UNIVERSITY OF CALIFORNIA

Los Angeles

Building Accountable Natural Language Processing Models: on Social Bias Detection and Mitigation

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy in Computer Science

by

Jieyu Zhao

2021

ABSTRACT OF THE DISSERTATION

Building Accountable Natural Language Processing Models:
on Social Bias Detection and Mitigation

by

Jieyu Zhao
Doctor of Philosophy in Computer Science
University of California, Los Angeles, 2021
Professor Kai-Wei Chang, Chair

Natural Language Processing (NLP) plays an important role in many applications, including resume filtering, text analysis, and information retrieval. Despite the remarkable accuracy enabled by the advances of machine learning methods, recent studies show that these techniques also capture and generalize the societal biases in the data. For example, an automatic resume filtering system may unconsciously select candidates based on their gender and race due to implicit associations between applicant names and job titles, causing the societal disparity as indicated in [BCZ16]. Various laws and policies have been designed and created to ensure societal equality and diversity. However, there is a lack of such a mechanism to restrict machine learning models from making bias predictions in sensitive applications. My research goal is to analyze potential stereotypes exhibited in various machine learning models and to develop computational approaches to enhance fairness in a wide range of NLP applications. The broader impact of my research aligns well with the goal of fairness in machine learning – in recognizing the value of diversity and underrepresented groups.

The dissertation of Jieyu Zhao is approved.

Cho-Jui Hsieh

Yizhou Sun

Wei Wang

Kai-Wei Chang, Committee Chair

University of California, Los Angeles

2021

To my family.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

First and foremost, I'd like to express my greatest appreciate for my advisor Kai-Wei Chang, who has always been supportive. Before joining his group, I have zero background in NLP. I cannot remember how many times I have asked him stupid questions. But he is always patient and willing to answer. He shows me how to be an independent researcher from scratch and keeps encouraging me all the time. I cannot say enough how grateful and lucky I am to have Kai-Wei as my PhD advisor. He has taught me to be a critical thinker, and a responsible collaborator. From him, I have learned what a good advisor is like, not only in research but also in life. That has inspired me so much for my future career plan.

I would also thank all the professors in my committee, Prof. Cho-jui Hsieh, Prof. Yizhou Sun, and Prof. Wei Wang, for providing valuable feedback on my work. They have inspired me to look back to my research and and re-think from a bigger perspective.

I consider myself very lucky to have the opportunity to work with my amazing collaborators: Ahmed Hassan Awadallah, Ryan Cotterell, Saghar Hosseini, Daniel Khashabi, Tushar Khot, Subhabrata Mukherjee, Vicente Ordonez, Ashish Sabharwal, Tianlu Wang, William Yang Wang, Mark Yatskar and many more. Your help has made a remarkable difference to this thesis! Of course, my PhD life would be less colorful without my friends and labmates at UCLANLP: Wasi Ahamad, Kareem Ahmed, Muhao Chen, Junheng Hao, Tong He, Kuan-Hao Huang, Liunian Li, Qianru Li, Zeyu Li, Yitao Liang, Lu Lin, Zhongyu Lu, Tao Meng, Md. Rizwan Parvez, Yujia Shen, Bin Shi, Hua Wei, Borui Yang, Chi Zhang, Zhe Zeng, Honghua Zhang, Zhehui Zhang, Yichao Zhou and so on. In addition, I would like to thank Nanyun Peng for her help in both academia and real life. Thank you all!

Lastly, I want to say thank you to my family. I would not finish this PhD journey without the unconditional love, support and care from them. I am very thankful to have the accompany from my two cats during the pandemic time. In my heart, they are already part of my family. I love you all!

VITA

2009–2013	B.E. (Computer Science), Beihang University.
2013–2016	M.S. (Computer Science), Beihang University.
2016–2017	Ph.D. (Computer Science, withdrawal), University of Virginia.

PUBLICATIONS

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints. EMNLP. 2017.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. ”Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods.” NAACL (Short). 2018.

Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. Learning Gender-Neutral Word Embeddings. EMNLP. 2018.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. Gender Bias in Contextualized Word Embeddings. NAACL. 2019.

Jieyu Zhao, Subhabrata Mukherjee, Saghar Hosseini, Kai-Wei Chang, and Ahmed Hassan Awadallah. Gender Bias in Multilingual Embeddings and Cross-Lingual Transfer. ACL. 2020.

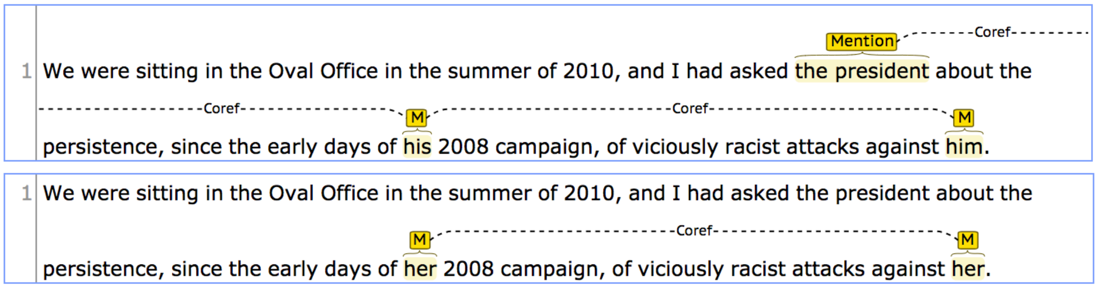
Jieyu Zhao, and Kai-Wei Chang. LOGAN: Local Group Bias Detection by Clustering. EMNLP. 2020.

Jieyu Zhao, Daniel Khashabi, Tushar Khot, Ashish Sabharwal, and Kai-Wei Chang. Ethical-Advice Taker: Do Language Models Understand Natural Language Interventions?. ACL Findings. 2021.

CHAPTER 1

Introduction

We are entering an era of Artificial Intelligence where we rely on machine learning models more than ever. Thanks to the development of both hardware and software, nowadays with affordable efforts, we can train a machine learning model to obtain an autonomous decision making agent that can largely benefits our daily life, such as an digital personal assistant. In a typical machine learning pipeline, given a set of *training data* (pairs of input and targeted output), an AI model learns underlying *representations* of input instances and conduct the *inference* to predict output labels based on the representations. Taking coreference resolution as an example, where the goal is to identify all phrases that refer to the same entity in a given text (e.g., in Fig. 1.1, both pronouns “his” and “him” refer to the same entity, “the president”). A coreference resolution system has to leverage syntactic and semantic information to cluster noun phrases into groups, such that the phrases in each group refer to the same entity. The state-of-the-art coreference system first converts each word in a document into a contextualized word vector. Then, based on the vectors, the model learns a representation of each noun span and leverages all information to make an inference in predicting the coreference clusters. Despite its wide usage in real word applications such as resume filtering or text summary, the model performs unequally for male and female entities. For example, in Fig. 1.1, when we change the gender pronoun to female ones, the model cannot predict the coreference link between “the president” and “her”. Similarly, the model will also fail to predict the link between female pronouns with some specific occupations such as “lawyer”, “doctor” and “leader”.



We were sitting in the Oval Office in the summer of 2010, and I had asked the president about the persistence, since the early days of his 2008 campaign, of viciously racist attacks against him.

We were sitting in the Oval Office in the summer of 2010, and I had asked the president about the persistence, since the early days of her 2008 campaign, of viciously racist attacks against her.

Figure 1.1: Illustration of gender bias in coreference resolution. The model can only discover the coreference link between "the president" and male pronouns ("him" and "his"). However, it fails to discover the link between "the president" and female pronouns ("her") even though it is provided with the same context.

Unfortunately, such biased behavior is not not specific to the coreference resolution system, but widely exists in various NLP models. In real world, humans have designed various policies and mechanisms to prevent discrimination on the basis of attributes such as gender or race when making sensitive decisions (e.g., hiring). Although there is no single effective way to prevent unconscious and implicit biases, great efforts have been made to promote diversity and fairness. However, for those machine learning models, they usually employ data-driven approaches, which learn to make decisions based on the statistics and diagnostic information from previously collected data. They thus risk causing the systems to potentially encourage unfair and discriminatory decision making. Such propagation of biases in NLP poses the danger of reinforcing damaging stereotypes in downstream applications. This has real-world consequences; for example, concerns have been raised about automatic resume filtering systems giving preference to male applicants when the only distinguishing factor is the applicants' gender.

My long-time research goal is to build accountable NLP models that are accessible to all people. During my PhD, I have been specifically focused on the social bias issue in NLP models. My research will benefit both the algorithm theory and the practice of machine learning and will especially help to reduce the potential social stereotypes. This thesis focuses on the following aspects:

- *Bias (amplification) Detection.* To deal with the bias issues in a model, the first and foremost step is to have a way to discover such biases. To do this, my research covers building essential evaluation datasets, proposing new evaluation metrics as well as understanding biases in multilingual scenarios. We analyze the bias in different NLP applications and reveal that all of them exhibit the stereotypes to some specific group.
- *Bias (amplification) Mitigation.* We propose various methods to mitigate the bias. Corresponding to the various source of biases, the proposed mitigation methods vary from modifying the training procedure to adding inference-phase-only constraints. The experimental results show that our methods can effectively reduce the bias without significant affect on the model's performance.

In Chapter 2, we review the bias issues in training dataset. We use the coreference resolution task as an example to demonstrate such an issue. And to promote the bias analysis research, we create a new dataset, WinoBias, for bias detection and evaluate different coreference models on it. Experimental results demonstrate the bias issues commonly exist in different systems.

In Chapter 3 we go through the biases in representations. It covers the biases in word embeddings, contextualized word embeddings as well as embeddings beyond English. In this chapter, we also discuss possible ways to mitigate those biases in both intrinsic and extrinsic levels.

In Chapter 4 we revisit the bias quantification metrics. In existing literature, group fairness is a widely used bias evaluation metric. In this chapter, we discuss the defect of such metrics and propose a new algorithm logan to discover a more fine-grained bias.

In Chapter 5, we consider the bias amplification problem in a vision-and-language task. We show that a machine learning model not only mimic the biases in the training dataset but further amplifies that. To reduce such bias amplification, we propose to add corpus level constraints in which way we do not need to retrain the model and the experimental results demonstrate that we can mitigate the bias amplification with a trivial affect on model performance.

In Chapter 6 we revisit the idea of building a highly self-adjustable machine. With the current advances in NLP models, we want to verify if those models with the ability to outperform humans can understand instructions and thus amend their behaviors. We use natural language to express those instructions and mimic human behavior to verify if models can understand and follow the instructions with respect to social stereotypes. It turns out to be a nontrivial topic and we propose our new task as a challenge for the community.

We summarize this thesis in Chapter 7.

CHAPTER 2

Bias in Training

In a typical machine learning pipeline, given a set of *training data* (pairs of input and targeted output), an AI model is learned to automatically discover underlying *representations* of input instances and conduct the *inference* in predicting output labels based on the representations. However, most of those collected dataset would be biased and hence the models trained on such datasets would also inherit the biases. In this chapter, I will use one NLP application – Coreference Resolution – as an example to demonstrate the biases in the training corpus and how to deal with such biases. This chapter is based on our work [ZWY18].

2.1 Introduction

Coreference resolution is a task aimed at identifying phrases (mentions) referring to the same entity. Various approaches, including rule-based [RLR10], feature-based [DK13, PCR15], and neural-network based [CM16, LHL17] have been proposed. While significant advances have been made, systems carry the risk of relying on societal stereotypes present in training data that could significantly impact their performance for some demographic groups.

In this work, we test the hypothesis that co-reference systems exhibit gender bias by creating a new challenge corpus, WinoBias.This dataset follows the winograd format [Hir81, RN12, PKR15], and contains references to people using a vocabulary of 40 occupations. It contains two types of challenge sentences that require linking gendered pronouns to either male or female stereotypical occupations (see the illustrative examples in Figure 2.1). None of the examples can be disambiguated by the gender of the pronoun but this cue can potentially distract the model. We consider a system to be gender biased if it links pronouns to occupations dominated by the gender of the pronoun (pro-stereotyped condition) more accurately than occupations not dominated by the gender of the pronoun (anti-stereotyped condition). The corpus can be used to certify a system has gender bias.[1]

We use three different systems as prototypical examples: the Stanford Deterministic Coreference System [RLR10], the Berkeley Coreference Resolution System [DK13] and the current best published system: the UW End-to-end Neural Coreference Resolution System [LHL17]. Despite qualitatively different approaches, all systems exhibit gender bias, showing an average difference in performance between pro-stereotypical and anti-stereotyped conditions of 21.1 in F1 score. Finally we show that given sufficiently strong alternative cues, systems can ignore their bias.

In order to study the source of this bias, we analyze the training corpus used by these systems, Ontonotes 5.0 [WPR12].[2] Our analysis shows that female entities are significantly underrepresented in this corpus. To reduce the impact of such dataset bias, we propose to generate an auxiliary dataset where all male entities are replaced by female entities, and vice versa, using a rule-based approach. Methods can then be trained on the union of the original and auxiliary dataset. In combination with methods that remove bias from fixed resources such as word embeddings [BCZ16], our data augmentation approach completely eliminates bias when evaluating on WinoBias, without significantly affecting overall coreference accuracy.

[1] Note that the counter argument (i.e., systems are gender bias free) may not hold.

[2] The corpus is used in CoNLL-2011 and CoNLL-2012 shared tasks, http://www.conll.org/previous-tasks

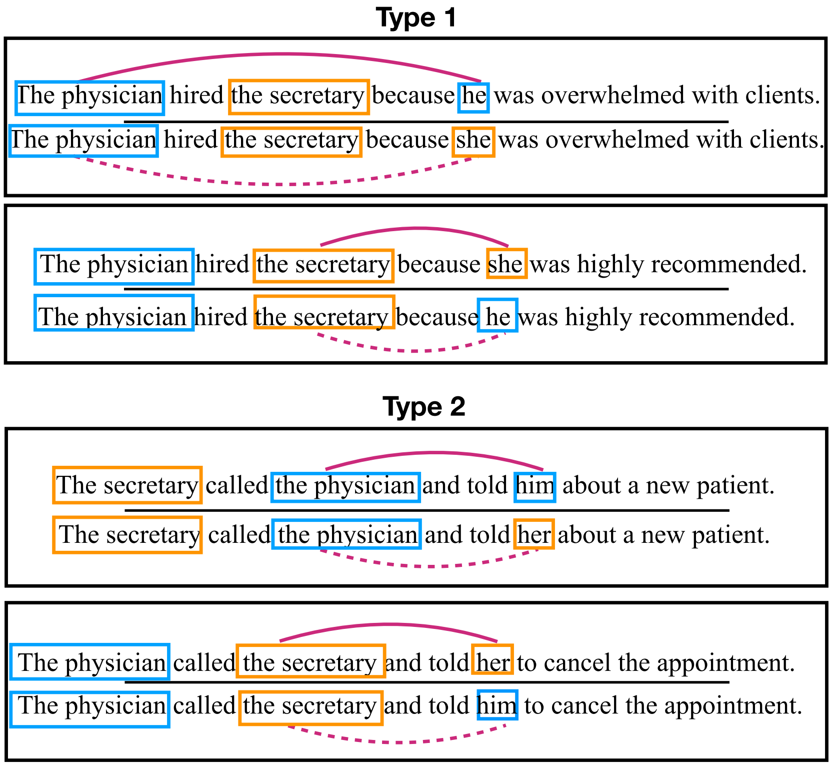

Type 1
The physician hired the secretary because he was overwhelmed with clients.
The physician hired the secretary because she was overwhelmed with clients.
The physician hired the secretary because she was highly recommended.
The physician hired the secretary because he was highly recommended.
Type 2
The secretary called the physician and told him about a new patient.
The secretary called the physician and told her about a new patient.
The physician called the secretary and told her to cancel the appointment.
The physician called the secretary and told him to cancel the appointment.

