School of Informatics

Informatics Project Proposal
How the Geometry of Embedding Spaces Affects Bias in NLP Systems

B154805
April 2020

Abstract

When we build intelligent systems that can understand and generate natural language using real-life data, we also allow the system to acquire cultural associations which can be unfair. Many new methods have been developed to tackle this problem of bias and unfairness. What is lacking is the evaluation of whether modification of word embeddings is sufficient to improve fairness in downstream applications. This project aims to evaluate this by correlating changes in word embedding space to bias in standard downstream NLP applications.

Date: Sunday 26th April, 2020

Tutor: Resul Tugay

Supervisor: Seraphina Goldfarb-Tarrant

1 Motivation

With the recent rise in the use of machine learning (ML) for a wide variety of tasks, unfairness in these models is a growing concern. Most datasets used to train ML systems are generated from human data. Human biases and stereotypes from the data can easily result in unfair systems due to skewed statistics. Further concerns may arise as artificial intelligence (AI) is given agency in our society.

We see a lot of examples of errors in applications due to the presence of bias. Machine translation systems often change the gender of female pronouns while translating. Cameras that use computer vision to identify when a person's eyes are closed often mistake Asian people's eyes as always closed. Such examples cause irritation and inconvenience but there are applications that can be dangerous. Face recognition systems used by the police in the USA is racially biased. It works well for white faces but performs poorly on black faces [1]. Recently, a system was created to aid jury members in making death sentence judgements. The system uses characteristics of a person to decide whether they are guilty or not [2]. Natural language processing (NLP) systems also show a lot of bias. Automatic CV parsing systems make decisions based on gender due to past gender statistics. [3].

This has led to a flurry of research on recognising and reducing these biases, so as to prevent systems from performing better for some users than for others. Word embeddings are vector representations of words in a language [4]. The geometry of the embedding space represents the relationship between words. While word embeddings are indispensable in NLP, they also propagate stereotypes and biases to real-world systems [5]. The majority of approaches to measuring bias do so via examining the spatial relationships between word embeddings. For example, some characterise gender as a direction in the word embedding vector space [6]. These spatial relationships have been shown to reflect and to be predictive of many human biases.

There is considerable research on mitigating bias in NLP systems by altering word embeddings using various techniques. While such debiasing is effective for target criteria that measure implicit bias like the Word Embedding Association Test (WEAT) [7] or gender as a dimension [6], Gonen et al. say that the geometry of the bias is transformed but as such the bias still remains [8]. Bias information is still reflected in the clustering pattern of words in the debiased embeddings and can be recovered from them despite their change with respect to the target criteria. This indicates that most words that represented similar bias still group together in the new debiased vector space. The geometry of the word embeddings remains largely the same except for the changes with respect to these words.

It is currently unknown how much the differences in spatial relationships really affect practical downstream NLP applications. There is existing work which shows that fine-tuning word embedding spaces can affect downstream tasks. However, there is no systematic study of how the different notions of bias impact diverse downstream tasks.

1.1 Problem Statement

It has been shown that word embeddings generated from real-world text corpora contain human stereotypes and bias. To counter this, researchers have been working on modifying word embeddings to get rid of the bias. Debiasing techniques at the moment modify the geometry of the vector space such that the magnitude of the distance between two clusters reduces. In some sense, the spatial differences between clusters in the vector space represent the intensity of bias. These clusters represent groups of words having some attribute in common (eg: cluster of words deemed feminine - pretty, elegant, delicate, soft, cute). Such methods of measuring bias are called intrinsic bias measures. We will use of one such well known intrinsic measure, WEAT [7].

An NLP text classification task may exhibit bias by falsely classifying a disproportionate number of samples in the data due to the presence of particular words. For example, "muslim" may unduly be classified as "violent", and "gay" as "hateful". A definition based on measurement of such errors by a classification task is termed as *extrinsic bias measure.* We will use two well-known extrinsic measures, Equality of opportunity [9] and pinned AUC metric [10].

Intrinsic metrics measure implicit bias present in word embeddings while extrinsic metrics measure the presence of bias in the output of a classifier model. Most work on debiasing NLP systems tries to reduce bias with respect to intrinsic bias measures. However, what actually matters in practice is the latter because ultimately we want our classification model to be less biased. In this project, we explore the relationship between intrinsic bias and extrinsic bias. Our research tries to answer the following questions:

1. Do all spatial differences in the word embedding space matter for downstream application?
2. Does the magnitude of the distance between clusters of words relating to the bias attribute have a noticeable effect on the bias in a downstream application?
3. Is the relationship between intrinsic and extrinsic bias task dependent? Is it different for different tasks such as toxicology, co-reference resolution, etc?
4. Does the relationship between intrinsic and extrinsic bias differ by the type of embedding algorithm used?

