Quiz 4 CSCE 771 at USC

Andrew Smith

8 December 2022

0 Instructions

1. This quiz is a mix of paper reading and running working code to understand the concepts underlying the paper
2. Code has to be submitted in a directory of your GitHub called "Quiz4" with sub-dir for code, data and doc. Code will have your source code, data will have any input or output generated, and doc will have a .pdf of this file (called Quiz4-CSCE771-answers.pdf) along with any answers
3. Complete quiz by 1:00 pm on Thursday, Dec 8, 2022. Hand over printout in class / in-person or send pdf as an email to biplav.s@sc.edu confirming completing the quiz and attaching your Quiz4-CSCE771-answers.pdf.
4. Total points = 30 + 60 + 10 = 100
5. Obtained =

Student Name: **Andrew Smith**

---

1 Objective

The objective of the Quiz is to learn bias issues with the usage of large language models on NLP tasks with hands-on experience.

2 Read paper

[10 + 10 + 10 = 30] Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings, Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, Adam T. Kalai, Neurips 2016

2.1 a) According to you, why does this behavior of word embeddings happen when running analogies?

Without delving into the details of creating the Word2Vec model, distance between two vectors in the embedding space is determined by "co-occurrence in text corpora". The theory underlying Word2Vec is necessary and sufficient. If two words co-occur often, then the word vectors are close to one another. If two word vectors are close to one another, then the words represented by the vectors co-occur often. The question is, at what level of abstraction does gender bias enter the system. Is is a product of the necessary and sufficient assumption of the Word2Vec model? It might be, but probably not, assuming that words that co-occur are related to one another (this is an assumption). Another grave assumption is the one of vector addition being related to relationships between words. The authors have cited that this has been "proven", but we must be careful when basing an entire scientific investigation on such an assumption. The authors even admit that vector representation of words is suprisingly simple for it to work. Does the bias arise from the assumption of vector multiplication in the embedding space? Further, the authors note in their final two sentences (which is far too late for my liking, and should have been the first two sentences of the paper) that the bias may a) reflect bias in society, or b) may capture useful statistics (these are not mutually exclusive claims). The authors have done nothing to refute these claims other than assume it is better to remove bias than to keep it. This is a grave mistake in my estimation. Fundamentally, the author's mitigation approaches to words that are assumed to be gender neutral *lose information*. Suppose I want to ask Word2Vec whether the majority of programmers are men or women. In the case that the supposed gender bias has been properly mitigated, the model would not know the answer to this question, though a simple Google search will reveal that 70% of programmers are men. The bias of word embeddings when running analogies comes from societal bias. Some of the societal bias is good (that is statistical bias) and some of the societal bias is bad (that is stereotypes). The authors need to address that there are statistical biases that are useful and not claim that all bias is bad.

2.2 b) What approach is proposed in the paper to mitigate it?

The authors propose neutralization and equalization. Neutralization *zeros* gender neutral words in the gender subspace. Equalization forces the distances between sets of words outside the the gender subspace. Equalize fundamentally loses information related to polysemy.

2.3 c) Do you think the approaches actually mitigate? Any problems you anticipate in practice?

Of course the approach *actually* mitigates the gender bias. The authors have clearly, empirically defined a *gender subspace* and produce methods that provably equalize (or neutralize) that bias. In general, the problems that will arise is that Word2Vec is no longer a valid embedding technique for the language. It is a *gender netural* word embedding, that is a

word embedding that has a purpose, but does not correspond to the real word. In fact, romance languages in general have gender woven into the fabric of the language. Objects in romance languages are fundamentally masculine or feminine and are denoted as such by adjusting the preceeding article. Languages are *fundamentally* gendered.

3 Create and run notebook

[10 + 10 + 20 + 20 = 60] Create your own copy python notebook and execute from the tutorial python notebook. Activities to do are: a. load word embedding b. do visualization c. run analogies: examples and your own (at least 3) d. run one mitigation method Link to your completed notebook is at: this jupyter notebook in my github repo. **The notebook is completed and has loaded the word embedding, visualized the embedding, run many analogies (the examples and my own, and run one mitigation method (neutralization).**

4 Apply to your project

[3 + 4 + 3 = 10] Now consider your course project.

4.1 a. Does gender bias in word embedding is relevant to your project? How?

Gender bias in word embedding does not directly relate to my project. The project takes audio files, converts them to spectrograms, and processes them through a neural network to convert the audio to text. I will describe in section (c) how this relates to gender generally and approaches that can be used to mitigate unfairness.

4.2 b. If yes, what strategy will you use to improve to the situation?

Word embeddings aren't used directly in my course project; however, I will describe in the next section how to ensure fairness related to gender.

4.3 c. If no, what ethical issue(s) do you anticipate for your project and the strategy one may use?

It has been observed that my project uses audio files and their spectrograms to convert speech to text at *something like the phoneme level*. This has been observed by looking at target prediction pairs and noting that the model produces sentences that would *sound* like the target sentences, but the words are not necessarily proper english. Since the model fundamentally converts phonemes into text, bias may arise there. The question is, what does

the underlying statistical distribution of phonemes look like across gender? The distributions may be identical and identically distribution; however, *this is the least likely scenario*. How do we ensure that the speech to text model is *fair* with reference to gender. If we can measure the underlying distribution of phonemes related to gender in a few different ways that seem fair to the community (empirically, crowd-sourced, etc), then we can ensure that the output of the model *aligns* with the agreed upon underlying distribution. The bias would not present itself in the same way as word-embeddings with analogies and such; however, the bias may present itself in the following manner. Suppose we have all genders utter the same word. We would expect the model to produce the same output for all genders. If the model does not, it is biased for some reason. We need to also ensure that the training data is i.i.d as described previously. If we ensure that the training data is i.i.d and observe that the output of the model is biased (more than the underlying statistical distribution), then there is an issue and we need to develop mitigation techniques. One technique may be developing separate models for each gender and forcing the output of the models to have the same distribution across phonemes.