Figure 2.1: Pairs of gender balanced co-reference tests in the WinoBias dataset. Male and female entities are marked in solid blue and dashed orange, respectively. For each example, the gender of the pronominal reference is irrelevant for the co-reference decision. Systems must be able to make correct linking predictions in pro-stereotypical scenarios (solid purple lines) and anti-stereotypical scenarios (dashed purple lines) equally well to pass the test. Importantly, stereotypical occupations are considered based on US Department of Labor statistics.

2.2 WinoBias for Bias Evaluation

To better identify gender bias in coreference resolution systems, we build a new dataset centered on people entities referred by their occupations from a vocabulary of 40 occupations gathered from the US Department of Labor, shown in Table 2.1.[3] We use the associated occupation statistics to determine what constitutes gender stereotypical roles (e.g. 90% of nurses are women in this survey). Entities referred by different occupations are paired and used to construct test case scenarios. Sentences are duplicated using male and female pronouns, and contain equal numbers of correct co-reference decisions for all occupations. In total, the dataset contains 3,160 sentences, split equally for development and test, created by researchers familiar with the project. Sentences were created to follow two prototypical templates but annotators were encouraged to come up with scenarios where entities could be interacting in plausible ways. Templates were selected to be challenging and designed to cover cases requiring semantics and syntax separately.[4]

Type 1: [`entity1`] [interacts with] [`entity2`] [conjunction] [pronoun] [circumstances]. Prototypical WinoCoRef style sentences, where co-reference decisions must be made using world knowledge about given circumstances (Figure 2.1; Type 1). Such examples are challenging because they contain no syntactic cues.

Type 2: [`entity1`] [interacts with] [`entity2`] and then [interacts with] [pronoun] for [circumstances]. These tests can be resolved using syntactic information and understanding of the pronoun (Figure 2.1; Type 2). We expect systems to do well on such cases because both semantic and syntactic cues help disambiguation.

Evaluation To evaluate models, we split the data in two sections: one where correct co-reference decisions require linking a gendered pronoun to an occupation stereotypically associated with the gender of the pronoun and one that requires linking to the anti-stereotypical occupation. We say that a model passes the WinoBias test if for both Type 1 and Type 2 examples, pro-stereotyped and anti-stereotyped co-reference decisions are

[3] Labor Force Statistics from the Current Population Survey, 2017. https://www.bls.gov/cps/cpsaat11.htm

[4] We do not claim this set of templates is complete, but that they provide representative examples that, pratically, show bias in existing systems.

Occupation	%	Occupation	%	Occupation	%	Occupation	%
carpenter	2	chief	27	editor	52	teacher	78
mechanician	4	janitor	34	designers	54	sewer	80
construction worker	4	lawyer	35	accountant	61	librarian	84
laborer	4	cook	38	auditor	61	assistant	85
driver	6	physician	38	writer	63	cleaner	89
sheriff	14	ceo	39	baker	65	housekeeper	89
mover	18	analyst	41	clerk	72	nurse	90
developer	20	manager	43	cashier	73	receptionist	90
farmer	22	supervisor	44	counselors	73	hairdressers	92
guard	22	salesperson	48	attendant	76	secretary	95

Table 2.1: Occupations statistics used in WinoBias dataset, organized by the percent of people in the occupation who are reported as female. When woman dominate profession, we call linking the noun phrase referring to the job with female and male pronoun as 'pro-stereotypical', and 'anti-stereotypical', respectively. Similarly, if the occupation is male dominated, linking the noun phrase with the male and female pronoun is called, 'pro-stereotypical' and 'anti-steretypical', respectively.

made with the same accuracy.

2.3 Gender Bias in Coreference

In this section, we highlight two sources of gender bias in co-reference systems that can cause them to fail WinoBias: training data and auxiliary resources and propose strategies to mitigate them.

Method	Anon.	Resour.	Aug.	OntoNotes	T1-p	T1-a	Avg	| Diff |	T2-p	T2-a	Avg	| Diff |
E2E				**67.7**	**76.0**	49.4	62.7	26.6*	**88.7**	75.2	82.0	13.5*
E2E	✓			66.4	73.5	51.2	62.6	21.3*	86.3	70.3	78.3	16.1*
E2E	✓	✓		66.5	67.2	59.3	63.2	7.9*	81.4	82.3	81.9	0.9
E2E	✓		✓	66.2	65.1	59.2	62.2	5.9*	86.5	**83.7**	**85.1**	2.8*
E2E	✓	✓	✓	66.3	63.9	**62.8**	**63.4**	**1.1**	81.3	83.4	82.4	**2.1**
Feature				**61.7**	**66.7**	56.0	61.4	10.6*	**73.0**	57.4	65.2	15.7*
Feature	✓			61.3	65.9	56.8	61.3	9.1*	72.0	58.5	65.3	13.5*
Feature	✓	✓		61.2	61.8	**62.0**	**61.9**	**0.2**	67.1	63.5	65.3	3.6
Feature	✓		✓	61.0	65.0	57.3	61.2	7.7*	72.8	63.2	68.0	9.6*
Feature	✓	✓	✓	61.0	62.3	60.4	61.4	1.9*	71.1	**68.6**	**69.9**	**2.5**
Rule				57.0	76.7	37.5	57.1	39.2*	50.5	29.2	39.9	21.3*

Table 2.2: F1 on OntoNotes and WinoBias development set. WinoBias results are split between Type-1 and Type-2 and in pro/anti-stereotypical conditions. * indicates the difference between pro/anti stereotypical conditions is significant ($p < .05$) under an approximate randomized test [GMB14]. Our methods eliminate the difference between pro-stereotypical and anti-stereotypical conditions (Diff), with little loss in performance (OntoNotes and Avg).

2.3.1 Training Data Bias

Bias in OntoNotes 5.0 Resources supporting the training of co-reference systems have severe gender imbalance. In general, entities that have a mention headed by gendered pronouns (e.g."he", "she") are over 80% male.[5] Furthermore, the way in which such entities are referred to, varies significantly. Male gendered mentions are more than twice as likely to contain a job title as female mentions.[6] Moreover, these trends hold across genres.

[5] To exclude mentions such as "his mother", we use Collins head finder [Col03] to identify the head word of each mention, and only consider the mentions whose head word is gender pronoun.

[6] We pick more than 900 job titles from a gazetteer.

Method	Anon.	Resour.	Aug.	OntoNotes	T1-p	T1-a	Avg	| Diff |	T2-p	T2-a	Avg	| Diff |
E2E				**67.2**	**74.9**	47.7	61.3	27.2*	**88.6**	77.3	**82.9**	11.3*
E2E	✓	✓	✓	66.5	62.4	**60.3**	**61.3**	**2.1**	78.4	**78.0**	78.2	**0.4**
Feature				**64.0**	**62.9**	58.3	60.6	4.6*	**68.5**	57.8	63.1	10.7*
Feature	✓	✓	✓	63.6	62.2	**60.6**	**61.4**	**1.7**	**70.0**	**69.5**	**69.7**	**0.6**
Rule				58.7	72.0	37.5	54.8	34.5*	47.8	26.6	37.2	21.2*

Table 2.3: F1 on OntoNotes and Winobias test sets. Methods were run once, supporting development set conclusions.

Gender Swapping To remove such bias, we construct an additional training corpus where all male entities are swapped for female entities and vice-versa. Methods can then be trained on both original and swapped corpora. This approach maintains non-gender-revealing correlations while eliminating correlations between gender and co-reference cues.

We adopt a simple rule based approach for gender swapping. First, we anonymize named entities using an automatic named entity finder [LBS16]. Named entities are replaced consistently within document (i.e. "Barak Obama ... Obama was re-elected." would be annoymized to "E1 E2 ... E2 was re-elected."). Then we build a dictionary of gendered terms and their realization as the opposite gender by asking workers on Amazon Mechnical Turk to annotate all unique spans in the OntoNotes development set.[7] Rules were then mined by computing the word difference between initial and edited spans. Common rules included "she → he", "Mr." → "Mrs.", "mother" → "father." Sometimes the same initial word was edited to multiple different phrases: these were resolved by taking the most frequent phrase, with the exception of "her → him" and "her → his" which were resolved using part-of-speech. Rules were applied to all matching tokens in the OntoNotes. We maintain anonymization so that cases like "John went to his house" can be accurately swapped to "E1 went to her house."

[7] Five turkers were presented with anonymized spans and asked to mark if it indicated male, female, or neither, and if male or female, rewrite it so it refers to the other gender.

Model	Original	Gender-reversed
E2E	66.4	65.9
Feature	61.3	60.3

Table 2.4: Performance on the original and the gender-reversed developments dataset (anonymized).

2.3.2 Resource Bias

Word Embeddings Word embeddings are widely used in NLP applications however recent work has shown that they are severely biased: "man" tends to be closer to "programmer" than "woman" [BCZ16, CBN17]. Current state-of-art co-reference systems build on word embeddings and risk inheriting their bias. To reduce bias from this resource, we replace GloVe embeddings with debiased vectors [BCZ16].

Gender Lists While current neural approaches rely heavily on pre-trained word embeddings, previous feature rich and rule-based approaches rely on corpus based gender statistics mined from external resources [BL06]. Such lists were generated from large unlabeled corpora using heuristic data mining methods. These resources provide counts for how often a noun phrase is observed in a male, female, neutral, and plural context. To reduce this bias, we balance male and female counts for all noun phrases.

2.4 Result

In this section we evaluate of three representative systems: rule based, Rule, [RLR10], feature-rich, Feature, [DK13], and end-to-end neural (the current state-of-the-art), E2E, [LHL17]. The following sections show that performance on WinoBias reveals gender bias in all systems, that our methods remove such bias, and that systems are less biased on OntoNotes data.

WinoBias Reveals Gender Bias Table 5.2 summarizes development set evaluations using all three systems. Systems were evaluated on both types of sentences in WinoBias (T1 and T2), separately in pro-stereotyped and anti-stereotyped conditions (T1-p vs. T1-a, T2-p vs T2-a). We evaluate the effect of named-entity anonymization (Anon.), debiasing supporting resources[8] (Resour.) and using data-augmentation through gender swapping (Aug.). E2E and Feature were retrained in each condition using default hyper-parameters while Rule was not debiased because it is untrainable. We evaluate using the coreference scorer v8.01 [PLR14] and compute the average (Avg) and absolute difference (Diff) between pro-stereotyped and anti-stereotyped conditions in WinoBias.

All initial systems demonstrate severe disparity between pro-stereotyped and anti-stereotyped conditions. Overall, the rule based system is most biased, followed by the neural approach and feature rich approach. Across all conditions, anonymization impacts E2E the most, while all other debiasing methods result in insignificant loss in performance on the OntoNotes dataset. Removing biased resources and data-augmentation reduce bias independently and more so in combination, allowing both E2E and Feature to pass WinoBias without significantly impacting performance on either OntoNotes or WinoBias . Qualitatively, the neural system is easiest to de-bias and our approaches could be applied to future end-to-end systems. Systems were evaluated once on test sets, Table 2.3, supporting our conclusions.

Systems Demonstrate Less Bias on OntoNotes While we have demonstrated coreference systems have severe bias as measured in WinoBias , this is an out-of-domain test for systems trained on OntoNotes. Evaluating directly within OntoNotes is challenging because sub-sampling documents with more female entities would leave very few evaluation data points. Instead, we apply our gender swapping system (Section 3.2.2), to the OntoNotes development set and compare system performance between swapped and

[8] Word embeddings for E2E and gender lists for Feature

unswapped data.[9] If a system shows significant difference between original and gender-reversed conditions, then we would consider it gender biased on OntoNotes data.

Table 2.4 summarizes our results. The E2E system does not demonstrate significant degradation in performance, while Feature loses roughly 1.0-F1.[10] This demonstrates that given sufficient alternative signal, systems often do ignore gender biased cues. On the other hand, WinoBias provides an analysis of system bias in an adversarial setup, showing, when examples are challenging, systems are likely to make gender biased predictions.

2.5 Discussion

Bias in NLP systems has the potential to not only mimic but also amplify stereotypes in society. For a prototypical problem, coreference, we provide a method for detecting such bias and show that three systems are significantly gender biased. We also provide evidence that systems, given sufficient cues, can ignore their bias. Finally, we present general purpose methods for making co-reference models more robust to spurious, gender-biased cues while not incurring significant penalties on their performance on benchmark datasets.

[9] This test provides a lower bound on OntoNotes bias because some mistakes can result from errors introduce by the gender swapping system.

[10] We do not evaluate the Rule system as it cannot be train for anonymized input.

CHAPTER 3

Bias in Representations

In this section, we revisit the biases in representations. More specifically, we will review the problem in both word embeddings and contextualized word embeddings. Other than English, we will also cover the bias analysis in multilingual embeddings. In each case, we will discuss possible ways to mitigate the biases. This section is based on our work [ZZL18, ZWY18, ZMH20].

3.1 Learning Gender-Neutral Word Embeddings

Word embedding models have become a fundamental component in a wide range of NLP applications. However, embeddings trained on human-generated corpora have been demonstrated to inherit strong gender stereotypes that reflect social constructs. To address this concern, in this section, we propose a novel training procedure for learning gender-neutral word embeddings. Our approach aims to preserve gender information in certain dimensions of word vectors while compelling other dimensions to be free of gender influence. Based on the proposed method, we generate a Gender-Neutral variant of GloVe (GN-GloVe). Quantitative and qualitative experiments demonstrate that GN-GloVe successfully isolates gender information without sacrificing the functionality of the embedding model.

3.1.1 Introduction

Word embedding models have been designed for representing the meaning of words in a vector space. These models have become a fundamental NLP technique and have been widely used in various applications. However, prior studies show that such models learned from human-generated corpora are often prone to exhibit social biases, such as gender stereotypes [BCZ16, CBN17]. For example, the word "programmer" is neutral to gender by its definition, but an embedding model trained on a news corpus associates "programmer" closer with "male" than "female".

To alleviate gender stereotype in word embeddings, [BCZ16] propose a post-processing method that projects gender-neutral words to a subspace which is perpendicular to the gender dimension defined by a set of gender-definition words.[1] However, their approach has two limitations. First, the method is essentially a pipeline approach and requires the gender-neutral words to be identified by a classifier before employing the projection. If the classifier makes a mistake, the error will be propagated and affect the performance of the model. Second, their method completely removes gender information from those words which are essential in some domains such as medicine and social science [BPS10, MMP92].

To overcome these limitations, we propose a learning scheme, Gender-Neutral Global Vectors (GN-GloVe) for training word embedding models with protected attributes (e.g., gender) based on GloVe [PSM14].[2] GN-GloVe represents protected attributes in certain dimensions while neutralizing the others during training. As the information of the protected attribute is restricted in certain dimensions, it can be removed from the embedding easily. By jointly identifying gender-neutral words while learning word vectors, GN-GloVe does not require a separate classifier to identify gender-neutral words; therefore, the error propagation issue is eliminated. The proposed approach is generic and can be incorporated with other word embedding models and be applied in reducing other societal stereotypes.

[1] Gender-definition words are the words associated with gender by definition (e,g., mother, waitress); the remainder are gender-neutral words.

[2] The code and data are released at `https://github.com/uclanlp/gn_glove`

3.1.2 Methodology

In this paper, we take GloVe [PSM14] as the base embedding model and gender as the protected attribute. It is worth noting that our approach is general and can be applied to other embedding models and attributes. Following GloVe [PSM14], we construct a word-to-word co-occurrence matrix X, denoting the frequency of the j-th word appearing in the context of the i-th word as $X_{i,j}$. $w, \tilde{w} \in \mathbb{R}^d$ stand for the embeddings of a center and a context word, respectively, where d is the dimension.

In our embedding model, a word vector w consists of two parts $w = [w^{(a)}; w^{(g)}]$. $w^{(a)} \in \mathbb{R}^{d-k}$ and $w^{(g)} \in \mathbb{R}^k$ stand for neutralized and gendered components respectively, where k is the number of dimensions reserved for gender information.[3] Our proposed gender neutralizing scheme is to reserve the gender feature, known as "protected attribute" into $w^{(g)}$. Therefore, the information encoded in $w^{(a)}$ is independent of gender influence. We use $v_g \in \mathbb{R}^{d-k}$ to denote the direction of gender in the embedding space. We categorize all the vocabulary words into three subsets: male-definition Ω_M, female-definition Ω_F, and gender-neutral Ω_N, based on their definition in WordNet [MF98].