This project will involve training different word embeddings and implementing a standard NLP classification task. It will involve examining the correlation between bias as measured geometrically in the original embedding, and as measured in the downstream task.

1.2 Research Hypothesis and Objectives

The main objective of this project is to study the effect of bias reduction in word embeddings on the bias present in a downstream NLP application. We hypothesise that bias in word embeddings is directly proportional to bias observed in our downstream task, but not linearly.

The study will not involve new debiasing techniques or comparison of language-specific bias reduction effects. It is mainly an evaluation study.

1.3 Timeliness and Novelty

Of late, the field of fairness in machine learning has become very important. With AI and NLP systems being used rampantly in the real world, it is high time that we ensure that our systems are fair and unbiased. We must ensure that these systems do not propagate and magnify bias and stereotypes. If such studies are not conducted, we could potentially be releasing new systems into the real world that seem to solve fairness issues superficially but are still inherently biased.

Ethics and fairness in AI are blowing up. Hence, a lot of research is going into debiasing systems. However, a thorough investigation of the effects of this debiasing has not been conducted. Our research will be the first study on the evaluation aspect of debiasing embeddings.

1.4 Significance

With NLP systems being used in all fields, it is of utmost importance that these systems are fair. They need to be modelled in a way that gives equal opportunity to everyone. They shouldn't be biased towards or against any set of people. This need is being sated by research into ethics and bias. However, the new techniques have not been studied in terms of their effect on downstream tasks. It is important to do this evaluation before deploying systems that use these techniques in case the bias has been removed only superficially in actuality.

We believe that our study will help further research in this area as it gives us some understanding of whether this approach is good enough and research should continue in this direction or there is need to develop new techniques that do more than transform the bias.

1.5 Feasibility

Our project pipeline is given below.

1. Collecting data: Existing datasets need to be identified as suitable to train the embedding algorithms.
2. Embedding algorithms: code base for GloVe and FastText already exist. We will need to fine-tune the code to fit our needs.
3. Downstream task: existing models for hate speech detection will be used.
4. Evaluation: Use WEAT and Equality of Opportunity to measure changes in bias in the downstream task

Thus, our baseline system should not be hard to implement. We can run tests on this to get baseline measures of bias in the embeddings and hate speech classifier.

Debiasing the embeddings will be our most time-consuming task. However, except for this added task, the pipeline remains the same. Therefore, we feel this project is feasible given the circumstances, time frame, and resources.

1.6 Beneficiaries

We believe that our work will be relevant and will benefit the NLP community whether we prove our hypothesis or not.

If we manage to prove that implicit bias removal in word embedding space is enough, more applications can employ the existing debiasing methods. If we prove otherwise, more research will need to be conducted to find techniques that effectively reduce bias in downstream applications. Either way, there will be some indication of how to move forward. If we get clear results we will, of course, present our work to the NLP community.

2 Background and Related Work

Word embeddings are used extensively in NLP applications. Semantically similar words tend to have vectors that are close together [4]. These vectors are built using co-occurrence counts.

The spatial differences between word vectors portray relationships between words. While there are different algorithms to generate these embeddings, GloVe is one of the most popular. GloVe [11] combines count-based and prediction-based models to create a global bi-linear regression model. Count-based techniques efficiently represent global word statistics and prediction-based techniques capture meaningful linear substructures. Unlike other embedding algorithms that use context windows and document level co-occurrence counts [4], Pennington et al. build a model using global co-occurrence counts [11]. The model also trains only on non-zero elements of the co-occurrence matrix. Interestingly, they use ratios of co-occurrence probabilities instead of just co-occurrence probabilities. They also learn word vectors using ratios of co-occurrences probabilities instead of just the probabilities of co-occurrence.

Another relatively new approach to creating word embeddings is making use of subword information. When using only word-level information, important characteristics like morphology are ignored. These characteristics could help in modelling low-frequency, and out of vocabulary words. Using subword information also makes word embedding generation more efficient as vectors for every word in the vocabulary are unnecessary. The work in [12] is modelled on skip-gram in which words are represented as groups of n-grams made up of the characters in the word. This means that each vector in the embedding space represents an n-gram and a word is the sum of all the n-grams of its constituent characters. Morphologically rich languages benefit greatly from such a representation.

