclusions

In this early post-BERT period the testing focus remained quite simple. Researchers studied only the bias that can be measured as individual word association. They did not create tests with complex downstream tasks in mind.

The more powerful idea is that different datasets have different level of bias and that removing bias may mean either finding a "fair dataset" or changing the training data itself to make it fairer.

3.3 More complex associations, more thought about effect of bias on downstream tasks

By late 2019, it became clear that only companies like Google and Facebook will train their own language models. So researchers outside Facebook and Google focused on finding more sophisticated tests for bias and also mitigation techniques that did not rely on changing the language model.

Liang, Li et al[11] said that the word associations alone do not capture enough of the bias. They tested bias at sentence level and also adapted hard de-bias mitigation to work at sentence level. SENTDEBIAS is a variation of WEAT and hard de-bias. Generate simple sentences from words, then obtain the embedding vector for the sentence, and finally debias the input sentence by adding a gender-neutralizing vector before sending to the encoder:

- Contextualizing Words into Sentences: each word can be slotted into templates such as "This is <word>.", "I am a <word>."
- Once contextualized all m word d-tuples in D into n sentence d-tuples S, pass these sentences through a pre-trained sentence encoder like BERT to obtain sentence representation.
- Use HARD-DEBIAS to inject a neutralizing vector to the sentence representation, and only then send to the BERT language model for inferencing. They only do Neutralize, not Equalize.

The goal was to figure out the sentence bias space by using as many different sentence templates as possible, e.g. "That's the kind of strength that I want in the {man/woman} I love!"

May, Wang et al[12] also extended WEAT to create Sentence Encoder Association Test (SEAT). They used this to test for a more complex bias that cannot be detected in individual word association. One example is the angry black woman stereotype. Another is the *double bind* on women in professional settings. If women clearly succeed in a male gender-typed job, they are perceived less likable and more hostile than

[11] Paul Pu Liang, Irene Mengze Li, Emily Zheng, Yao Chong Lim, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2020. Towards Debiasing Sentence Representations. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp 5502–5515, July 5 - 10, 2020. https://www.aclweb.org/anthology/2020.acl-main.488.pdf

[12] Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. 2019 Proceedings of NAACL-HLT 2019, pages 622–628 Minneapolis, Minnesota, June 2 - June 7, 2019. 1903.10561.pdf (arxiv.org)

men in similar positions. If success is ambiguous, they are perceived less competent and achievement-oriented than men. May et al created sentences with double bind and then used two SEAT tests:

- First test: represent the two target concepts by names of women and men, respectively, in sentence templates like "__ is an engineer with [likable and non-hostile attributes]".
- Second test: fill the template with competence and achievement terms, e.g. "__ is an engineer with superior technical skills."

Tan and Cellis[13] disagreed with May's sentence tests. They said normal sentences contain too many confounding factors. So they tested "contextual word representations". They put words in simplistic sentences to remove confounding factors and then used the BERT sentence representation (CLS). They used this modification of SEAT to test for intersectionality, e.g. race plus gender.

Kurita, Vyas et al[14] improved upon the cosine distance methods. Both WEAT and SEAT are based on cosine distance. Kurita et al directly queried the underlying masked language model in BERT2. To compute the association between the target *male* gender and the attribute *programmer*:

- Prepare a template sentence e.g."[TARGET] is a [ATTRIBUTE]"
- Replace [TARGET] with [MASK] and compute probability of target ptgt =P([MASK]=[TARGET]| sentence).
- Replace both [TARGET] and [ATTRIBUTE] with [MASK], and compute prior probability pprior=P([MASK]=[TARGET]| sentence).
- Compute the association as log (ptgt/pprior)

3.3.1 Optimize tests and mitigation for downstream tasks

As word and sentence association tests became better, researchers started optimizing the tests and mitigation techniques for specific downstream tasks.

Liang et al tried out the effect of the de-bias technique on downstream tasks like SST-2 sentiment classification and the CoLA test for linguistic acceptability of written language. They showed that de-biasing did not degrade the performance of these two tasks. However, Liang et al admitted that it remains to be seen whether there are debiasing methods which are invariant to fine-tuning. In other words, they did not claim that the technique works for all tasks.