Gender Neutral Word Embedding Our minimization objective is designed in accordance with above insights. It contains three components:

$$J = J_G + \lambda_d J_D + \lambda_e J_E, \tag{3.1}$$

where λ_d and λ_e are hyper-parameters.

The first component J_G is originated from GloVe [PSM14], which captures the word proximity:

$$J_G = \sum_{i,j=1}^{V} f(X_{i,j}) \left(w_i^T \tilde{w}_j + b_i + \tilde{b}_j - \log X_{i,j} \right)^2 .$$

Here, $f(X_{i,j})$ is a weighting function to reduce the influence of extremely large co-occurrence frequencies. b and $\tilde{b}$ are the respective linear biases for w and $\tilde{w}$.

[3] We set $k = 1$ in this paper.

The other two terms are aimed to restrict gender information in $w^{(g)}$, such that $w^{(a)}$ is neutral. Given male- and female-definition seed words Ω_M and Ω_F, we consider two distant metrics and form two types of objective functions.

In J_D^{L1}, we directly minimizing the negative distances between words in the two groups:

$$J_D^{L1} = -\left\| \sum_{w \in \Omega_M} w^{(g)} - \sum_{w \in \Omega_F} w^{(g)} \right\|_1 .$$

In J_D^{L2}, we restrict the values of word vectors in $[\beta_1, \beta_2]$ and push $w^{(g)}$ into one of the extremes:

$$J_D^{L2} = \sum_{w \in \Omega_M} \left\| \beta_1 \boldsymbol{e} - w^{(g)} \right\|_2^2 + \sum_{w \in \Omega_F} \left\| \beta_2 \boldsymbol{e} - w^{(g)} \right\|_2^2 ,$$

where $\boldsymbol{e} \in \mathcal{R}^k$ is a vector of all ones. β_1 and β_2 can be arbitrary values, and we set them to be 1 and -1, respectively.

Finally, for words in Ω_N, the last term encourages their $w^{(a)}$ to be retained in the null space of the gender direction v_g:

$$J_E = \sum_{w \in \Omega_N} \left(v_g^T w^{(a)} \right)^2 ,$$

where v_g is estimating on the fly by averaging the differences between female words and their male counterparts in a predefined set,

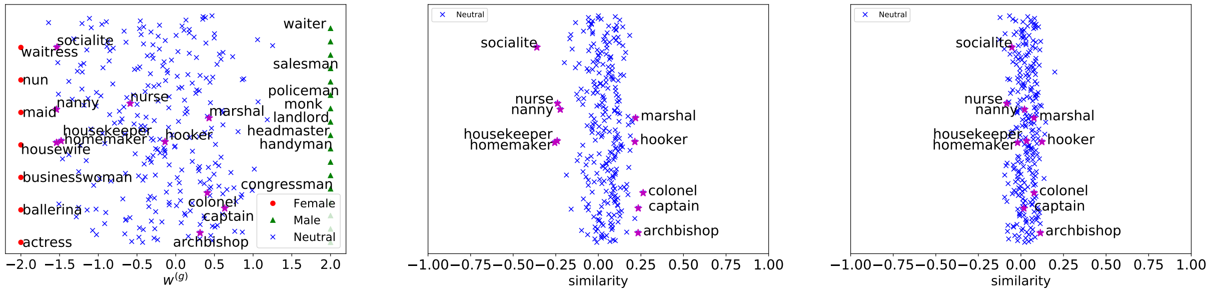
$$v_g = \frac{1}{|\Omega'|} \sum_{(w_m, w_f) \in \Omega'} (w_m^{(a)} - w_f^{(a)}),$$

where Ω' is a set of predefined gender word pairs.

We use stochastic gradient descent to optimize Eq. (3.1). To reduce the computational complexity in training the wording embedding, we assume v_g is a fixed vector (i.e., we do not derive gradient w.r.t v_g in updating $w^{(a)}, \forall w \in \Omega'$) and estimate v_g only at the beginning of each epoch.

3.1.3 Experiments

In this section, we conduct the following qualitative and quantitative studies: 1) We visualize the embedding space and show that GN-GloVe separates the protected gender attribute



(a) $w^{(g)}$ dimension for all the professions

(b) Gender-neutral profession words projected to gender direction in GloVe

(c) Gender-neutral profession words projected to gender direction in GN-GloVe

Figure 3.1: Cosine similarity between the gender direction and the embeddings of gender-neutral words. In each figure, negative values represent a bias towards female, otherwise male.

from other latent aspects; 2) We measure the ability of GN-GloVe to distinguish between gender-definition words and gender-stereotype words on a newly annotated dataset; 3) We evaluate GN-GloVe on standard word embedding benchmark datasets and show that it performs well in estimating word proximity; 4) We demonstrate that GN-Glove reduces gender bias on a downstream application, coreference resolution.

We compare GN-GloVe with two embedding models, GloVe and Hard-GloVe. GloVe is a widely-used model [PSM14], and we apply the post-processing step introduced in [BCZ16] to reduce gender bias in GloVe and name it after Hard-GloVe. All the embeddings are trained on *2017 English Wikipedia dump* with the default hyper-parameters decribed in [PSM14]. When training GN-GloVe, we constrain the value of each dimension within $[-1, 1]$ to avoid numerical difficulty. We set λ_d and λ_e both to be 0.8. In our preliminary study on development data, we observe that the model is not sensitive to these parameters. Unless other stated, we use J_D^{L1} in the GN-GloVe model.

Separate protected attribute First, we demonstrate that GN-GloVe preserves the gender association (either definitional or stereotypical associations) in $w^{(g)}$[4]. To illustrate the distribution of gender information of different words, we plot Fig. 3.1a using $w^{(g)}$ for the x-axis and a random value for the y-axis to spread out words in the plot. As shown in the figure, the gender-definition words, e.g. "waiter" and "waitress", fall far away from each other in $w^{(g)}$. In addition, words such as "housekeeper" and "doctor" are inclined to different genders and their $w^{(g)}$ preserves such information.

Next, we demonstrate that GN-GloVe reduces gender stereotype using a list of profession titles from [BCZ16]. All these profession titles are neutral to gender by definition. In Fig. 3.1b and Fig. 3.1c, we plot the cosine similarity between each word vector $w^{(a)}$ and the gender direction v_g (i.e., $\frac{w^T v_g}{\|w\| \|v_g\|}$). Result shows that words, such as "doctor" and "nurse", possess no gender association by definition, but their GloVe word vectors exhibit strong gender stereotype. In contrast, the gender projects of GN-GloVe word vectors $w^{(a)}$ are closer to zero. This demonstrates the gender information has been substantially diminished from $w^{(a)}$ in the GN-GloVe embedding.

We further quantify the gender information exhibited in the embedding models. For each model, we project the word vectors of occupational words into the gender sub-space defined by "he-she" and compute their average size. A larger projection indicates an embedding model is more biased. Results show that the average projection of GloVe is 0.080, the projection of Hard-GloVe is 0.019, and the projection of Gn-Glove is 0.052. Comparing with GloVe, GN-GloVe reduces the bias by 35%. Although Hard-GloVe contains less gender information, we will show later GN-GloVe can tell difference between gender-stereotype and gender-definition words better.

Gender Relational Analogy To study the quality of the gender information present in each model, we follow SemEval 2012 Task2 [JTM12] to create an analogy dataset, *Sem-*

[4] We follow the original GloVe implementation using the summation of word vector and context vector to represent a word. Therefore, the elements of the word vectors are constrained in [-2, 2]

Dataset	Embeddings	Definition	Stereotype	None
SemBias	GloVe	80.2	10.9	8.9
	Hard-Glove	84.1	6.4	9.5
	GN-GloVe	97.7	1.4	0.9
SemBias (subset)	GloVe	57.5	20	22.5
	Hard-Glove	25	27.5	47.5
	GN-GloVe	75	15	10

Table 3.1: Percentage of predictions for each category on gender relational analogy task.

Bias, with the goal to identify the correct analogy of "he - she" from four pairs of words. Each instance in the dataset consists of four word pairs: a gender-definition word pair (Definition; e.g., "waiter - waitress"), a gender-stereotype word pair (Stereotyp; e.g., "doctor - nurse") and two other pairs of words that have similar meanings (None; e.g., "dog - cat", "cup - lid")[5]. We consider 20 gender-stereotype word pairs and 22 gender-definition word pairs and use their Cartesian product to generate 440 instances. Among the 22 gender-definition word pairs, there are 2 word pairs that are not used as a seed word during the training. To test the generalization ability of the model, we generate a subset of data (SemBias (subset)) of 40 instances associated with these 2 pairs.

Table 3.1 lists the percentage of times that each class of pair is on the top based on a word embedding model [MYZ13]. GN-GloVe achieves 97.7% accuracy in identifying gender-definition word pairs as an analogy to "he - she". In contrast, GloVe and Hard-GloVe makes significantly more mistakes. On the subset, GN-GloVe also achieves significantly better performance than Hard-Glove and GloVe, indicating that it can generalize the gender pairs on the training set to identify other gender-definition word pairs.

[5] The pair is sampled from the list of word pairs with "SIMILAR: Coordinates" relation annotated in [JTM12]. The original list has 38 pairs. After removing gender-definition word pairs, 29 are left.

Embeddings	Analogy		Similarity					
	Google	MSR	WS353-ALL	RG-65	MTurk-287	MTurk-771	RW	MEN-TR-3k
GloVe	**70.8**	**45.8**	62.0	75.3	64.8	64.9	37.3	72.2
Hard-GloVe	**70.8**	**45.8**	61.2	74.8	64.4	64.8	37.3	72.2
GN-GloVe-L1	68.9	43.7	**62.8**	74.1	66.2	**66.2**	**40.0**	**74.5**
GN-GloVe-L2	68.8	43.6	62.5	**76.4**	**66.8**	65.6	39.3	74.4

Table 3.2: Results on the benchmark datasets. Performance is measured in accuracy and in Spearman rank correlation for word analogy and word similarity tasks, respectively.

Word Similarity and Analogy In addition, we evaluate the word embeddings on the benchmark tasks to ensure their quality. The word similarity tasks measure how well a word embedding model captures the similarity between words comparing to human annotated rating scores. Embeddings are tested on multiple datasets: WS353-ALL [FGM01], RG-65 [RG65], MTurk-287 [RAG11], MTurk-771 [HDG12], RW [LSM13], and MEN-TR-3k [BBB12] datasets. The analogy tasks are to answer the question "*A* is to *B* as *C* is to _?" by finding a word vector w that is closest to $w_A - w_B + w_C$ in the embedding space. Google [MCC13] and MSR [MYZ13] datasets are utilized for this evaluation. The results are shown in Table 3.2, where the suffix "-L1" and "-L2" of GN-GloVe stand for the GN-GloVe using J_D^{L1} and J_D^{L2}, respectively. Compared with others, GN-GloVe achieves a higher accuracy in the similarity tasks and its analogy score slightly drops indicating that GN-GloVe is capable of preserving proximity among words.

Coreference Resolution Finally, we investigate how the gender bias in word embeddings affects a downstream application, such as coreference resolution. Coreference resolution aims at clustering the denotative noun phrases referring to the same entity in the given text. We evaluate our models on the Ontonotes 5.0 [WPR12] benchmark dataset and the WinoBias dataset [ZWY18].[6] In particular, the WinoBias dataset is composed of

[6] Specifically, we conduct experiments on the Type 1 version.

Embeddings	OntoNotes-test	PRO	ANTI	Avg	Diff
GloVe	66.5	76.2	46.0	61.1	30.2
Hard-Glove	66.2	70.6	54.9	62.8	15.7
GN-GloVe	66.2	72.4	51.9	62.2	20.5
GN-GloVe(w_a)	65.9	70.0	53.9	62.0	16.1

Table 3.3: F1 score (%) on the coreference system.

pro-stereotype (PRO) and anti-stereotype (ANTI) subsets. The PRO subset consists of sentences where a gender pronoun refers to a profession, which is dominated by the same gender. Example sentences include "The CEO raised the salary of the receptionist because he is generous." In this sentence, the pronoun "he" refers to "CEO" and this reference is consistent with societal stereotype. The ANTI subset contains the same set of sentences, but the gender pronoun in each sentence is replaced by the opposite gender. For instance, the gender pronoun "he" is replaced by "she" in the aforementioned example. Despite the sentence is almost identical, the gender pronoun now refers to a profession that is less represented by the gender. Details about the dataset are in [ZWY18].

We train the end-to-end coreference resolution model [LHL17] with different word embeddings on OntoNote and report their performance in Table 3.3. For the WinoBias dataset, we also report the average (Avg) and absolute difference (Diff) of F1 scores on two subsets. A smaller Diff value indicates less bias in a system. Results show that GN-GloVe achieves comparable performance as Glove and Hard-GloVe on the OntoNotes dataset while distinctly reducing the bias on the WinoBias dataset. When only the $w^{(a)}$ potion of the embedding is used in representing words, GN-GloVe($w^{(a)}$) further reduces the bias in coreference resolution.

3.1.4 Discussion

In this section, we introduced an algorithm for training gender-neutral word embedding. Our method is general and can be applied in any language as long as a list of gender definitional words is provided as seed words (e.g., gender pronouns). Future directions include extending the proposed approach to model other properties of words such as sentiment and generalizing our analysis beyond binary gender.

3.2 Bias Analysis in Contextualized Word Embeddings

In this section, we quantify, analyze and mitigate gender bias exhibited in contextualized word vectors (ELMo specifically). First, we conduct several intrinsic analyses and find that (1) training data for ELMo contains significantly more male than female entities, (2) the trained ELMo embeddings systematically encode gender information and (3) ELMo unequally encodes gender information about male and female entities. Then, we show that a state-of-the-art coreference system that depends on ELMo inherits its bias and demonstrates significant bias on the WinoBias probing corpus. Finally, we explore two methods to mitigate such gender bias and show that the bias demonstrated on WinoBias can be eliminated.

3.2.1 Introduction

Distributed representations of words in the form of word embeddings [MSC13, PSM14] and contextualized word embeddings [PNI18, DCL18, RNS18, MBX17, RWC19] have led to huge performance improvement on many NLP tasks. However, several recent studies show that training word embeddings in large corpora could lead to encoding societal biases present in these human-produced data [BCZ16, CBN17]. In this work, we extend these analyses to the ELMo contextualized word embeddings.

Our work provides a new intrinsic analysis of how ELMo represents gender in biased ways. First, the corpus used for training ELMo has a significant gender skew: male entities are nearly three times more common than female entities, which leads to gender bias in the downloadable pre-trained contextualized embeddings. Then, we apply principal component analysis (PCA) to show that after training on such biased corpora, there exists a low-dimensional subspace that captures much of the gender information in the contextualized embeddings. Finally, we evaluate how faithfully ELMo preserves gender information in sentences by measuring how predictable gender is from ELMo representations of occupation words that co-occur with gender revealing pronouns. Our results show that ELMo embeddings perform unequally on male and female pronouns: male entities can be predicted from occupation words 14% more accurately than female entities.

In addition, we examine how gender bias in ELMo propagates to the downstream applications. Specifically, we evaluate a state-of-the-art coreference resolution system [LHZ18] that makes use of ELMo's contextual embeddings on WinoBias [ZWY18] (as described in Chap. 2), a coreference diagnostic dataset that evaluates whether systems behave differently on decisions involving male and female entities of stereotyped or anti-stereotyped occupations. We find that in the most challenging setting, the ELMo-based system has a disparity in accuracy between pro- and anti-stereotypical predictions, which is nearly 30% higher than a similar system based on GloVe [LHL17].