Debiasing ML systems is becoming more relevant as they start being used in daily life. Research on detection and mitigation of bias in NLP systems has become important in recent times. We need metrics to detect bias and measure bias reduction. Caliskan et al. built a test for word embeddings based on the Implicit Association Test (IAT) [7]. They used the cosine similarity of pairs of word vectors as the difference between the two with respect to a set of words used by the IAT to recognise associations. The Word Embedding Association Test (WEAT) takes two target words (eg surgeon and nurse) and two attribute words say, male or female. Then they calculate the difference in the cosine similarity of the target words with the attribute words. This difference is a measure of bias. They found that word embeddings contain human-like biases and stereotypes with differences between demographic groups such as gender, race etc. This metric of bias measures implicit bias in word embeddings. Intrinsic bias metrics are agnostic to any downstream application of word embeddings. In practice, however, we need to measure and reduce the bias in applications that make use of these embeddings. Such measurements are made using extrinsic metrics. One such metric is the Equality of Odds, proposed in [9]. It defines unintended bias in a system as an imbalance in the false-positive and false-negative rate for samples containing different identity terms.

Using a related measure, Dixon et al's study on toxicology showed that bias gets amplified for certain words in the vocabulary [10]. Statements containing words (identity terms) that are often used negatively (eg: gay) got a very high toxicity score. Such bias is present due to a disproportionate number of toxic and non-toxic samples containing these identity terms in the training data. The model learned to associate these word with toxicity. In this paper, they also mitigate the bias using the Equality of odds measure which ensures that false-positive rates and false-negative rates are equal for samples that contain identity terms. They use the difference in these rates to measure the unintended bias. In another work, Zhao et al. studied gender bias in co-reference Resolution [13]. A system was said to be biased if it resolved pro-stereotypical samples (eg: he-surgeon, she-receptionist) more accurately than anti-stereotypical samples (eg: she-astronaut, he-nurse). They showed that there was bias in all commonly used co-reference resolution systems and provided a method for reducing the bias. They used dataset balancing to create a new dataset containing additional sentences with gender-swapped pronoun-occupation pairs.

3 Programme and Methodology

Work Package 1

First, we will need to set up and implement a baseline system. The results we get on running this system will be compared with the results of our debiased word embeddings system. We follow the pipeline mentioned in section 1.5.

1. **Collecting data:** The common corpora for training word embeddings are usually from Wikipedia, online magazines, and news articles. These datasets are used because they are convenient sources of a lot of data comprising standard English. Off-the-shelf models for hate speech detection do exist. However, not all of them make use of word embeddings. In case we need to build a classifier from scratch, we will need data for this too. The corpus for this task would contain text from social media such as Facebook, Twitter, and Reddit. These platforms are usually what people use to express their opinions and hence they are perfect for our task. This task will take us up to 1 week.
2. **Embedding algorithms:** We are going to use GloVe and FastText word embeddings for our project. We do this because we want to know if different embedding strategies model bias differently. Code bases for both already exist. FastText code is written in python and we will need to fine-tune the code to fit our needs. GloVe code is written in C++ but there should be python wrappers available. Again, we will need to fine-tune the code to fit our needs. Finding a suitable python wrapper might take more time than we anticipate so our worst-case duration for this task is 2 weeks.
3. **Downstream task:** Hate speech detection is a fairly popular application in NLP so there are existing models for hate speech detection. We will make use of one of these for our work. However, if the model we select does not have a provision for word embeddings we will need to add this and retrain the model. Training and fine-tuning the model to the best settings will take up to 1 week.
4. **Evaluation:** We will use WEAT and Equality of Opportunity to measure the bias present in the hate speech detection classifier. Familiarising ourselves with the two metrics and taking the actual measurements will take us 1 week.

Since some of our tasks are independent of each other we can perform them simultaneously. The time we have allotted to each task in this work package is highly exaggerated. However, if we do take 5 weeks to do this we will still have enough time to complete the rest of the project.

Work Package 2

Next, we need to set up our main experiment. This involves performing classification using debiased embeddings. The pipeline for this is given below.