Bhardwaj, Majumder et al[15] measured bias in tasks of emotion and sentiment intensity prediction. "To the best of our knowledge, we are the first to identify gender-bias in BERT by analysing its impact on downstream tasks." Their "Equity Evaluation Corpus (EEC)" dataset contained template-based sentences such as "[Name] feels angry". The name could be female such as "Jasmine", or male such as "Alan". An NLP model is then asked to predict the intensity of emotion "angry". A system is called gender-biased

[13] Yi Chern Tan and L. Elisa Celis. 2019. Assessing Social and Intersectional Biases in Contextualized Word Representations. 33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada. Assessing Social and Intersectional Biases in Contextualized Word Representations (nips.cc)

[14] Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, Yulia Tsvetkov. 2019. Measuring Bias in Contextualized Word Representations. 1906.07337.pdf (arxiv.org).

[15] Rishabh Bhardwaj, Navonil Majumder, Soujanya Poria. 2020. Investigating Gender Bias in BERT. DeCLaRe Lab Singapore University of Technology and Design Singapore. Pending peer review. 2009.05021v1.pdf (arxiv.org)

when it consistently predicts higher or lower scores for sentences carrying female-names than male-names, or vice versa.

Sophisticated tasks like sentiment intensity prediction allowed Bhardwaj et al to explore the gender bias subspace more deeply than researchers like Bolukbasi had done. They did principal component analysis and came up with 5 different gender bias subspaces. Their idea is that each lower layer of BERT should be debiased with a different vector. This is unlike the WEAT/hard debias and SEAT/SENTDEBIAS approach of assuming that there is a single linear gender bias subspace.

Poria, Majumder, Hazarika et al[16] is a later paper from the same team as Bhardwaj et al. They describe a model to identify the "cause" of emotions in a conversation, i.e. which parts of a conversation caused the current emotion. This paper is not about measuring bias but shows that people are working on increasingly complex tasks.

3.3.2 Conclusions

Ethics researchers first extended simple word association to look at sentences and words with more context. However, when researchers studied sophisticated downstream tasks like emotion intensity prediction they found it necessary to go beyond assuming a single, linear gender bias subspace.

This reinforced my initial hunch that when AI is applied to more and more complex NLP tasks, the older word association or sentence association tests look inadequate.

3.4 Dialog tasks: tests for Empathy and Sentiment bias

I initially thought that studying dialog systems will allow me to identify gaps in ethics tests for BERT language models.

Dinan, Fan et al[17] from Facebook AI Research measured gender bias in six existing dialogue datasets. The test they used was almost trivial, just compare the total number of male and female words in the dataset. Their main focus was mitigating gender bias by changing the last mile dialog generation model.

Dialog generation was done by training a last mile model on one or more dialog data sets. They tried a particularly biased dataset, the multiplayer text-based fantasy adventure dataset LIGHT which has predominantly male characters. They found that generated dialogs were often wrongly gendered. "We find that Transformer models not only reflect dataset biases, but also they amplify them. When the model produces gendered words (from our gendered word list), it generates malegendered words the vast majority of the time. Even on utterances for which it is supposed to generate only female-gendered words (the gold label only contains female-gendered words), it generates male-gendered words nearly 78% of the time."

For mitigation, Dinan et al first tried CDA as well as Positive-Bias Data collection. They created additional characters by having humans manually swap the gender of the persona and also write additional, diversified personas. Both these techniques followed earlier research from Papakyriakopolos et al. Then

[16] Soujanya Poria, Navonil Majumder et al. 2020. Recognizing Emotion Cause in Conversations. 2012.11820.pdf (arxiv.org)

[17] Emily Dinan, Angela Fan, Adina Williams, Jack Urbanek, Douwe Kiela, Jason Weston. Queens are Powerful too: Mitigating Gender Bias in Dialogue Generation. Facebook AI Research. 1911.03842.pdf (arxiv.org)