Finally, we investigate approaches for mitigating the bias which propagates from the contextualized word embeddings to a coreference resolution system. We explore two different strategies: (1) a training-time data augmentation technique [ZWY18] (as described in Chap. 2), where we augment the corpus for training the coreference system with its gender-swapped variant (female entities are swapped to male entities and vice versa) and, afterwards, retrain the coreference system; and (2) a test-time embedding neutralization technique, where input contextualized word representations are averaged with word representations of a sentence with entities of the opposite gender. Results show that test-time embedding neutralization is only partially effective, while data augmentation largely mitigates bias demonstrated on WinoBias by the coreference system.

3.2.2 Gender Bias in ELMo

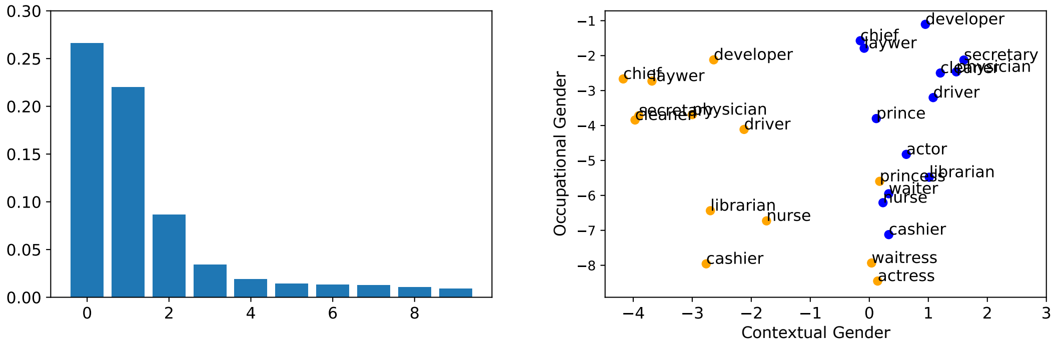


Figure 3.2: Left: Percentage of explained variance in PCA in the embedding differences. Right: Selected words projecting to the first two principle components where the blue dots are the sentences with male context and the orange dots are from the sentences with female context.

In this section we describe three intrinsic analyses highlighting gender bias in trained ELMo contextual word embeddings [PNI18]. We show that (1) training data for ELMo contains significantly more male entities compared to female entities leading to gender bias in the pre-trained contextual word embeddings (2) the geometry of trained ELMo embeddings systematically encodes gender information and (3) ELMo propagates gender information about male and female entities unequally.

3.2.2.1 Training Data Bias

Table 3.4 lists the data analysis on the One Billion Word Benchmark [CMS13] corpus, the training corpus for ELMo. We show counts for the number of occurrences of male pronouns (*he*, *his* and *him*) and female pronouns (*she* and *her*) in the corpus as well as the co-occurrence of occupation words with those pronouns. We use the set of occupation words defined in the WinoBias corpus and their assignments as prototypically male or female [ZWY18]. The analysis shows that the Billion Word corpus contains a signifi-

	#occurrence	#M-biased occs.	#F-biased occs.
M	5,300,000	170,000	81,000
F	1,600,000	33,000	36,000

Table 3.4: Training corpus for ELMo. We show total counts for male (M) and female (F) pronouns in the corpus, and counts corresponding to their co-occurrence with occupation words where the occupations are stereotypically male (M-biased) or female (F-biased).

cant skew with respect to gender: (1) male pronouns occur three times more than female pronouns and (2) male pronouns co-occur more frequently with occupation words, irrespective of whether they are prototypically male or female.

3.2.2.2 Geometry of Gender

Next, we analyze the gender subspace in ELMo. We first sample 400 sentences with at least one gendered word (e.g., *he* or *she* from the OntoNotes 5.0 dataset [WPR12] and generate the corresponding gender-swapped variants (changing *he* to *she* and vice-versa). We then calculate the difference of ELMo embeddings between occupation words in corresponding sentences and conduct principal component analysis for all pairs of sentences. Figure 3.2 shows there are two principal components for gender in ELMo, in contrast to GloVe which only has one [BCZ16]. The two principal components in ELMo seem to represent the gender from the contextual information (Contextual Gender) as well as the gender embedded in the word itself (Occupational Gender).

To visualize the gender subspace, we pick a few sentence pairs from WinoBias [ZWY18]. Each sentence in the corpus contains one gendered pronoun and two occupation words, such as “The developer corrected the secretary because she made a mistake” and also the same sentence with the opposite pronoun (he). In Figure 3.2 on the right, we project the ELMo embeddings of occupation words that are co-referent with the pronoun (e.g. *secretary* in the above example) for when the pronoun is male (blue dots) and female (orange dots) on the two principal components from the PCA analysis. Qualitatively, we can see the first component separates male and female contexts while the second component groups male related words such as *lawyer* and *developer* and female related words such as *cashier* and *nurse*.

3.2.2.3 Unequal Treatment of Gender

To test how ELMo embeds gender information in contextualized word embeddings, we train a classifier to predict the gender of entities from occupation words in the same sentence. We collect sentences containing gendered words (e.g., *he-she*, *father-mother*) and occupation words (e.g., *doctor*)[7] from the OntoNotes 5.0 corpus [WPR12], where we treat occupation words as a mention to an entity, and the gender of that entity is taken to the gender of a co-referring gendered word, if one exists. For example, in the sentence "the engineer went back to her home," we take *engineer* to be a female mention. Then we split all such instances into training and test, with 539 and 62 instances, respectively and augment these sentences by swapping all the gendered words with words of the opposite gender such that the numbers of male and female entities are balanced.

We first test if ELMo embedding vectors carry gender information. We train an SVM classifier with an RBF kernel[8] to predict the gender of a mention (i.e., an occupation word) based on its ELMo embedding. On development data, this classifier achieves 95.1% and 80.6% accuracy on sentences where the true gender was male and female respectively. For both male and female contexts, the accuracy is much larger than 50%, demonstrating that ELMo does propagate gender information to other words. However, male information is more than 14% more accurately represented in ELMo than female information, showing that ELMo propagates the information unequally for male and female entities.

[7] We use the list collected in [ZWY18]

[8] We use the ν-SVC formulation and tune the hyper-parameter ν [CL11] in the range of $[0.1, 1]$ with a step 0.1.

Embeddings	Data Augmentation	Neutralization		OntoNotes	Semantics Only				w/ Syntactic Cues			
		GloVe	ELMo		Pro.	Anti.	Avg.	| Diff |	Pro.	Anti.	Avg.	| Diff |
GloVe				67.7	76.0	49.4	62.7	26.6*	88.7	75.2	82.0	13.5*
GloVe	✓			65.8	63.9	62.8	63.4	1.1	81.3	83.4	82.4	2.1
GloVe+ELMo				72.7	79.1	49.5	64.3	29.6*	93.0	85.9	89.5	7.1*
GloVe+ELMo	✓			71.0	65.9	64.9	65.4	1.0	87.8	88.9	88.4	1.2
GloVe+ELMo		✓		71.0	72.6	57.8	64.9	14.3*	90.2	88.6	89.4	1.6
GloVe+ELMo		✓	✓	71.1	71.7	60.6	66.2	11.1*	90.3	89.2	89.8	1.1

Table 3.5: F1 on OntoNotes and WinoBias development sets. WinoBias dataset is split Semantics Only and w/ Syntactic Cues subsets. ELMo improves the performance on the OntoNotes dataset by 5% but shows stronger bias on the WinoBias dataset. Avg. stands for averaged F1 score on the pro- and anti-stereotype subsets while "Diff." is the absolute difference between these two subsets. * indicates the difference between pro/anti stereotypical conditions is significant ($p < .05$) under an approximate randomized test [GMB14]. Mitigating bias by data augmentation reduces all the bias from the coreference model to a neglect level. However, the neutralizing ELMo approach only mitigates bias when there are other strong learning signals for the task.

3.2.3 Bias in Coreference Resolution

In this section, we establish that coreference systems that depend on ELMo embeddings exhibit significant gender bias. Then we evaluate two simple methods for removing the bias from the systems and show that the bias can largely be reduced.

3.2.3.1 Setup

We evaluate bias with respect to the WinoBias dataset, a benchmark of paired male and female coreference resolution examples following the Winograd format [Hir81, RN12, PKR15]. It contains two different subsets, pro-stereotype, where pronouns are associated with occupations predominately associated with the gender of the pronoun, or anti-stereotype, when the opposite relation is true. Each subset consists of two types of sentences: one that requires semantic understanding of the sentence to make coreference resolution (Semantics Only) and another that relies on syntactic cues (w/ Syntactic Cues). Gender bias is measured by taking the difference of the performance in pro- and anti-stereotypical subsets (see Chap. 2 for more details). Previous work [ZWY18] evaluated the systems based on GloVe embeddings but here we evaluate a state-of-the-art system that trained on the OntoNotes corpus with ELMo embeddings [LHZ18].

3.2.3.2 Bias Mitigation Methods

Next, we describe two methods for mitigating bias in ELMo for the purpose of coreference resolution: (1) a train-time data augmentation approach and (2) a test-time neutralization approach.

Data Augmentation In Chap. 2, we propose a method to reduce gender bias in coreference resolution by augmenting the training corpus for this task. Data augmentation is performed by replacing gender revealing entities in the OntoNotes dataset with words indicating the opposite gender and then training on the union of the original data and this swapped data. In addition, they find it useful to also mitigate bias in supporting resources and therefore replace standard GloVe embeddings with bias mitigated word embeddings from [BCZ16]. We evaluate the performance of both aspects of this approach.

Neutralization We also investigate an approach to mitigate bias induced by ELMo embeddings without retraining the coreference model. Instead of augmenting training corpus by swapping gender words, we generate a gender-swapped version of the test instances. We then apply ELMo to obtain contextualized word representations of the original and the gender-swapped sentences and use their average as the final representations.

3.2.3.3 Results

Table 3.5 summarizes our results on WinoBias.

ELMo Bias Transfers to Coreference Row 3 in Table 3.5 summarizes performance of the ELMo based coreference system on WinoBias. While ELMo helps to boost the coreference resolution F1 score (OntoNotes) it also propagates bias to the task. It exhibits large differences between pro- and anti-stereotyped sets (|Diff|) on both semantic and syntactic examples in WinoBias.

Bias Mitigation Rows 4-6 in Table 3.5 summarize the effectiveness of the two bias mitigation approaches we consider. Data augmentation is largely effective at mitigating bias in the coreference resolution system with ELMo (reducing |Diff | to insignificant levels) but requires retraining the system. Neutralization is less effective than augmentation and cannot fully remove gender bias on the Semantics Only portion of WinoBias, indicating it is effective only for simpler cases. This observation is consistent with [GG19], where they show that entirely removing bias from an embedding is difficult and depends on the manner, by which one measures the bias.

3.2.4 Discussion

Like word embedding models, contextualized word embeddings inherit implicit gender bias. We analyzed gender bias in ELMo, showing that the corpus it is trained on has significant gender skew and that ELMo is sensitive to gender, but unequally so for male and female entities. We also showed this bias transfers to downstream tasks, such as coreference resolution, and explored two bias mitigation strategies: 1) data augmentation and 2) neutralizing embeddings, effectively eliminating the bias from ELMo in a state-of-the-art system. With increasing adoption of contextualized embeddings to get better results on core NLP tasks, e.g. BERT [DCL18], we must be careful how such unsupervised methods perpetuate bias to downstream applications and our work forms the basis of evaluating and mitigating such bias.

3.3 Gender Bias in Multi-lingual Embeddings and Cross-lingual Transfer Learning

Multilingual representations embed words from many languages into a single semantic space such that words with similar meanings are close to each other regardless of the language. These embeddings have been widely used in various settings, such as cross-lingual transfer, where an NLP model trained on one language is deployed to another language. While the cross-lingual transfer techniques are powerful, they carry gender bias from the source to target languages. In this section, we study gender bias in multilingual embeddings and how it affects transfer learning for NLP applications. We create a multilingual dataset for bias analysis and propose several ways for quantifying bias in multilingual representations from both the intrinsic and extrinsic perspectives. Experimental results show that the magnitude of bias in the multilingual representations changes differently when we align the embeddings to different target spaces and that the alignment direction can also have an influence on the bias in transfer learning. We further provide recommendations for using the multilingual word representations for downstream tasks. This section is based on our work [ZWY19].

3.3.1 Introduction

Natural Language Processing (NLP) plays a vital role in applications used in our daily lives. Despite the great performance inspired by the advanced machine learning techniques and large available datasets, there are potential societal biases embedded in these NLP tasks – where the systems learn inappropriate correlations between the final predictions and sensitive attributes such as gender and race. For example, [ZWY18] and [RNL18] demonstrate that coreference resolution systems perform unequally on different gender groups. Other studies show that such bias is exhibited in various components of the NLP systems, such as the training dataset [ZWY18, RNL18], the embeddings [BCZ16, CBN17, ZSZ19, MYB19] as well as the pre-trained models [ZWY19, KVP19].

Recent advances in NLP require large amounts of training data. Such data may be available for resource-rich languages such as English, but they are typically absent for many other languages. Multilingual word embeddings align the embeddings from various languages to the same shared embedding space which enables transfer learning by training the model in one language and adopting it for another one [AMT16, AZM19, MPC19, CHH19]. Previous work has proposed different methods to create multilingual word embeddings. One common way is to first train the monolingual word embeddings separately and then align them to the same space [CLR17, JBM18]. While multiple efforts have focused on improving the models' performance on low-resource languages, less attention is given to understanding the bias in cross-lingual transfer learning settings.

In this section, we aim to understand the bias in multilingual word embeddings. In contrast to existing literature that mostly focuses on English, we conduct analyses in multilingual settings. We argue that the bias in multilingual word embeddings can be very different from that in English. One reason is that each language has its own properties. For example, in English, most nouns do not have grammatical gender, while in Spanish, all nouns do. Second, when we do the alignment to get the multilingual word embeddings, the choice of target space may cause bias. Third, when we do transfer learning based on multilingual word embeddings, the alignment methods, as well as the transfer procedure can potentially influence the bias in downstream tasks. Our experiments confirm that bias exists in the multilingual embeddings and such bias also impacts the cross-lingual transfer learning tasks. We observe that the transfer model based on the multilingual word embeddings shows discrimination against genders. To discern such bias, we perform analysis from both the corpus and the embedding perspectives, showing that both contribute to the bias in transfer learning. Our contributions are summarized as follows:

- We build datasets for studying the gender bias in multilingual NLP systems.[9]
- We analyze gender bias in multilingual word embeddings from both intrinsic and extrinsic perspectives. Experimental results show that the pre-trained monolingual word embeddings, the alignment method as well as the transfer learning can have an impact on the gender bias.
- We show that simple mitigation methods can help to reduce the bias in multilingual word embeddings and discuss directions for future work to further study the problem. We provide several recommendations for bias mitigation in cross-lingual transfer learning.

3.3.2 Intrinsic Bias Quantification and Mitigation

In this section, we analyze the gender bias in multilingual word embeddings. Due to the limitations of the available resources in other languages, we analyze the bias in English, Spanish, German and French. However, our systematic evaluation approach can be easily extended to other languages. We first define an evaluation metric for quantifying gender bias in multilingual word embeddings. Note that in this work, we focus on analyzing gender bias from the perspective of occupations. We then show that when we change the target alignment space, the bias in multilingual word embeddings also changes. Such observations provide us a way to mitigate the bias in multilingual word embeddings – by choosing an appropriate target alignment space.

3.3.2.1 Quantifying Bias in Multilingual Embeddings

We begin with describing inBias, our proposed evaluation metric for quantifying intrinsic bias in multilingual word embeddings from word-level perspective. We then introduce the dataset we collected for quantifying bias in different languages.

[9] Code and data will be available at `https://aka.ms/MultilingualBias`.