1. **Debias embeddings:** We need to debias the word embeddings using some of the current debiasing techniques. We have decided to use dataset balancing and attract-repel. Dataset balancing involves either subsampling the dataset so that we have an equal number of

positive and negative samples, or it involves generating synthetic data so that there are equal positive and negative samples. Attract-repel, on the other hand, is language agnostic and more task-specific. We will select a WEAT attribute around which we want to change the bias. This task will take us 2 weeks because we will need to implement it from scratch for dataset balancing. We can fine-tune the existing code for attract-repel.

2. **Retrain embeddings:** Since we have already got suitable training algorithms for the word embeddings in our baseline system, we only need to feed it our new dataset. No other changes will be required. This task will take us <1 week.
3. **Downstream task:** We will use the same model we built in the baseline system but provide it with the debiased embeddings instead. This task will take us <1 week.
4. **Evaluation:** This task remains the same as the evaluation task in the baseline system. Since we will have understood the metrics by now, this task will take <1 week.

None of our tasks is independent of each other so a delay in any pushes the timeline. However, the time we have allotted to each task in this work package is highly exaggerated. If we do take 2-2.5 weeks to complete this we will still have enough time to complete the rest of the project.

Work Package 3

At this stage, we will have all the results we need. We will now need to compare them to measure the correlation between intrinsic bias and extrinsic bias. We will now perform the following analysis:

1. **Change in bias using GloVe:** We will compare the results of the baseline system that makes use of the GloVe algorithm with the results of the debiased system that makes use of the GloVe algorithm.
2. **Change in bias using FastText:** We will compare the results of the baseline system that makes use of the FastText algorithm with the results of the debiased system that makes use of the FastText algorithm.
3. **Analysis of bias change for GloVe vs FastText:** We will analyse the difference in results of task 1 and task 2 and understand why this happens.
4. **Bias change for different downstream tasks:** We will analyse the difference in results of task 3 for hate speech and results of task 3 for co-reference resolution. Another member of this project will be working on a co-reference resolution and thus we can use their results to compare against ours.

3.1 Risk Assessment

We can categorise our risks into 3 types.

The first type of risk is that of inconclusive results. As conclusive results would have led to useful insights into current bias reduction techniques in NLP applications, these significant advantages would be lost.

A more project-specific risk is that off-the-shelf models for hate speech detection may not have a provision for word embeddings. This could lead to a delay in our timeline as we would need to train a model to be able to use word embeddings. Another risk is the use of FastText. Though it is a very efficient algorithm, it may not behave as we want it to in terms of either capturing language characteristics due to it's use of subword information. The use of two different datasets for training the word embeddings and downstream task could result in correlations that require further analysis.

The final risk is that of failure to access resources. As we are now working remotely, we could lose access to the university servers and GPU clusters.

3.2 Ethics

Since we are making use of existing, openly available, and widely used corpora for our work, we do not need to get permission to use them. The topic of our project indeed contributes to ethical and fair use of machine learning techniques in society.

4 Evaluation

We require data that are natural language texts from sources such as Wikipedia, Reddit, news articles etc. These datasets are used because they are convenient sources of a lot of data comprising standard English. Well-known datasets exist for data from such sources. We will use one of these to train our embedding algorithms. The size of the corpus must be large enough to train embeddings, so a minimum of 10^7 tokens is required.

We plan to use off-the-shelf models for hate speech detection. However, in case we need to train a model from scratch, we will need data from sources like Twitter, Reddit etc. There are common datasets that are used for hate speech detection purposes and we would most likely use one of those. Our evaluation datasets will also be derived from the same sources as our training datasets.

The results of our study are the changes in amount of bias in the downstream task after debiasing the word embeddings measured using WEAT and Equality of Opportunity. We use these results to evaluate the correlation between intrinsic bias and extrinsic bias.

5 Expected Outcomes

We are hoping to gain some insight into how modification of the vector space to reduce bias in word embeddings affects bias in downstream applications. We predict that the bias in our downstream task will decrease with the decrease in bias in the word embeddings. However, this decrease in bias in the downstream task will not have a linear relationship with decrease in bias in the embeddings. We cannot say this with certainty, though.

Our research delves into what happens when the geometry of word embeddings changes and what these changes mean for applications that make use of these new word embeddings. It will help to understand if current methods of implicit bias reduction are sufficient for debiasing whole NLP systems.