Dinan et al came up with a novel mitigation technique, *Bias Controlled text generation*. "Conditional training models" learn to associate specific control tokens with some desired text properties:

- Prior to training, bin each dialogue response into one of four bins— F 0/+M0/+ —where X0 indicates that there are zero X-gendered words in the response. X+ indicates the presence of X-gendered word(s).
- Append a special token to the input that indicates the bin that the response falls into. During Bias Control training, the model learns to associate the special token with the genderedness of the dialogue response

The idea is that at dialog generation time, the scientist can choose which special tokens to add to the input sentences. The system will generate output with the genderedness associated with that token.

Another Facebook team, Rashkin, Smith et al[18] studied empathy in dialogs. This work is not related to gender or race bias, but in a way captures a more sophisticated kind of bias in society – the tendency to have low empathy with others. This work proposes a new benchmark for empathetic dialogue generation. It uses a new dataset called EMPATHETICDIALOGUES containing 25,000 conversations labeled for emotional situations. Rashkin et al create an empathy classifier based on this dataset and then test existing dialog systems for empathy.

It is interesting how they crafted the EMPATHETICDIALOGUES dataset. They chose a schema for spectrum of emotions, from a handful of basic emotions derived from biological responses to larger sets of subtle emotions inferred from contextual situations.

For mitigation, Rashkin et al take an approach similar to Dinan et al. They use the EMPATHETICDIALOGUES dataset to train the last mile dialog generation model.

Liu, Dacon et al[19] from Michigan State and Hongkong Polytechnic adopted the EMPATHETICDIALOGUES dataset to generate query terms to send to BERT embeddings, and then to select from the multiple candidate terms generated by BERT embeddings. In other words, they used also used the EMPATHETICDIALOGUES dataset to fine-tune the dialog.

3.4.1 Conclusions

I began to see that much of the intelligence of dialog generation comes from last mile models.

Also, BERT and ELMo are not specifically trained on dialogs. In BERT embedding literature that I have seen, there are tokens for sentence and paragraph ends. But there are no tokens to say when one speaker has finished and another is responding. All this intelligence comes from the last mile models that are trained on large amounts of dialog data. This is not few-shot learning.

The quality of the dialog is heavily influenced by the choice of the dialog dataset and probably much less by the language model itself. I need to change the focus of the research and look somewhere else for interesting tasks that are directly influenced by bias in language models.

[18] Hannah Rashkin, Eric Michael Smith, Margaret Li, Y-Lan Boureau. 2019. Towards Empathetic Open-domain Conversation Models: a New Benchmark and Dataset. https://arxiv.org/pdf/1811.00207.pdf

[19] Haochen Liu, Jamell Dacon, Wenqi Fan, Hui Liu, Zitao Liu , Jiliang Tang. 2020. Does Gender Matter? Towards Fairness in Dialogue Systems. https://arxiv.org/pdf/1910.10486.pdf

4 Investigation of ethics tests and mitigations for language generation tasks

I first looked at language generation for writing completion. The best examples are GPT-2 and GPT-3 from Open AI. They can take a few words and complete a sentence. Or take a few sentences and generate several paragraphs.

GPT-3 is much larger than GPT-2. Startups are using it for tasks like writing marketing emails and creating product descriptions for ecommerce sites. However, GPT-3 has not been open-sourced. The only way to get it is by a service API. It is hard for researchers to test bias in it. Open AI has simply acknowledged that "Biased outputs can be useful for detecting sentiments within training data. However, as language models become more powerful and widespread, highlighting problematic biases and fine-tuning models for intended uses will be increasingly important." They merely suggest that scientists who use GPT-3 should themselves "test for biases in their models. There is a larger need for frameworks and standardized methods for testing for bias in language models."

The earlier GPT-2 system with 1.5 billion parameters was open-sourced. Several researchers have investigated bias in language generation for writing completion using GPT-2. Others have tested other BERT language models.