Bias Definition Given a set of masculine and feminine words, we define inBias as:

$$\text{inBias} = \frac{1}{N} \sum_{i=1}^{N} |dis(O_{M_i}, S_M) - dis(O_{F_i}, S_F)|, \tag{3.2}$$

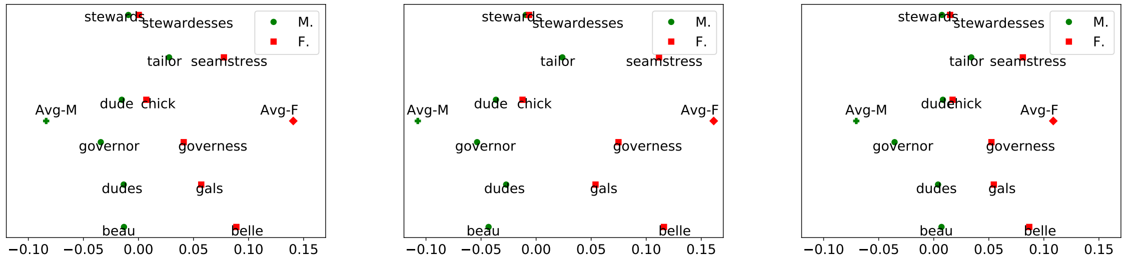
where

$$dis(O_{G_i}, S) = \frac{1}{|S|} \sum_{s \in S} (1 - \cos(O_{G_i}, s)).$$

Here (O_{M_i}, O_{F_i}) stands for the masculine and feminine format of the i-th occupation word, such as ("doctor", "doctora"). S_M and S_F are a set of gender seed words that contain male and female gender information in the definitions such as "he" or "she".

Intuitively, given a pair of masculine and feminine words describing an occupation, such as the words "*doctor*" (Spanish, masculine doctor) and "*doctora*" (Spanish, feminine doctor), the only difference lies in the gender information. As a result, they should have similar correlations to the corresponding gender seed words such as "*él*" (Spanish, he) and "*ella*" (Spanish, she). If there is a gap between the distance of occupations and corresponding gender, (i.e., the distance between "*doctor*" and "*él*" against the distance between "*doctora*" and "*ella*"), it means such occupation shows discrimination against gender. Note that such metric can also be generalized to other languages without grammatical gender, such as English, by just using the same format of the occupation words. It is also worth noting that our metric is general and can be used to define other types of bias with slight modifications. For example, it can be used to detect age or race bias by providing corresponding seed words (e.g., "*young*" - "*old*" or names correlated with different races). In this paper we focus on gender bias as the focus of study. We provide detailed descriptions of those words in the dataset collection subsection.

Unlike previous work [BCZ16] which requires calculating a gender direction by doing dimensionality reduction, we do not require such a step and hence we can keep all the information in the embeddings. The goal of inBias is aligned to that of WEAT [CBN17]. It calculates the difference of targets (occupations in our case) corresponding to different attributes (gender). We use paired occupations in each language, reducing the influence



(a) Original es embeddings.
(b) In es-en embeddings.
(c) In es-de embeddings.

Figure 3.3: Most biased occupations in ES projected to the gender subspace defined by the difference between two gendered seed words. Green dots are masculine (M.) occupations while the red squares are feminine (F.) ones. We also show the average projections of the gender seed words for male and female genders denoted by “Avg-M” and “Avg-F”. Compared to EN, aligning to DE makes the distance between the occupation word and corresponding gender more symmetric.

of grammatical gender. Compared to [ZSZ19], we do not need to separately generate the two gender directions, as in our definition, the difference of the distance already contains such information. In addition, we no longer need to collect the gender neutral word list. In multilingual settings, due to different gender assignments to each word (e.g., “spoon” is masculine is DE but feminine in ES), it is expensive to collect such resources which can be alleviated by the inBias metric.

Multilingual Intrinsic Bias Dataset To conduct the intrinsic bias analysis, we create the MIBs dataset by manually collecting pairs of occupation words and gender seed words in four languages: English (EN), Spanish (ES), German (DE) and French (FR). We choose these four languages as they come from different language families (EN and DE belong to the Germanic language family while ES and FR belong to the Italic language family) and exhibit different gender properties (e.g., in ES, FR and DE, there is grammatical gender).[10] We refer to languages with grammatical gender as gender-rich languages; and otherwise, as gender-less languages. Among these three gender-rich languages, ES and FR only have feminine and masculine genders while in DE, there is also a neutral gender. We obtain the feminine and masculine words in EN from [ZZL18] and extend them by manually adding other common occupations. The English gender seed words are from [BCZ16]. For all the other languages, we get the corresponding masculine and feminine terms by using online translation systems, such as Google Translate. We refer to the words that have both masculine and feminine formats in EN (e.g., "waiter" and "waitress") as *strong gendered* words while others like "doctor" or "teacher" as *weak gendered* words. In total, there are 257 pairs of occupations and 10 pairs of gender seed words for each language. In the gender-rich languages, if the occupation only has one lexical format, (e.g., "prosecutor" in ES only has the format "fiscal"), we add it to both the feminine and the masculine lists.

3.3.2.2 Characterizing Bias in Multilingual Embeddings

As mentioned in Sec. 3.3.1, multilingual word embeddings can be generated by first training word embeddings for different languages individually and then aligning those embeddings to the same space. During the alignment, one language is chosen as target and the embeddings from other languages are projected onto this target space. We conduct comprehensive analyses on the MIBs dataset to understand: 1) how gender bias exhibits in embeddings of different languages; 2) how the alignment target affects the gender bias in the embedding space; and 3) how the quality of multilingual embeddings is affected by choice of the target language.

For the monolingual embeddings of individual languages and the multilingual embeddings that used English as the target language (*-en),[11] we use the publicly available

[10] We also do analyses with Turkish where there is no grammatical gender and no gendered pronoun. Details are in Sec. 3.3.2.2.

[11] We refer to the aligned multilingual word embeddings using the format src-tgt. For example, "es-en"

Source	Target			
	EN	ES	DE	FR
EN	**0.0830**	0.0639*	0.0699*	0.0628*
ES	0.0889*	**0.0803**	0.0634*	0.0642*
DE	0.1124	0.0716*	**0.1079**	0.0805*
FR	0.1027	0.0768*	0.0782*	**0.0940**

Table 3.6: inBias score before and after alignment to different target spaces. Rows stands for the source languages while columns are the target languages. The diagonal values stand for the bias in the original monolingual word embeddings. Here * indicates the difference between the bias before and after alignment is statistically significant ($p < 0.05$).

fastText embeddings trained on 294 languages in Wikipedia [BGJ17, JBM18]. For all other embeddings aligned to a target space other than EN, we adopt the RCSLS alignment model [JBM18] based on the same hyperparameter setting.

Analyzing Bias before Alignment We examine the bias using four languages mentioned previously based on all the word pairs in the MIBs. Table 3.6 reports the inBias score on this dataset. The diagonal values here stand for the bias in each language before alignment. Bias commonly exists across all the four languages. Such results are also supported by WEAT in [ZSZ19], demonstrating the validity of our metric. What is more, comparing those four languages, we find DE and FR have stronger biases comparing to EN and ES.

How will the bias change when aligned to different languages? Commonly used multilingual word embeddings align all languages to the English space. However, our analysis shows that the bias in the multilingual word embeddings can change if we choose a different target space. All the results are shown in Table 3.6. Specifically, when

means we align the ES embeddings to the EN space. An embedding not following such format refers to a monolingual embedding.

Source	Target			
	EN	ES	DE	FR
EN	-	83.08	78.60	83.00
ES	86.40	-	72.40	87.27
DE	76.33	69.80	-	78.13
FR	84.27	84.80	75.53	-

Table 3.7: Performance (accuracy %) of the BLI task for the aligned embeddings. Row stands for the source language and column is the target language. The values in the first row are from [JBM18].

we align the embeddings to the gender-rich languages, the bias score will be lower compared to that in the original embedding space. In the other situation, when aligning the embeddings to the gender-less language space (i.e., EN in our case), the bias increases. For example, in original EN, the bias score is 0.0830 and when we align EN to ES, the bias decreases to 0.0639 with 23% reduction in the bias score. However, the bias in ES embeddings increases to 0.0889 when aligned to EN while only 0.0634 when aligned to DE.[12] In Fig. 3.3, we show the examples of word shifting along the gender direction when aligning ES to different languages. The gender direction is calculated by the difference of male gendered seeds and female gendered seeds. We observe the feminine occupations are further away from female seed words than masculine ones, causing the resultant bias. In comparison to using EN as target space, when aligning ES to DE, the distance between masculine and feminine occupations with corresponding gender seed words become more symmetric, therefore reducing the inBias score.

What words changed most after the alignment? We are interested in understanding how the gender bias of words changes after we do the alignment. To do this,

[12] We show the bias for all the 257 pairs of words in EN. In the appendix, we also show the bias for strong gendered words and weak gendered words separately.

we look at the top-15 most and least changed words. We find that in each language, the strongest bias comes from the strong gendered words; while the least bias happens among weak gendered words. When we align EN embeddings to gender-rich languages, bias in the strong gendered words will change most significantly; and the weak gendered words will change least significantly. When we align gender-rich languages to EN, we observe a similar trend. Among all the alignment cases, gender seed words used in Eq. (3.2) do not change significantly.

Bilingual Lexicon Induction To evaluate the quality of word embeddings after the alignment, we test them on the bilingual lexicon induction (BLI) task [CLR17] goal of which is to induce the translation of source words by looking at their nearest neighbors. We evaluate the embeddings on the MUSE dataset with the CSLS metric [CLR17].

We conduct experiments among all the pair-wise alignments of the four languages. The results are shown in Table 3.7. Each row depicts the source language, while the column depicts the target language. When aligning languages to different target spaces, we do not observe a significant performance difference in comparison to aligning to EN in most cases. This confirms the possibility to use such embeddings in downstream tasks. However, due to the limitations of available resources, we only show the result on the four languages and it may change when using different languages.

Languages of Study In this paper, we mainly focus on four European languages from different language families, partly caused by the limitations of the currently available resources. We do a simplified analysis on Turkish (TR) which belongs to the Turkic language family. In TR, there is no grammatical gender for both nouns and pronouns, i.e., it uses the same pronoun "o" to refer to "he", "she" or "it". The original bias in TR is 0.0719 and when we align it to EN, the bias remains almost the same at 0.0712. When aligning EN to TR, we can reduce the intrinsic bias in EN from 0.0830 to 0.0592, with 28.7% reduction. However, the BLI task shows that the performance on such aligned embeddings drops significantly: only 53.07% when aligned to TR but around 80% when aligned to the other four languages. Moreover, as mentioned in **?**], some other languages such as Chinese and Japanese cannot align well to English. Such situations require more investigations and forming a direction for future work.

Source	Target			
	ENDEB	ES	DE	FR
ENDEB	0.0501*	0.0458*	0.0524*	0.0441*
ES	0.0665*	0.0803	-	-
DE	0.0876*	-	0.1079	-
FR	0.0905	-	-	0.0940

Table 3.8: inBias score before and after alignment to ENDEB. * indicates statistically significant difference between the bias in original and aligned embeddings.

3.3.2.3 Bias after Mitigation

Researchers have proposed different approaches to mitigate the bias in EN word embeddings [BCZ16, ZZL18]. Although these approaches cannot entirely remove the bias [GG19], they significantly reduce the bias in English embeddings. We refer to such embedding as *ENDEB*. We analyze how the bias changes after we align the embeddings to such ENDEB space. The ENDEB embeddings are obtained by adopting the method in [BCZ16] on the original fastText monolingual word embeddings. Table 3.8 and 3.9 show the bias score and BLI performance when we do the alignment between ENDEB and other languages. Similar to [ZSZ19], we find that when we align other embeddings to the ENDEB space, we can reduce the bias in those embeddings. What is more, we show that we can reduce the bias in ENDEB embeddings further when we align it to a gender-rich language such as ES while keeping the functionality of the embeddings, which is consistent with our previous observation in Table 3.6. Besides, comparing aligning to gender-rich languages and to ENDEB, the former one can reduce the bias more.

Source	Target			
	ENDEB	ES	DE	FR
ENDEB	-	84.07	79.13	83.27

Target	Source			
	ENDEB	ES	DE	FR
ENDEB	-	86.07	76.27	84.33

Table 3.9: Performance (accuracy %) on the BLI task using the aligned embeddings based on ENDEB embeddings. The top one is the result of aligning ENDEB to other languages while the bottom is to align other languages to ENDEB.

Language	EN	ES	DE	FR
#occupation	28	72	27	27
#instance	397,907	82,863	12,976	59,490

Table 3.10: Statistics of the MLBs for each language.

3.3.3 Extrinsic Bias Quantification and Mitigation

In addition to the intrinsic bias in multilingual word embeddings, we also analyze the downstream tasks, specifically in the cross-lingual transfer learning. One of the main challenges here is the absence of appropriate datasets. To motivate further research in this direction, we build a new dataset called MLBs. Experiments demonstrate that bias in multilingual word embeddings can also have an effect on models transferred to different languages. We further show how mitigation methods can help to reduce the bias in the transfer learning setting.

3.3.3.1 Quantifying Bias in Multilingual Models

In this section, we provide details of the dataset we collected for the extrinsic bias analysis as well as the metric we use for the bias evaluation.

Multilingual BiosBias Datasets

[DRW19] built an English BiosBias dataset to evaluate the bias in predicting the occupations of people when provided with a short biography on the bio of the person written in third person. To evaluate the bias in cross-lingual transfer settings, we build the Multilingual BiosBias (MLBs) Dataset which contains bios in different languages.

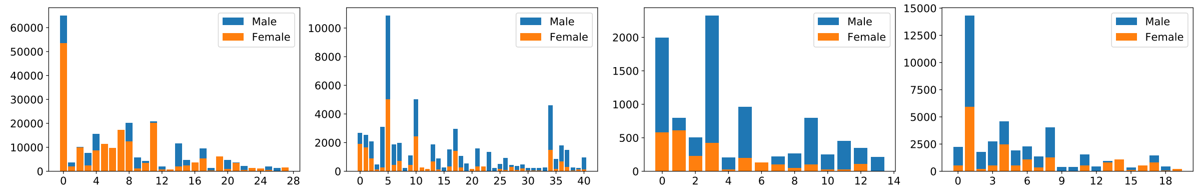
Dataset Collection Procedure We collect a list of common occupations for each language and follow the data collection procedure used for the English dataset [DRW19]. To identify bio paragraphs, we use the pattern "NAME is an OCCUPATION-TITLE" where name is recognized in each language by using the corresponding Named Entity Recognition model from spaCy.[13] To control for the same time period for datasets across languages, we process the same set of Common Crawl dumps ranging from the year 2014 to 2018. For the occupations, we use both the feminine and masculine versions of the word in the gender-rich languages. For EN, we use the existing BiosBias dataset.

The number of occupations in each language is shown in Table 3.10. As the bios are written in third person, similar to [DRW19], we extract the binary genders based on the gendered pronouns in each language, such as "he" and "she".

Bias Evaluation

We follow the method in [ZWY18] to measure the extrinsic bias: using the performance gap between different gender groups as a metric to evaluate the bias in the MLBs dataset. We split the dataset based on the gender attribute. A gender-agnostic model should have similar performance in each group. To be specific, we use the average performance gap across each occupation in the male and female groups aggregated across all occupations (|Diff| in Table 3.11) to measure the bias. However, as described in [SDH19], people's names are potentially indicative of their genders. To eliminate the influence of names

[13] https://spacy.io/usage/models



(a) EN (b) ES (c) DE (d) FR

Figure 3.4: Gender statistics of MLBs dataset for different occupations where each occupation has at least 200 instances. X-axis here stands for the occupation index and y-axis is the number of instances for each occupation. Among all the languages, EN corpus is the most gender balanced one. All the corresponding occupations will be provided in the appendix.

as well as the gender pronouns on the model predictions, we use a "scrubbed" version of the MLBs dataset by removing the names and some gender indicators (e.g., gendered pronouns and prefixes such as "*Mr.*" or "*Ms.*").

To make predictions of the occupations, we adopt the model used in [DRW19] by taking the fastText embeddings as the input and encoding the bio text with bi-directional GRU units following by an attention mechanism. The predictions are generated by a softmax layer. We train such models using standard cross-entropy loss and keep the embeddings frozen during the training.