6 Research Plan, Milestones and Deliverables

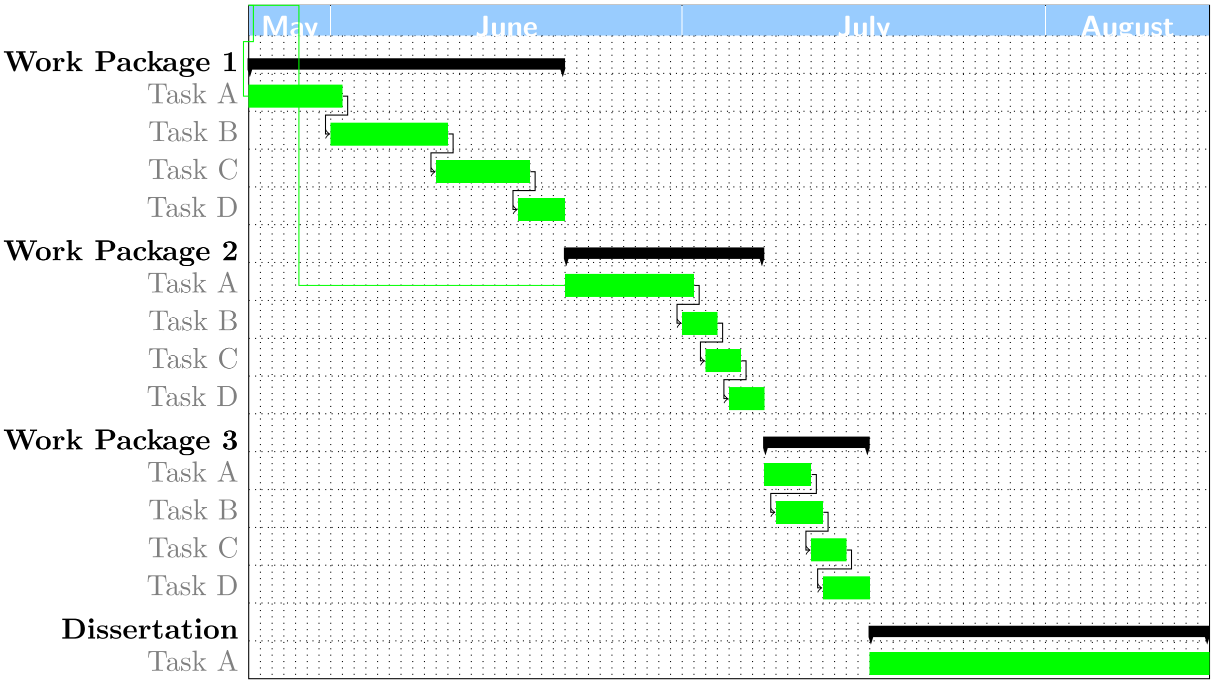

Figure 1: Gantt Chart of the activities defined for this project.

Milestone	Week	Description
M_1	4	Baseline system completed
M_2	6	Debiased system completed
M_3	7	Evaluation completed
M_4	10	Submission of dissertation

Table 1: Milestones defined in this project.

Deliverable	Week	Description
D_1	4	Baseline bias results
D_2	6	Debiased system results
D_3	7	Evaluation report
D_3	10	Dissertation

Table 2: List of deliverables defined in this project.

References

[1] Fabio Bacchini and Ludovica Lorusso. Race, again. how face recognition technology reinforces racial discrimination. *Journal of Information, Communication and Ethics in Society*, 2019.

[2] Yu Li, Tieke He, Ge Yan, Shu Zhang, and Hui Wang. Using case facts to predict penalty with deep learning. In *International Conference of Pioneering Computer Scientists, Engineers and Educators*, pages 610–617. Springer, 2019.

[3] Dena F Mujtaba and Nihar R Mahapatra. Ethical considerations in ai-based recruitment. In *2019 IEEE International Symposium on Technology and Society (ISTAS)*, pages 1–7. IEEE, 2019.

[4] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[5] Aylin Caliskan Islam, Joanna J Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora necessarily contain human biases. *CoRR, abs/1608.07187*, 2016.

[6] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in neural information processing systems*, pages 4349–4357, 2016.

[7] Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186, 2017.

[8] Hila Gonen and Yoav Goldberg. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. In *Proceedings of NAACL-HLT*, pages 609–614, 2019.

[9] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, pages 3315–3323, 2016.

[10] Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 67–73, 2018.

[11] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[12] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 2017.

[13] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Gender bias in coreference resolution: Evaluation and debiasing methods. *arXiv preprint arXiv:1804.06876*, 2018.