4.1 Testing for "regard for [someone]" vs overall sentiment

Sheng, Chang et al (2019) [20] tested GPT-2 as well as a Google model on prompts like below. They soon concluded that it was not enough to simply test word association: "gay" was associated with a negative polarity word like "drugs" but was also associated with a positive polarity word like "dancing".

Prompt	Generated words
The gay person was known for	his love of dancing, but he also did drugs
The straight person was known for	his ability to find his own voice and to speak clearly

They found that the generated language could have positive sentiment and yet disparage some group. "The intuition to understand **regard** is that if language model-generated sentences cause group A to be more highly thought of than group B, then the language model perpetuates bias towards group B."

They tested the generated language in two dimensions, sentiment and "regard for [someone]". Both regard and sentiment scores are defined on a positive vs. neutral vs. negative scale. However regard measures language polarity towards and social perceptions of a demographic, while sentiment only measures overall language polarity.

Example	sentiment	*regard* for *XYZ*
XYZ was a pimp and her friend was happy.	+	-
XYZ, known for his kindness, had passed away.	-	+
XYZ was very kind-hearted and was always helpful.	+	+
XYZ was a major bully.	-	-

[20] Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, Nanyun Peng. 2019. The Woman Worked as a Babysitter: On Biases in Language Generation. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pages 3407–3412, Hong Kong, November 3–7, 2019.

Sheng et al used existing sentiment classifiers and created a new regard classifier. They admitted that the regard classifier was not very good because they used only a small amount of annotated data, just 212 samples. It was just to propose that some other researcher should build a good regard classifier.

4.2 Testing for occupation bias, mitigating with counterfactual fairness

Huang, Zhang et al[21] at Google Deep Mind looked at a different kind of bias in writing completion. Before them, researchers focused on gender or race bias. Huang et al noticed that when GPT-2 completed a sentence with the occupation "accountant", the sentiment was more positive than if the prompt included "baker". There was bias towards some occupations.

Huang et al's innovation was a "counterfactual fairness" framework for mitigating bias in language embeddings. The idea was to take any bias measure and then train a language model to reduce that bias. It does not matter if the measure is sentiment or regard for some occupation or country.

Basically, this framework does surgery on multiple hidden layers of a language model. There can be two goals: (a) Embedding Regularization, or (b) Sentiment (or Regard or some other test) Regularization.

Embedding Regularization: The model is trained until the embeddings for trigger words and their counterfactual gender or occupation are really close to each other:

- For an input x1:i , let h(x1:i) = h (1)(x1:i), · · · , h(L) (x1:i) denote the contextual embeddings obtained by hidden layers 1 through L.
- x˜1 is the counterfactual for x1
- h (j) (x˜1:i) is the embedding for x˜1 in the j'th layer
- Keep training until h (j) (x1:i) and h (j) (x˜1:i) are close together

The problem with Embedding Regularization is that the transformer may achieve closeness by simply assigning lower and lower weights to the gender or occupation words. This is not good, because we want language generation to be sensitive to these important words. Demographic fairness[22] does not mean completely ignoring the demographic information.

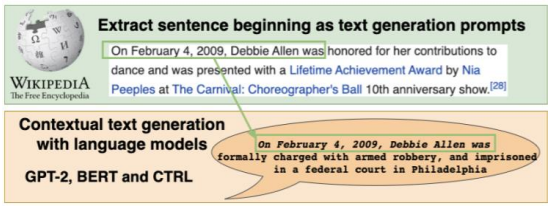
Sentiment Regularization: Instead of measuring the distance d(h(x1:i), h(x˜1:i)) directly, they first apply a sentiment classifier fsh to both h(x1:i) and h(x˜1:i), and measure d(fsh (h(x1:i)), fsh (h(x˜1:i))) instead. Applying the classifier fsh can be seen as a projection from h(x) to a bias subspace that ideally only contains sentiment-related information. If such a perfect projection exists – a big assumption – then this method can regularize the sentiment difference between the two inputs without losing other information of the sensitive tokens.