3.3.3.2 Characterizing Bias in Multilingual Models

In this section, we analyze the bias in the multilingual word embeddings from the extrinsic perspective. We show that bias exists in cross-lingual transfer learning and the bias in multilingual word embeddings contributes to such bias.

The gender distribution of the MLBs dataset is shown in Fig. 3.4. Among the three languages, EN corpus is most gender neutral one where the ratio between male and female instances is around $1.2 : 1$. For all the other languages, male instances are far larger than

MLBs	Emb.	Avg.	Female	Male	|Diff|
EN	en	82.82	84.69	80.70	**7.26**
	endeb	83.00	84.71	81.06	6.09 ↓
	en-es	83.43	85.14	81.51	6.72 ↓
	en-de	82.85	84.64	80.84	6.37 ↓
	en-fr	82.66	84.34	80.78	5.87 ↓
ES	es	63.83	64.47	63.56	6.56
	es-en	61.47	61.42	61.49	**7.13** ↑
	es-endeb	61.91	62.98	61.45	5.61 ↓
	es-de	61.61	62.82	61.11	5.51 ↓
	es-fr	62.91	63.31	62.73	4.32 ↓

Table 3.11: Results on scrubbed MLBs. "Emb." stands for the embeddings used in model training. "Avg.", "Female" and "Male" refer to the overall average accuracy (%), and average accuracy for different genders respectively. " |Diff|" stands for the average absolute accuracy gap between each occupation in the male and female groups aggregated across all the occupations. The results of FR and DE are in the original paper.

female ones. In ES, the ratio between male and female is $2.7 : 1$, in DE it is $3.53 : 1$, and in FR, it is $2.5 : 1$; all are biased towards the male gender.

Bias in Monolingual BiosBias We first evaluate the bias in the MLBs monolingual dataset by predicting the occupations of the bios in each language.[14] From Table 3.11 we observe that: 1) Bias commonly exists across all languages ($|\text{Diff}| > 0$) when using different aligned embeddings, meaning that the model works differently for male and female groups. 2) When training the model using different aligned embeddings, it does not affect the overall average performance significantly ("Avg." column in the table). 3) The alignment direction influences the bias. On training the model based on the embeddings

[14] The results of DE and FR are in the original paper.

Trans.	Src.	Tgt.	Avg.	Female	Male	|Diff|
EN→ES	en	es-en	41.68	42.29	41.42	2.83
	en-es	es	34.15	33.97	34.22	3.49
ES→EN	es	en-es	57.33	59.61	54.75	8.33
	es-en	en	57.05	59.32	54.47	10.13

Table 3.12: Results of transfer learning on the scrubbed MLBs. "Src." and "Tgt." stand for the embeddings in source model and fine tuning procedure respectively.

aligned to different target space, we find that aligning the embeddings to ENDEB or a gender-rich language reduces the bias in the downstream task. This is aligned with our previous observation in Section 3.3.2.

Bias in Transfer Learning Multilingual word embeddings are widely used in cross-lingual transfer learning [RVS19]. In this section, we conduct experiments to understand how the bias in multilingual word embeddings impacts the bias in transfer learning. To do this, we train our model in one language (i.e., source language) and transfer it to another language based on the aligned embeddings obtained in Section 3.3.2.2. For the transfer learning, we train the model on the training corpus of the source language and randomly choose 20% of the dataset from the target language and use them to fine-tune the model.[15] Here, we do not aim at achieving state-of-the-art transfer learning performance but pay more attention to the bias analysis. Table 3.12 shows that the bias is present when we do the transfer learning regardless of the direction of transfer learning.

Bias from Multilingual Word Embeddings The transfer learning bias in Table 3.12 is a combined consequence of both corpus bias and the multilingual word embedding bias. To better understand the influence of the bias in multilingual word embeddings on the transfer learning, we make the training corpus gender balanced for each occupation by

[15] As there are fewer examples in DE, we use the whole datasets for transfer learning.

Trans.	Src.	Tgt.	Avg.	Female	Male	|Diff|
EN→ES	en	es-en	39.17	41.30	38.70	**7.97**
	en-es	es	35.66	36.11	35.47	4.53
	en-de	es-de	34.12	34.46	33.98	4.07
	en-fr	es-fr	37.63	38.75	37.16	4.87
ES→EN	es	en-es	58.41	61.78	54.60	9.03
	es-en	en	55.62	58.00	52.93	**9.52**
	es-de	en-de	57.98	60.47	55.17	9.13
	es-fr	en-fr	55.04	57.85	51.86	8.47

Table 3.13: Results of transfer learning on gender balanced scrubbed MLBs. The bias in the last column demonstrates that the bias in the multilingual word embeddings also influences bias in transfer learning.

Trans.	Src.	Tgt.	Avg.	Female	Male	|Diff|
EN→ES	endeb	es-endeb	37.44	39.90	36.40	5.93
ES→EN	es-endeb	endeb	52.51	54.45	50.03	9.06

Table 3.14: Bias mitigation results of transfer learning when we aligned the embeddings to the ENDEB space on gender balanced scrubbed MLBs.

upsampling to approximately make the model free of the corpus bias. We then test the bias for different languages with differently aligned embeddings. The results are shown in Table 3.13. When we adopt the embeddings aligned to gender-rich languages, we could reduce the bias in the transfer learning, whereas adopting the embeddings aligned to EN results in an increased bias.

Bias after Mitigation Inspired by the method in [ZWY18], we mitigate the bias in the downstream tasks by adopting the bias-mitigated word embeddings. To get the less biased multilingual word embeddings, we align other embeddings to the ENDEB space

MLBs	Avg.	Female	Male	|Diff|
EN	84.35	85.54	83.01	7.31
ES	67.93	65.79	68.82	4.16
DE	72.68	73.68	72.28	4.89
FR	79.18	78.80	79.35	8.75

Table 3.15: Bias in monolingual MLBs using M-BERT.

Trans.	Avg.	Female	Male	|Diff|
EN→ES	66.56	65.70	66.92	5.48
EN→DE	76.21	75.66	76.42	7.51
EN→FR	76.46	75.73	76.81	8.97

Table 3.16: Bias in MLBs using M-BERT when transferring from EN to other languages. Comparing to multilingual word embeddings, M-BERT achieves better transfer performance on the MLBs dataset across different languages. But the bias can be higher comparing to the multilingual word embeddings.

previously obtained in Section 3.3.2. Table 3.14 demonstrates that by adopting such less biased embeddings, we can reduce the bias in transfer learning. Comparing to Table 3.13, aligning the embeddings to a gender-rich language achieves better bias mitigation and, at the same time, remains the overall performance.

3.3.3.3 Bias Analysis Using Contextualized Embeddings

Contextualized embeddings such as ELMo [PNI18], BERT [DCL18] and XLNet [YDY19] have shown significant performance improvement in various NLP applications. Multilingual BERT (M-BERT) has shown its great ability for the transfer learning. As M-BERT provides one single language model trained on multiple languages, there is no longer a need for alignment procedure. In this section, we analyze the bias in monolingual MLBs dataset as well as in transfer learning by replacing the fastText embeddings with M-BERT embeddings. Similar to previous experiments, we train the model on the English dataset and transfer to other languages. Table 3.15 and 3.16 summarizes our results: comparing to results by fastText embeddings in Table 3.11, M-BERT improves the performance on monolingual MLBs dataset as well as the transfer learning tasks. When it comes to the bias, using M-BERT gets similar or lower bias in the monolingual datasets, but sometimes achieves higher bias than the multilingual word embeddings in transfer learning tasks such as the EN → ES (in Table 3.12).

3.3.4 Discussion

Recently bias in embeddings has attracted much attention. However, most of the work only focuses on English corpora and little is known about the bias in multilingual embeddings. In this section, we build different metrics and datasets to analyze gender bias in the multilingual embeddings from both the intrinsic and extrinsic perspectives. We show that gender bias commonly exists across different languages and the alignment target for generating multilingual word embeddings also affects such bias. In practice, we can choose the embeddings aligned to a gender-rich language to reduce the bias.

However, due to the limitation of available resources, this study is limited to the European languages. We hope this study can work as a foundation to motivate future research about the analysis and mitigation of bias in multilingual embeddings. We encourage researchers to look at languages with different grammatical gender (such as Czech and Slovak) and propose new methods to reduce the bias in multilingual embeddings as well as in cross-lingual transfer learning

3.4 Discussion

In this chapter, we review the bias problem in an important NLP component – the embeddings. We propose a new embedding learning schema to disentangle the protected attributes in certain dimensions in the embedding space (Sec. 3.1); we reveal the bias problems in the contextualized word embeddings (Sec. 3.2) and demonstrate the bias in multilingual word embeddings and its impact on cross-lingual transfer learning (Sec. 3.3).

CHAPTER 4

Bias Evaluation Metric

Machine learning techniques have been widely used in NLP. However, as revealed by many recent studies, machine learning models often inherit and amplify the societal biases in data. Various metrics have been proposed to quantify biases in model predictions. In particular, several of them evaluate disparity in model performance between protected groups and advantaged groups in the test corpus. However, we argue that evaluating bias at the corpus level is not enough for understanding how biases are embedded in a model. In fact, a model with similar aggregated performance between different groups on the entire data may behave differently on instances in a local region. To analyze and detect such *local bias*, we propose LOGAN, a new bias detection technique based on clustering. Experiments on toxicity classification and object classification tasks show that LOGAN identifies bias in a local region and allows us to better analyze the biases in model predictions. This chapter is from our work [ZC20].

4.1 Introduction

Machine learning models such as deep neural networks have achieved remarkable performance in many NLP tasks. However, as noticed by recent studies, these models often inherit and amplify the biases in the datasets used to train the models [ZWY17, BCZ16, CBN17, ZSZ19, MYB19, BBD20].

To quantify bias, researchers have proposed various metrics to study algorithmic fairness at both individual and group levels. The former measures if a model treats similar individuals consistently no matter which groups they belong to, while the latter requires the model to perform similarly for protected groups and advantaged groups in the corpus.[1] In this paper, we argue that studying algorithmic fairness at either level does not tell the full story. A model that reports similar performance across two groups in a corpus may behave differently between these two groups in a local region.

For example, the performance gap of a toxicity classifier for sentences mentioning black and white race groups is 4.8%.[2] This gap is only marginally larger than the performance gap of 2.4% when evaluating the model on two randomly split groups. However, if we evaluate the performance gap on the sentences containing the token "racist", the performance gap between these two groups is as large as 19%. Similarly, [ZWY17] report that a visual semantic role labeling system tends to label an image depicting cooking as *woman cooking* than *man cooking*. However, the model is, in fact, more likely to produce an output of *man cooking* when the agent in the image wears a chef hat. We call these biases exhibited in a neighborhood of instances **local group bias** in contrast with **global group bias** which is evaluated on the entire corpus.

To detect *local group bias*, we propose LOGAN, a LOcal Group biAs detectioN algorithm to identify biases in local regions. LOGAN adapts a clustering algorithm (e.g., K-Means) to group instances based on their features while maximizing a bias metric (e.g., performance gap across groups) within each cluster. In this way, local group bias is highlighted, allowing a developer to further examine the issue.

Our experiments on toxicity classification and MS-COCO object classification demonstrate the effectiveness of LOGAN. We show that besides successfully detecting local group bias, our method also provides interpretations for the detected bias. For example, we find that different topics lead to different levels of local group bias in the toxicity classification.

[1] For example, [ZWY18] and [RNL18] evaluate the bias in coreference resolution systems by measuring the difference in F_1 score between cases where a gender pronoun refers to an occupation stereotypical to the gender and the opposite situation.

[2] Performance in accuracy on the unintended bias detection task [Con19]

4.2 Methodology

In this section, we first provide formal definitions of local group bias and then the details of the detection method LOGAN.

Performance Disparity Assume we have a trained model f and a test corpus $\mathcal{D} = \{(x_i, y_i)\}_{i=1\ldots n}$ that is used to evaluate the model. Let $P_f(\mathcal{D})$ represents the performance of the model f evaluated on the corpus $\mathcal{D}$. Based on the applications, the performance metric can be accuracy, AUC, false positive rates, etc. For the sake of simplicity, we assume each input example x_i is associated with one of demographic groups (e.g., male or female), i.e., $x_i \in \mathcal{A}_1$ or $x_i \in \mathcal{A}_2$.[3] As a running example, we take performance disparity as the bias metric. That is, if $\|P_f(\mathcal{A}_1) - P_f(\mathcal{A}_2)\| > \epsilon$, then we consider that the model exhibits bias, where ϵ is a given threshold.

Definition of local group bias We define *local group bias* as the bias exhibits in certain local region of the test examples. Formally, given a centroid c in the input space, let $\mathcal{A}_1^c = \{x \in \mathcal{A}_1 | \|x - c\|^2 < \gamma\}$ and $\mathcal{A}_2^c = \{x \in \mathcal{A}_2 | \|x - c\|^2 < \gamma\}$ be the neighbor instances of c in each group, where γ is a threshold. We call a model has local group bias if

$$\|P_f(\mathcal{A}_1^c) - P_f(\mathcal{A}_2^c)\| > \epsilon. \tag{4.1}$$

While this definition is based on performance disparity, it is straightforward to extend the notion of local group bias to other bias metrics.

LOGAN The goal of LOGAN is to cluster instances in $\mathcal{D}$ such that (1) similar examples are grouped together, and (2) each cluster demonstrates local group bias contained in f. To achieve this goal, LOGAN generates cluster $\mathcal{C} = \{C_{i,j}\}_{i=1\ldots n, j=1\ldots k}$ by optimizing the

[3] In this paper, we consider only binary attributes such as gender = {male, female}, race = {white, black}. However, our approach is general and can be incorporated with any bias metric presented as a loss function. Therefore, it can be straightforwardly extended to a multi-class case by plugging the corresponding bias metric.

following objective:

$$\min_{\mathcal{C}} L_c + \lambda L_b, \tag{4.2}$$

where L_c is the clustering loss and L_b is local group bias loss. $\lambda \geq 0$ is a hyper-parameter to control the trade-offs between the two objectives. $C_{ij} = 1$ if x_i is assigned to the cluster j; $C_{ij} = 0$ otherwise. We introduce these two loss terms in the following.

Clustering objective The loss L_c is derived from a standard clustering technique. In this paper, we consider the K-Means clustering method [Llo82]. Specifically, the loss L_c of K-Means is

$$L_c = \sum_{j=1}^{k} \sum_{i=1}^{n} \|C_{ij} x_i - \mu_j\|^2 \quad \forall i, \sum_{j=1}^{k} C_{ij} = 1, \tag{4.3}$$

$\mu_j = (\sum_{ij} C_{ij} x_i) / \sum_{i,j} C_{ij}$ is the mean of cluster j. Note that our framework is general and other clustering techniques, such as Spectral clustering [SM00], DBSCAN [EKS96], or Gaussian mixture model can also be applied in generating the clusters. Besides, the features used for creating the clusters can be different from the features used in the model f.

Local group bias objective For the local group bias loss L_b, the goal is to obtain a clustering that maximizes the bias metric within each cluster. In the following descriptions, we take the performance gap between different attributes (see Eq. (4.1)) as an example to describe the bias metric.

Let $\hat{y}_i = f(x_i)$ be the prediction of f on x_i. The local group bias loss L_b is defined as the negative summation of performance gaps over all the clusters. If accuracy is used as the performance evaluation metric, $L_b =$

$$-\sum_{j=1}^{k} \left| \frac{\sum_{x_i \in \mathcal{A}_1} C_{ij} \mathcal{I}_{\hat{y}_i = y_i}}{\sum_{x_i \in \mathcal{A}_1} C_{ij}} - \frac{\sum_{x_i \in \mathcal{A}_2} C_{ij} \mathcal{I}_{\hat{y}_i = y_i}}{\sum_{x_i \in \mathcal{A}_2} C_{ij}} \right|^2,$$

where $\mathcal{I}$ is the indicator function.