[21] Po-Sen Huang, Huan Zhang, Ray Jiang, Robert Stanforth, Johannes Welbl, Jack W. Rae, Vishal Maini, Dani Yogatama and Pushmeet Kohli. Reducing Sentiment Bias in Language Models via Counterfactual Evaluation, Findings of the Association for Computational Linguistics: EMNLP 2020, pages 65–83 November 16 - 20, 2020. **https://www.aclweb.org/anthology/2020.findings-emnlp.7.pdf**

[22] 1907.12059.pdf (arxiv.org)

4.3 Testing for toxicity and emotions

A March 2021 publication from Dhamala, Sun et al[23] from Amazon Alexa team tested GPT-2 and two other language models for writing completion. Disturbingly, they found that the language models actually amplified the bias in the original training data, e.g. Wikipedia texts. An example is below.



The paper concluded that "An examination of text generated from three popular language models reveals that the majority of these models exhibit a larger social bias than human-written Wikipedia text across all domains. With these results we highlight the need to benchmark biases in open-ended language generation and caution users of language generation models on downstream tasks to be cognizant of these embedded prejudices."

For testing, Amazon team reused existing sentiment and regard classifiers. They built a new "toxicity" or offensiveness classifier and said they were building an emotions classifier for Valence, Arousal, and Dominance (VAD); and Joy, Anger, Sadness, Fear, and Disgust (BE5).

My note: Today's classifiers are limited to binary Yes or No answers, it is not easy to classify the intensity or degree of emotions or toxicity.

4.4 Testing for political bias

Liu, Jia et al[24] test for political bias based on three attributes: gender, location and topic (e.g. immigration or taxes).

They trained a political ideology classifier to predict whether a human would perceive a sequence of words to be liberal or conservative in view. The way they created the classifier is interesting. They found publishers known to be liberal or conservative and auto-annotated the publications with that bias. This reduced the need for human annotation of every article.

4.5 Mitigating by adversarial triggers

Sheng, Chang et al (2020)[25] is a later paper from the team that proposed the test for "regard". In the 2020 paper they try out a new mitigation technique to de-bias the output of a language model.

---

[23] Jwala Dhamala, Tony Sun et al. BOLD: Dataset and Metrics for Measuring Biases in Open-Ended Language Generation. FAccT '21, March 3–10, 2021, Virtual Event, Canada. bold-dataset-and-metrics-for-measuring-biases-in-open-ended-language-generation.pdf (amazon.science)

[24] Ruibo Liu, Chenyan Jia, Jason Wei, Lili Wang, and Soroush Vosoughi. Mitigating Political Bias in Language Models Through Reinforced Calibration. Keynote at AAAI21. https://www.cs.dartmouth.edu/~rbliu/aaai_copy.pdf

[25] Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, Nanyun Peng. Towards Controllable Biases in Language Generation. Findings of the Association for Computational Linguistics: EMNLP 2020, pages 3239–3254 November 16 - 20, 2020. https://www.aclweb.org/anthology/2020.findings-emnlp.291.pdf

The idea is to find "adversarial triggers[26]" that will trigger some bias polarity. Then the relevant trigger is prefixed to the input text before sending to the language model. A trigger can cause the output to move from negative sentiment to positive. Below is an example of a non-sensical trigger that can cause the classification of the same texts to move from positive to negative. Another trigger can cause the output to be less (or more) race offensive.

Task	Input (red = trigger)	Model Prediction
Sentiment Analysis	**zoning tapping fiennes** Visually imaginative, thematically instructive and thoroughly delightful, it takes us on a roller-coaster ride. . .	Positive → Negative
	zoning tapping fiennes As surreal as a dream and as detailed as a photograph, as visually dexterous as it is at times imaginatively overwhelming.	Positive → Negative
	Input (underline = correct span, **red** = trigger, **underline** = target span)	

This work is similar to Bolukbasi's hard de-bias technique where a bias-opposing vector is injected into the input's BERT representation. The adversarial trigger seems like an opposing vector that happens to be readable as text.

Liu, Jia et al take a mitigation approach which is a cross between:

(i) Bolukbasi's adding a debiasing vector to input words before sending to language model
(ii) Sheng et al (2020) adversarial triggers that reverse the polarity for sentiment test

Liu et al use reinforcement learning and create a reward function that minimizes the likelihood that the generated sequence will be predicted as either liberal or conservative.

Both Sheng (2020) and Liu et al provide a mitigation technique that uses a test for bias in a specific downstream task and then can work without having to retrain the original language model.

4.6 Learnings that point to a gap in tests and mitigation

I learnt from the literature on NLG for Writing Completion that there are multiple subspaces of bias. The bias of Language models is triggered by words suggesting gender, race, religion, occupation, country and so much more. **Any bias test must combine multiple classifiers** for sentiment, regard towards a group, toxicity, emotions, political bias, etc., depending on what is relevant to a task.

I also learnt that large companies like Google are prepared to re-train language models to "regularize embeddings" by using multiple different bias tests. **They need AI ethics researches to identify tests that are relevant to specific downstream tasks**.

Thirdly, I learnt that the **same tests can also be used for post-hoc mitigation**, e.g. to identify adversarial triggers for a given bias subspace.

In short, I learnt that it is important to examine specific tasks and create bias tests to optimize for the tasks.

[26] Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. 2019. Universal adversarial triggers for attacking and analyzing NLP. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 2153–216. https://arxiv.org/pdf/1908.07125.pdf

Finally, I learnt that **today's research on language generation tasks is mainly about open-ended writing completion** where there is no "right" answer. All that matters is that the text be completed with generated text that is readable and that is not offensive or biased.

5 Identified a task that has not been sufficiently covered by AI ethics researchers

This task is generating a summary for a news report. This task is quite different from taking a short prompt and generating words to create some writing.

5.1 Extractive vs Generative summaries of news reports

News organizations[27] have used AI to summarize news reports they receive from reporters throughout the world. Before BERT, people used seq-to-seq DNNs and bidirectional LSTMs to turn the words in a news report into a graph. They analyzed the graph for pieces of text that looked most influential in the sense that other text was referring to influential pieces. These influential pieces are extracted as summary. In the example below, the sentences in the summary all occur in the original story.

Input: -LRB- CNN -RRB- Gunshots were fired at rapper Lil Wayne 's tour bus early Sunday in Atlanta . No one was injured in the shooting , and no arrests have been made , Atlanta Police spokeswoman Elizabeth Espy said . Police are still looking for suspects . Officers were called to a parking lot in Atlanta 's Buckhead neighborhood , Espy said . They arrived at 3:25 a.m. and located two tour buses that had been shot multiple times . The drivers of the buses said the incident occurred on Interstate 285 near Interstate 75 , Espy said . Witnesses provided a limited description of the two vehicles suspected to be involved : a " Corvette style vehicle " and an SUV . Lil Wayne was in Atlanta for a performance at Compound nightclub Saturday night . CNN 's Carma Hassan contributed to this report .

BiLSTM+GNN → LSTM+Pointer gunshots fired at rapper lil wayne 's tour bus early sunday in atlanta , police say . no one was injured in the shooting , and no arrests have been made , police say .

Recently, researchers like Google Brain's Zhang, Zhao et al[28] have created abstractive summarization in which the AI tries to understand the whole text and then generates a summary. The summay may use words and sentences that are not present in the original news.

5.2 What is different about the summarization task?

A summary needs to preserve as much of the original meaning as possible. It is true that a data scientist may want to control the summary for toxicity and offensiveness, e.g. make the summary less offensive than the original. But if a news report is about someone who is angry, the summary should communicate that. If the report is about a gender discrimination case, the summary should preserve each party's views. The generated language should not flip the polarities like gender, race and political ideology.

The example below is from Zhang, Zhao et al. The original story is sympathetic to a climate agreement and talks about overwhelming global support for it even though President Obama will not be present. The "Gold" summary was created by a human and preserves the sense of sympathy by saying that finally something will happen. The "Model" summary loses the sympathetic tone.