Similar to K-Means algorithm, we solve Eq. (4.2) by iterating two steps: first, assign x_i to its closest cluster j based on current μ_j; second, update μ_j based on current label assignment. We use k-means++ [AV07] for the cluster initialization and stop when the model converges or reaches enough iterations. To make sure each cluster contains enough instances, in practice, we choose a large k ($k = 10$ in our case) and merge a small cluster to its closest neighbor. [4] For local group bias detection, we only consider clusters with at least 20 examples from each group.

4.3 Experiment

In this section, we show that LOGAN is capable of identifying local group bias, and the clusters generated by LOGAN provide an insight into how bias is embedded in the model.

4.3.1 Toxicity Classification

This task aims at detecting whether a comment is toxic (e.g. abusive or rude). Previous work has demonstrated that this task is biased towards specific identities such as "gay" [DLS18]. In our work, we use toxicity classification as one example to detect local group bias in texts and show that such local group bias could be caused by different topics in the texts.

Dataset We use the official train and test datasets from [Con19]. As the dataset is extremely imbalanced, we down-sample the training dataset and reserve 20% of it as the development set. In the end, we have $204,000$, $51,000$ and $97,320$ examples for train, development and test, respectively. We tune $\lambda = \{1, 5, 10, 100\}$ and choose the one with the largest number of clusters showing local group bias.

Model We fine-tune a BERT sequence classification model from [WDS19] for 2 epochs with a learning rate 2×10^{-5}, max sequence length 220 and batch size 20. The model achieves 90.2% accuracy on the whole test dataset.[5] We use sentence embeddings from

[4] We merge the clusters iteratively and stop the procedure when all the clusters have at least 20 examples or only 5 clusters are left.

[5] The source code is available at `https://github.com/uclanlp/clusters`.

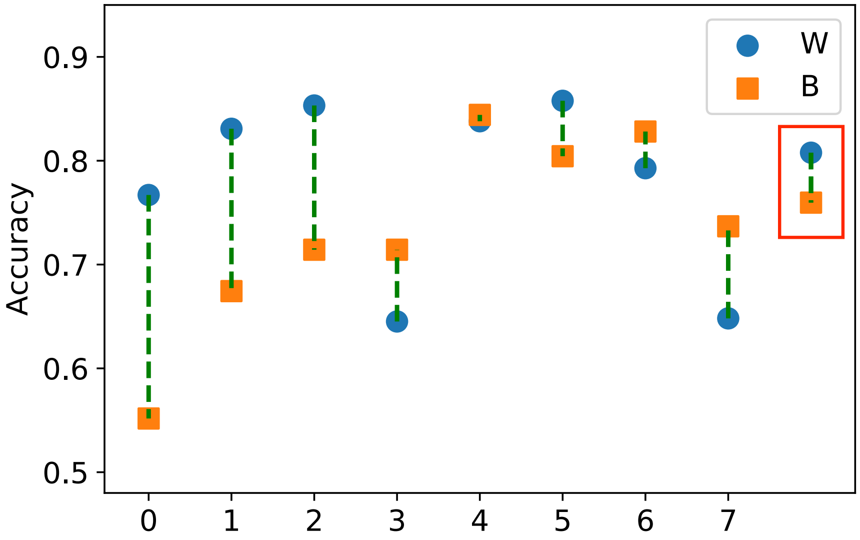


Figure 4.1: Accuracy for White (blue circle) and Black (orange square) groups in each cluster using LOGAN. The length of the dashed line shows the gap. Red box highlights the accuracy of these two groups on the entire corpus. Clusters 0 and 1 demonstrate strong local group bias. Full results are in the original paper.

the second to last layer of a pre-trained BERT model as features to perform clustering.

Bias Detection There are several demographic groups in the toxic dataset such as gender, race and religion. We focus on the binary gender (male/female) and binary race (black/white) in the experiments. For local group bias, we report the largest bias score among all the clusters. Figure 4.1 shows the accuracy of white and black groups in each cluster using LOGAN. The example bounded in the red box is the global accuracy of these two groups. Based on the results in Figure 4.1 and Table 4.1, we only detect weak global group bias in the model predictions. However, both K-Means and LOGAN successfully detect strong local group bias. In particular, LOGAN identifies a local region that the model has difficulties in making correct predictions for female group.

While we use the gap of accuracy as the bias metric, the clusters detected by LOGAN also exhibit local bias when evaluating using other metrics. Table 4.2 shows the gap of

	Method	Acc-W	Acc-B	|Bias|
Race	Global	80.8	76.0	4.8
	K-Means	75.9	53.8	**22.1**
	LOGAN	76.7	55.2	**21.5**
	Method	Acc-M	Acc-F	|Bias|
Gender	Global	79.8	81.6	1.8
	K-Means	70.2	82.8	**12.6**
	LOGAN	80.2	57.1	**23.1**

Table 4.1: Bias detection in toxic classification. Results are shown in %. “Global” stands for global group bias detection. W, B, M, F refer to White, Black, Male and Female groups respectively.

subgroup AUC scores over the clusters. Similar to the results in Table 4.1, K-Means and LOGAN detect local group bias. In particular, the first and the third clusters in Figure 4.1 also have larger AUC disparity than the global AUC gap. Similarly, the first three clusters in Figure 4.1 have a significantly larger gap of False Positive Rate across different groups than when evaluating on the entire dataset.

Bias Interpretation To better interpret the local group bias, we run a Latent Dirichlet Allocation topic model [BNJ03] to discover the main topic of each cluster. Table 4.3 lists the top 20 topic words for the most and least biased clusters using LOGAN under race attributes. We remove the words related to race attributes such as “white” and “black”. We find that different topics in each cluster may lead to different levels of local group bias. For example, compared with the less biased group, the most biased group includes a topic on supremacy.

Comparison between K-Means and LOGAN We compare LOGAN with K-Means using the following 3 metrics. “Inertia” sums over the distances of all instances to their

	Method	AUC-W	AUC-B	|Bias|
Race	Global	0.870	0.846	0.024
	K-Means	0.836	0.679	**0.157**
	LOGAN	0.844	0.691	**0.153**
	Method	AUC-M	AUC-F	|Bias|
Gender	Global	0.896	0.924	0.028
	K-Means	0.828	0.922	**0.094**
	LOGAN	0.910	0.818	**0.092**

Table 4.2: Bias detection using subgroup AUC. "Global" stands for global group bias detection. W, B, M, F refer to White, Black, Male and Female groups respectively.

Most Biased (21.5)	trump supremacist supremacists kkk
	people party america racist
	president support vote sessions
	voters republican said obama
	man base bannon nationalists
Least Biased (0.6)	people like get think know
	say men see racist way
	good point right go person
	well make time said much

Table 4.3: Top 20 topic words in the most and least biased cluster using LOGAN under RACE attributes. Number in parentheses is the bias score (%) of that cluster.

closest centers which is used to measure the clustering quality. We normalize it to make the inertia of K-Means 1.0. To measure the utility of local group bias detection, we look at

	Inertia	BCR	BIR	$\lvert$Bias$\rvert$
K-Means	1.0	62.5%	58.2%	12.4%
LOGAN	1.002	75.0%	71.8%	12.0%

Table 4.4: Comparison between K-Means and LOGAN under race attributes. " BCR" and "BIR" refer to the ratio of biased clusters and ratio of instances in those biased clusters, respectively. "$\lvert$Bias$\rvert$" here is the averaged absolute bias score for those biased clusters.

the ratio of clusters showing a bias score at least 5%[6] (BCR) as well as the ratio of instances within those biased clusters (BIR). Table 4.4 shows that LOGAN increases the ratio of clusters exhibiting non-trivial local group bias by a large margin with trivial trade offs in inertia.

4.3.2 Object Classification

We conduct experiments on object classification using MS-COCO [LMB14]. Given one image, the goal is to predict if one object appears in the image. Following the setup in [WZY19], we exclude person from the object labels.

Dataset Similar to [ZWY17] and [WZY19], we extract the gender label for one image by looking at the captions. For our analysis, we only consider images with gender labels. In the end, there are $22,800$, $5,400$ and $5,400$ images left for train, development and test, respectively.

Model We use the basic model from [WZY19] for this task, which adapts a standard ResNet-50 pre-trained on ImageNet with the last layer modified. We follow the default hyper-parameters of the original model.

[6] We choose 5% as it is close to the averaged bias score plus standard deviation when we randomly split the examples into two groups over 5 runs.

Bias Detection and Interpretation We evaluate bias in the predictions of the object classification model by looking at the accuracy gap between male and female groups for each object. In the analysis, we only consider objects with more than 100 images in the test set. This results in a total of 26 objects. Among the three methods, Global can only detect group bias at threshold 5% (i.e., performance gap $\geq 5\%$) for 14 objects, while K-Means and LOGAN increase the number to 19 and 21 respectively.

Comparing LOGAN with K-Means, among all the 26 objects, the average inertia is almost the same (the ratio is 1.001). On average, 34.0% and 35.7% of the clusters showing local group bias at threshold 5% (i.e. BCR) and the ratio of instances in those biased clusters (i.e., BIR) are 57.7% and 54.9% for K-Means and LOGAN, respectively.

We further investigate the local groups discovered by LOGAN by comparing the images in the less biased local groups with the strong biased ones. We find that, for example, in the most biased local groups, the images often contain "handbag" with a street scene. In such a case, the model is more likely to correctly predict the agent in the image is woman.

4.4 Discussion

Machine learning models risk inheriting the underlying societal biases from the data. In practice, many works use the global performance gap between different groups as a metric to detect the bias. In this work, we revisit the coarse-grained metric for group bias analysis and propose a new method, LOGAN, to detect local group bias by clustering. Our method can help detect model biases that previously are hidden from the global bias metrics and provide an explanation of such biases. But we notice there are some limitations in LOGAN. For example, the number of instances in clusters could be uneven.

CHAPTER 5

Bias Amplification

In previous chapters, we have shown that NLP models can implicitly learn the bias from the training dataset. In this chapter, we use one vision-and-language task as a running example to demonstrate besides mimicking the biases in the training corpus, a model can potentially enlarge those biases causing more severe problems. In the end, we show two possible ways to calibrate such bias amplifications without model retraining. This chapter is based on our work [ZWY17, JMZ20].

5.1 Bias Amplification in Top Prediction

5.1.1 Introduction

Visual recognition tasks involving language, such as captioning [VTB15], visual question answering [AAL15], and visual semantic role labeling [YZF16], have emerged as avenues for expanding the diversity of information that can be recovered from images. These tasks aim at extracting rich semantics from images and require large quantities of labeled data, predominantly retrieved from the web. Methods often combine structured prediction and deep learning to model correlations between labels and images to make judgments that otherwise would have weak visual support. For example, in the first image of Figure 5.1, it is possible to predict a `spatula` by considering that it is a common tool used for the activity `cooking`. Yet such methods run the risk of discovering and exploiting societal biases present in the underlying web corpora. Without properly quantifying and reducing the reliance on such correlations, broad adoption of these models can have the inadvertent

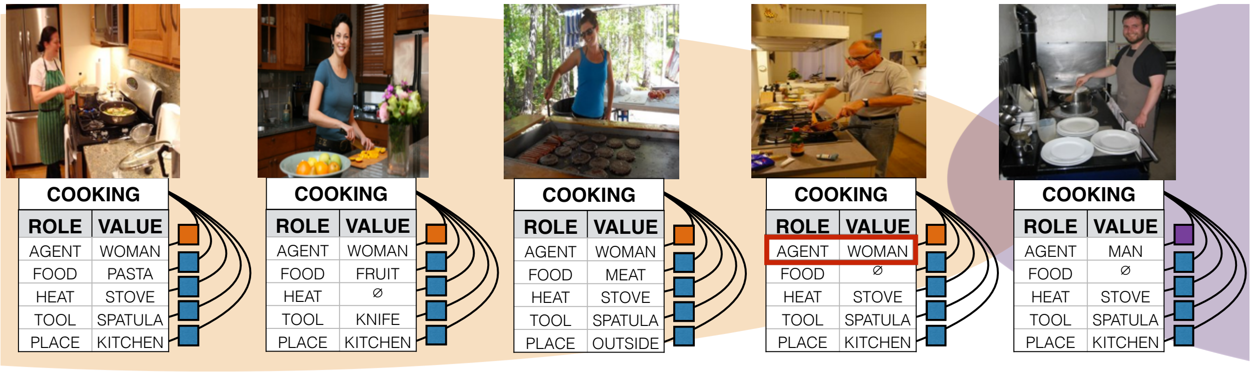


Figure 5.1: Five example images from the imSitu visual semantic role labeling (vSRL) dataset. Each image is paired with a table describing a situation: the verb, `cooking`, its semantic roles, i.e `agent`, and noun values filling that role, i.e. `woman`. In the imSitu training set, 33% of `cooking` images have `man` in the `agent` role while the rest have `woman`. After training a Conditional Random Field (CRF), bias is amplified: `man` fills 16% of `agent` roles in `cooking` images. To reduce this bias amplification our calibration method adjusts weights of CRF potentials associated with biased predictions. After applying our methods, `man` appears in the `agent` role of 20% of `cooking` images, reducing the bias amplification by 25%, while keeping the CRF vSRL performance unchanged.

effect of magnifying stereotypes.

In this paper, we develop a general framework for quantifying bias and study two concrete tasks, visual semantic role labeling (vSRL) and multilabel object classification (MLC). In vSRL, we use the imSitu formalism [YZF16, YOZ17], where the goal is to predict activities, objects and the roles those objects play within an activity. For MLC, we use MS-COCO [LMB14, CFL15], a recognition task covering 80 object classes. We use gender bias as a running example and show that both supporting datasets for these tasks are biased with respect to a gender binary[1].

[1] To simplify our analysis, we only consider a gender binary as perceived by annotators in the datasets. We recognize that a more fine-grained analysis would be needed for deployment in a production system. Also, note that the proposed approach can be applied to other NLP tasks and other variables such as identification with a racial or ethnic group.

Our analysis reveals that over 45% and 37% of verbs and objects, respectively, exhibit bias toward a gender greater than 2:1. For example, as seen in Figure 5.1, the `cooking` activity in imSitu is a heavily biased verb. Furthermore, we show that after training state-of-the-art structured predictors, models amplify the existing bias, by 5.0% for vSRL, and 3.6% in MLC.

To mitigate the role of bias amplification when training models on biased corpora, we propose a novel constrained inference framework, called **RBA**, for **R**educing **B**ias **A**mplification in predictions. Our method introduces corpus-level constraints so that gender indicators co-occur no more often together with elements of the prediction task than in the original training distribution. For example, as seen in Figure 5.1, we would like noun `man` to occur in the `agent` role of the `cooking` as often as it occurs in the imSitu training set when evaluating on a development set. We combine our calibration constraint with the original structured predictor and use Lagrangian relaxation [KV08, RC12] to reweigh bias creating factors in the original model.

We evaluate our calibration method on imSitu vSRL and COCO MLC and find that in both instances, our models substantially reduce bias amplification. For vSRL, we reduce the average magnitude of bias amplification by 40.5%. For MLC, we are able to reduce the average magnitude of bias amplification by 47.5%. Overall, our calibration methods do not affect the performance of the underlying visual system, while substantially reducing the reliance of the system on socially biased correlations[2].

5.1.2 Visualizing and Quantifying Biases

Modern statistical learning approaches capture correlations among output variables in order to make coherent predictions. However, for real-world applications, some implicit correlations are not appropriate, especially if they are amplified. In this section, we present a general framework to analyze inherent biases learned and amplified by a prediction model.

[2] Code and data are available at `https://github.com/uclanlp/reducingbias`

Identifying bias We consider that prediction problems involve several inter-dependent output variables $y_1, y_2, ...y_K$, which can be represented as a structure $y = \{y_1, y_2, ...y_K\} \in Y$. This is a common setting in NLP applications, including tagging, and parsing. For example, in the vSRL task, the output can be represented as a structured table as shown in Fig 5.1. Modern techniques often model the correlation between the sub-components in y and make a joint prediction over them using a structured prediction model. More details will be provided in Section 5.1.3.