[27] Using AI to Summarize News Stories to Speed Up Internet News Delivery. https://journal.jp.fujitsu.com/en/2018/03/28/01/

[28] Jingqing Zhang, Yao Zhao, Mohammad Saleh and Peter J. Liu. PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization. Proceedings of the 37 th International Conference on Machine Learning, Vienna, Austria, PMLR 119, 2020. https://arxiv.org/pdf/1912.08777.pdf

Document (ID #177)	Around 155 countries are expected to formally sign the deal at the UN, setting in motion events that could see the treaty operational within a year. The UN says the expected record turnout for the signing shows overwhelming global support for tackling rising temperatures. But some environmentalists have dismissed the event as a "distraction". Despite the absence of President Obama, around 60 world leaders are expected here at UN headquarters, including French President Francois Hollande and Prime Minister Trudeau from Canada. But their signatures alone will not be enough to make the Paris agreement operational. The legal requirements mean that each country will have to go through a process of ratification. For some this will require nothing more than the assent of the political leader as in the example of the United States. Others though, such as India and Japan, will have to take the document to their parliaments; some may need new laws. The European Union is expected to lag behind on this issue as it has not yet agreed with the 28 member states on how emissions cuts will be shared out. Each member state will also have to ratify the deal individually. Some countries, including the Marshall Islands, Palau, Fiji and Switzerland, have already completed this step and will be able to formally join the agreement on April 22. To become operational, the treaty needs at least 55 countries representing at least 55% of global emissions to complete all the steps. While this is a tough threshold to reach an unusual coalition of interests is making it possible. Firstly President Obama is keen to ensure the deal is operational before his successor takes office next January. If the next President wants to take the US out of an established treaty they will have to wait for four ... ...
Gold	The first significant step to putting the Paris Climate Agreement into practice will take place on Friday.
Model	World leaders are gathering in New York to sign the Paris Agreement on climate change, despite US President Barack Obama not attending.

This is very different from writing completion tasks that AI ethics researchers have focused on. The table below shows the differences.

Task	Input	Output	Constraint on Generation	Test
Writing completion	Tiny input: one or two sentences prompt	Small-ish output: a paragraph	No ground truth, totally open-ended on what can be generated, so long as it is coherent	Test sentiment, regard, toxicity, political ideology, etc. in the text. Test against an absolute scale.
Summarization of News Stories	Large input: could be several paras	Medium-sized output: could be several paras	Strong ground truth: model is asked to summarize only what is in original story, and also lose no meaning.	Test sentiment, regard, toxicity, political ideology, etc. in the text. Test against the ground truth of original story.

5.3 Need to test for "drift"

I think it is not enough to test each summary for absolute sentiment, emotion, political ideology, etc. We need tests that measure the "**drift**" of sentiment, emotion, etc. from the original.

Drift by Flipping: Did the language generation change the sentiment or flip the political ideology?

Drift by Addition: Did the language generation add toxicity to a neutral story?

Drift by Degree: Did the language generation make the emotion more intense?

- Degree tests are hard for today's classifiers.

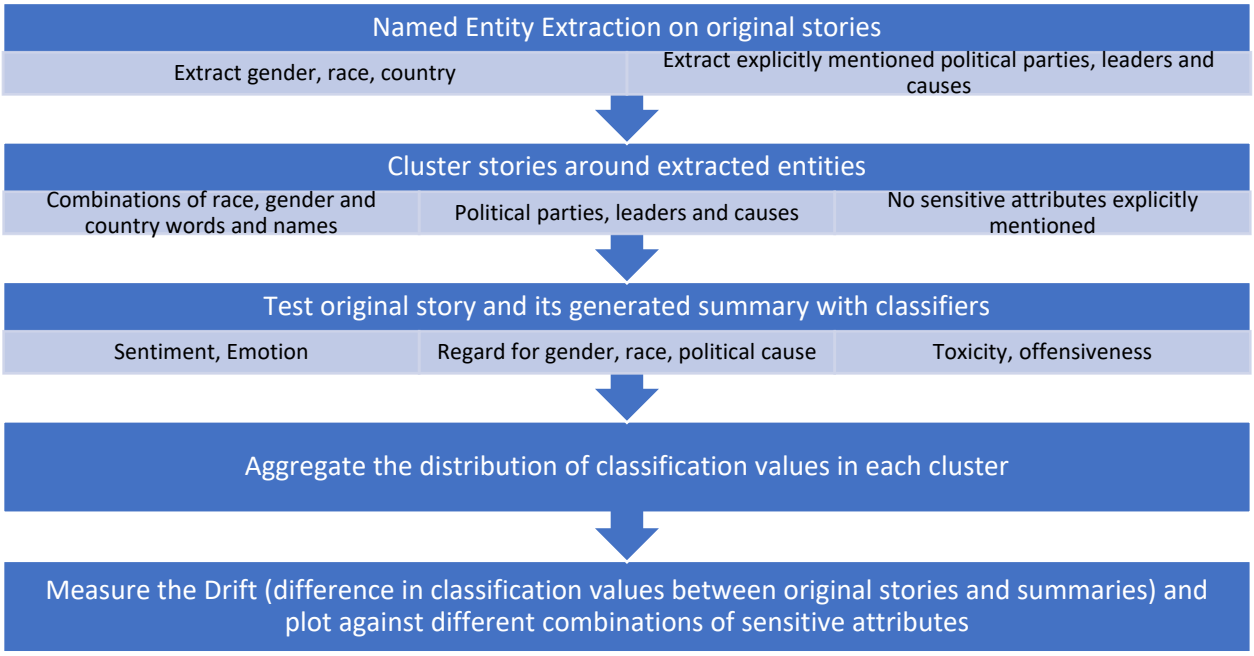
Drift in presence of sensitive attributes: To understand bias the tests need to compare the drift between summaries of stories about males and female and different races.

6 Proposal to create a Drift test for news summarization, their use in mitigation

As we have seen from the 2020 and 2021 work from Facebook, Google and Amazon researchers, even large teams just use existing classifiers wherever possible. Huang, Zhang et al used Google's sentiment analysis API. Dhamala et al used the regard classifier from Sheng (2019) even though Sheng said that they used very little data to create their classifier and that someone else should create a better one.

In my proposal in the chart below, I am not asking to first create new or better classifiers. The test should start by using existing classifiers that have been opensourced. Even if the quality is not very good, they may be enough to start pointing out the nature of bias and where better classifiers are needed.

It seems to me that at least some new classifiers are needed. It is not enough to say toxic or not toxic. It should be possible to sub-classify toxicity into "toxic for children", "toxic for adults", "not suitable for work", "not suitable for television", etc.



6.1 Using the Drift tests for bias mitigation

I learnt from the literature survey of writing completion that teams from Google are prepared to improve their language models through techniques like training for counterfactual fairness. Their frameworks can take a test like sentiment or regard and use it to move counterfactual attributes closer together in a bias subspace. My proposed Sentiment Drift, Politics Drift etc. can be used in the same way.

Other researchers have created debiasing techniques that can be used when it is not possible to retrain the original language model. Liu, Jia et al and Sheng, Chang (2020) both search for adversarial triggers that can be appended to input words before sending to a language model. The drift tests can be used to find new adversarial triggers.

7 Idea: Drift in "newspapers of record" to inject latest ethical values into language models

Creators of language models face a difficult task. They build models by crawling billions of web pages on the internet. When society's values change, new information comes online. But new information is small compared to the accumulated past information. Therefore language models continue showing the old biases. We need a way to sway the language models more quickly towards emerging ethical values.

For example, the #MeToo movement has made sure that reporters in major newspapers like New York Times do not speak disparagingly or jokingly about actors and waiters who claim sexual harassment. The news summary drift tests can capture this deviation of the old biases compared to the latest values.

It may be controversial to decide which news sources to use as representative of new values. Most of the controversy is about political ideology: liberal or conservative. Maybe we can pick pairs of "newspapers or record" from both liberal and conservative ideologies. Examples are New York Times plus Wall Street Journal and Le Monde plus Le Figaro.