We assume there is a subset of output variables $g \subseteq y, g \in G$ that reflects demographic attributes such as gender or race (e.g. $g \in G = \{\texttt{man}, \texttt{woman}\}$ is the agent), and there is another subset of the output $o \subseteq y, o \in O$ that are co-related with g (e.g., o is the activity present in an image, such as `cooking`). The goal is to identify the correlations that are potentially amplified by a learned model.

To achieve this, we define the bias score of a given output, o, with respect to a demographic variable, g, as:

$$b(o, g) = \frac{c(o, g)}{\sum_{g' \in G} c(o, g')},$$

where $c(o, g)$ is the number of occurrences of o and g in a corpus. For example, to analyze how genders of agents and activities are co-related in vSRL, we define the gender bias toward `man` for each verb $b(verb, \texttt{man})$ as:

$$\frac{c(verb, \texttt{man})}{c(verb, \texttt{man}) + c(verb, \texttt{woman})}. \tag{5.1}$$

If $b(o, g) > 1/\|G\|$, then o is positively correlated with g and may exhibit bias.

Evaluating bias amplification To evaluate the degree of bias amplification, we propose to compare bias scores on the training set, $b^*(o, g)$, with bias scores on an unlabeled evaluation set of images $\tilde{b}(o, g)$ that has been annotated by a predictor. We assume that the evaluation set is identically distributed to the training set. Therefore, if o is positively correlated with g (i.e, $b^*(o, g) > 1/\|G\|$) and $\tilde{b}(o, g)$ is larger than $b^*(o, g)$, we say bias has been amplified. For example, if $b^*(\texttt{cooking}, \texttt{woman}) = .66$, and $\tilde{b}(\texttt{cooking}, \texttt{woman}) = .84$, then the bias of `woman` toward `cooking` has been amplified. Finally, we define the mean bias amplification as:

$$\frac{1}{|O|}\sum_{g}\sum_{o\in\{o\in O|b^*(o,g)>1/\|G\|\}}\tilde{b}(o,g)-b^*(o,g).$$

This score estimates the average magnitude of bias amplification for pairs of o and g which exhibited bias.

5.1.3 Calibration Algorithm

In this section, we introduce **R**educing **B**ias **A**mplification, RBA, a debiasing technique for calibrating the predictions from a structured prediction model. The intuition behind the algorithm is to inject constraints to ensure the model predictions follow the distribution observed from the training data. For example, the constraints added to the vSRL system ensure the gender ratio of each verb in Eq. (5.1) are within a given margin based on the statistics of the training data. These constraints are applied at the corpus level, because computing gender ratio requires the predictions of all test instances. As a result, a joint inference over test instances is required[3]. Solving such a giant inference problem with constraints is hard. Therefore, we present an approximate inference algorithm based on Lagrangian relaxation. The advantages of this approach are:

- Our algorithm is iterative, and at each iteration, the joint inference problem is decomposed to a per-instance basis. This can be solved by the original inference algorithm. That is, our approach works as a meta-algorithm and developers do not need to implement a new inference algorithm.
- The approach is general and can be applied in any structured model.

[3] A sufficiently large sample of test instances must be used so that bias statistics can be estimated. In this work we use the entire test set for each respective problem.

- Lagrangian relaxation guarantees the solution is optimal if the algorithm converges and all constraints are satisfied.

In practice, it is hard to obtain a solution where all corpus-level constrains are satisfied. However, we show that the performance of the proposed approach is empirically strong. We use imSitu for vSRL as a running example to explain our algorithm.

Structured Output Prediction As we mentioned in Sec. 5.1.2, we assume the structured output $y \in Y$ consists of several sub-components. Given a test instance i as an input, the inference problem is to find

$$\arg\max_{y \in Y} \quad f_\theta(y, i),$$

where $f_\theta(y, i)$ is a scoring function based on a model θ learned from the training data. The structured output y and the scoring function $f_\theta(y, i)$ can be decomposed into small components based on an independence assumption. For example, in the vSRL task, the output y consists of two types of binary output variables $\{y_v\}$ and $\{y_{v,r}\}$. The variable $y_v = 1$ if and only if the activity v is chosen. Similarly, $y_{v,r} = 1$ if and only if both the activity v and the semantic role r are assigned [4]. The scoring function $f_\theta(y, i)$ is decomposed accordingly such that:

$$f_\theta(y, i) = \sum_v y_v s_\theta(v, i) + \sum_{v,r} y_{v,r} s_\theta(v, r, i),$$

represents the overall score of an assignment, and $s_\theta(v, i)$ and $s_\theta(v, r, i)$ are the potentials of the sub-assignments. The output space Y contains all feasible assignments of y_v and $y_{v,r}$, which can be represented as instance-wise constraints. For example, the constraint, $\sum_v y_v = 1$ ensures only one activity is assigned to one image.

Corpus-level Constraints Our goal is to inject constraints to ensure the output labels follow a desired distribution. For example, we can set a constraint to ensure the gender

[4] We use r to refer to a combination of role and noun. For example, one possible value indicates an `agent` is a `woman`.

ratio for each activity in Eq. (5.1) is within a given margin. Let $y^i = \{y^i_v\} \cup \{y^i_{v,r}\}$ be the output assignment for test instance i[5]. For each activity v^*, the constraints can be written as

$$b^* - \gamma \leq \frac{\sum_i y^i_{v=v^*,r\in M}}{\sum_i y^i_{v=v^*,r\in W} + \sum_i y^i_{v=v^*,r\in M}} \leq b^* + \gamma \tag{5.2}$$

where $b^* \equiv b^*(v^*, man)$ is the desired gender ratio of an activity v^*, γ is a user-specified margin. M and W are a set of semantic role-values representing the agent as a `man` or a `woman`, respectively.

Note that the constraints in (5.2) involve all the test instances. Therefore, it requires a joint inference over the entire test corpus. In general, these corpus-level constraints can be represented in a form of $A\sum_i y^i - b \leq 0$, where each row in the matrix $A \in R^{l\times K}$ is the coefficients of one constraint, and $b \in R^l$. The constrained inference problem can then be formulated as:

$$\begin{aligned} \max_{\{y^i\}\in\{Y^i\}} \quad & \sum_i f_\theta(y^i, i), \\ \text{s.t.} \quad & A\sum_i y^i - b \leq 0, \end{aligned} \tag{5.3}$$

where $\{Y^i\}$ represents a space spanned by possible combinations of labels for all instances. Without the corpus-level constraints, Eq. (5.3) can be optimized by maximizing each instance i

$$\max_{y_i \in Y^i} f_\theta(y^i, i),$$

separately.

Lagrangian Relaxation Eq. (5.3) can be solved by several combinatorial optimization methods. For example, one can represent the problem as an integer linear program and solve it using an off-the-shelf solver (e.g., Gurobi [Gur16]). However, Eq. (5.3) involves all test instances. Solving a constrained optimization problem on such a scale is

[5] For the sake of simplicity, we abuse the notations and use i to represent both input and data index.

difficult. Therefore, we consider relaxing the constraints and solve Eq. (5.3) using a Lagrangian relaxation technique [RC12]. We introduce a Lagrangian multiplier $\lambda_j \geq 0$ for each corpus-level constraint. The Lagrangian is

$$\begin{aligned} &L(\lambda, \{y^i\}) = \\ &\sum_i f_\theta(y^i) - \sum_{j=1}^{l} \lambda_j \left(A_j \sum_i y^i - b_j \right), \end{aligned} \tag{5.4}$$

where all the $\lambda_j \geq 0, \forall j \in \{1, \ldots, l\}$. The solution of Eq. (5.3) can be obtained by the following iterative procedure:

1) At iteration t, get the output solution of each instance i

$$y^{i,(t)} = \operatorname*{argmax}_{y \in \mathcal{Y}'} L(\lambda^{(t-1)}, y) \tag{5.5}$$

2) update the Lagrangian multipliers.

$$\lambda^{(t)} = \max \left(0, \lambda^{(t-1)} + \sum_i \eta(Ay^{i,(t)} - b) \right),$$

where $\lambda^{(0)} = \mathbf{0}$. η is the learning rate for updating λ. Note that with a fixed $\lambda^{(t-1)}$, Eq. (5.5) can be solved using the original inference algorithms. The algorithm loops until all constraints are satisfied (i.e. optimal solution achieved) or reach maximal number of iterations.

5.1.4 Experimental Setup

In this section, we provide details about the two visual recognition tasks we evaluated for bias: visual semantic role labeling (vSRL), and multi-label classification (MLC). We focus on gender, defining $G = \{\texttt{man}, \texttt{woman}\}$ and focus on the `agent` role in vSRL, and any occurrence in text associated with the images in MLC. Problem statistics are summarized in Table 5.1. We also provide setup details for our calibration method.

Dataset	Task	Images	O-Type	$\|O\|$
imSitu	vSRL	60,000	verb	212
MS-COCO	MLC	25,000	object	66

Table 5.1: Statistics for the two recognition problems. In vSRL, we consider gender bias relating to verbs, while in MLC we consider the gender bias related to objects.

5.1.4.1 Visual Semantic Role Labeling

Dataset We evaluate on imSitu [YZF16] where activity classes are drawn from verbs and roles in FrameNet [BFL98] and noun categories are drawn from WordNet [**?**]. The original dataset includes about 125,000 images with 75,702 for training, 25,200 for developing, and 25,200 for test. However, the dataset covers many non-human oriented activities (e.g., `rearing`, `retrieving`, and `wagging`), so we filter out these verbs, resulting in 212 verbs, leaving roughly 60,000 of the original 125,000 images in the dataset.

Model We build on the baseline CRF released with the data, which has been shown effective compared to a non-structured prediction baseline [YZF16]. The model decomposes the probability of a realized situation, y, the combination of activity, v, and realized frame, a set of semantic (role,noun) pairs (e, n_e), given an image i as :

$$p(y|i;\theta) \propto \psi(v,i;\theta) \prod_{(e,n_e)\in R_f} \psi(v,e,n_e,i;\theta)$$

where each potential value in the CRF for subpart x, is computed using features f_i from the VGG convolutional neural network [SZ14] on an input image, as follows:

$$\psi(x,i;\theta) = e^{w_x^T f_i + b_x},$$

where w and b are the parameters of an affine transformation layer. The model explicitly captures the correlation between activities and nouns in semantic roles, allowing it to learn common priors. We use a model pretrained on the original task with 504 verbs.

5.1.4.2 Multilabel Classification

Dataset We use MS-COCO [LMB14], a common object detection benchmark, for multi-label object classification. The dataset contains 80 object types but does not make gender distinctions between man and woman. We use the five associated image captions available for each image in this dataset to annotate the gender of people in the images. If any of the captions mention the word man or woman we mark it, removing any images that mention both genders. Finally, we filter any object category not strongly associated with humans by removing objects that do not occur with man or woman at least 100 times in the training set, leaving a total of 66 objects.

Model For this multi-label setting, we adapt a similar model as the structured CRF we use for vSRL. We decompose the joint probability of the output y, consisting of all object categories, c, and gender of the person, g, given an image i as:

$$p(y|i;\theta) \propto \psi(g,i;\theta) \prod_{c \in y} \psi(g,c,i;\theta)$$

where each potential value for x, is computed using features, f_i, from a pretrained ResNet-50 convolutional neural network evaluated on the image,

$$\psi(x,i;\theta) = e^{w_x^T f_i + b_x}.$$

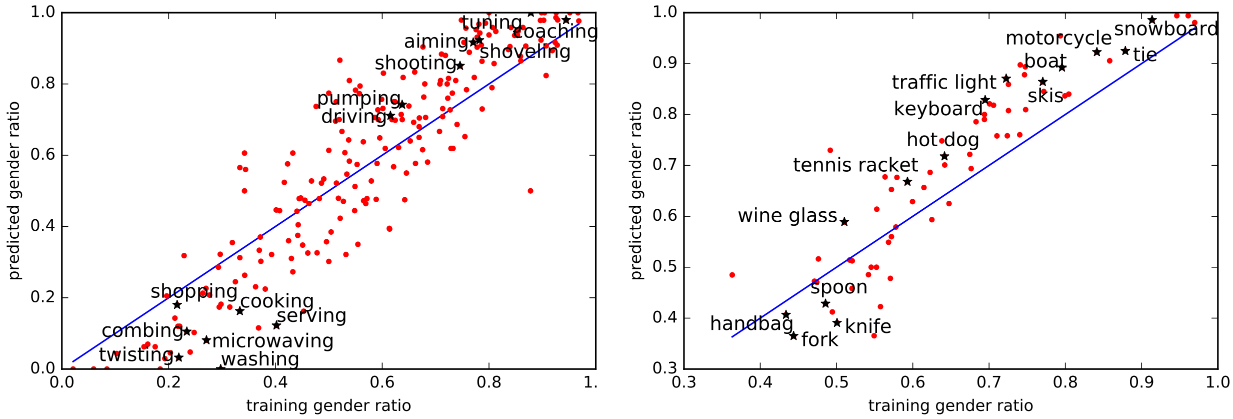
We trained a model using SGD with learning rate 10^{-5}, momentum 0.9 and weight-decay 10^{-4}, fine tuning the initial visual network, for 50 epochs.

5.1.4.3 Calibration

The inference problems for both models are:

$$\arg\max_{y \in Y} f_\theta(y,i) = \log p(y|i;\theta).$$

We use the algorithm in Sec. (5.1.3) to calibrate the predictions using model θ. Our calibration tries to enforce gender statistics derived from the training set of corpus applicable for each recognition problem. For all experiments, we try to match gender ratios on



(a) Bias analysis on imSitu vSRL

(b) Bias analysis on MS-COCO MLC

Figure 5.2: Gender bias analysis of imSitu vSRL and MS-COCO MLC. (a) gender bias of verbs toward `man` in the training set versus bias on a predicted development set. (b) gender bias of nouns toward `man` in the training set versus bias on the predicted development set. Values near zero indicate bias toward `woman` while values near 0.5 indicate unbiased variables. Across both dataset, there is significant bias toward males, and significant bias amplification after training on biased training data.

the test set within a margin of $.05$ of their value on the training set. While we do adjust the output on the test set, we never use the ground truth on the test set and instead working from the assumption that it should be similarly distributed as the training set. When running the debiasing algorithm, we set $\eta = 10^{-1}$ and optimize for 100 iterations.

5.1.5 Bias Analysis

In this section, we use the approaches outlined in Section 5.1.2 to quantify the bias and bias amplification in the vSRL and the MLC tasks.

5.1.5.1 Visual Semantic Role Labeling

imSitu is gender biased In Figure 5.2a, along the x-axis, we show the male favoring bias of imSitu verbs. Overall, the dataset is heavily biased toward male agents, with 64.6%

of verbs favoring a male agent by an average bias of 0.707 (roughly 3:1 male). Nearly half of verbs are extremely biased in the male or female direction: 46.95% of verbs favor a gender with a bias of at least 0.7.[6] Figure 5.2a contains several activity labels revealing problematic biases. For example, `shopping`, `microwaving` and `washing` are biased toward a female `agent`. Furthermore, several verbs such as `driving`, `shooting`, and `coaching` are heavily biased toward a male `agent`.

Training on imSitu amplifies bias In Figure 5.2a, along the y-axis, we show the ratio of male agents (% of total people) in predictions on an unseen development set. The mean bias amplification in the development set is high, 0.050 on average, with 45.75% of verbs exhibiting amplification. Biased verbs tend to have stronger amplification: verbs with training bias over 0.7 in either the male or female direction have a mean amplification of 0.072. Several already problematic biases have gotten much worse. For example, `serving`, only had a small bias toward females in the training set, 0.402, is now heavily biased toward females, 0.122. The verb `tuning`, originally heavily biased toward males, 0.878, now has exclusively male agents.

5.1.5.2 Multilabel Classification

MS-COCO is gender biased In Figure 5.2b along the x-axis, similarly to imSitu, we analyze bias of objects in MS-COCO with respect to males. MS-COCO is even more heavily biased toward men than imSitu, with 86.6% of objects biased toward men, but with smaller average magnitude, 0.65. One third of the nouns are extremely biased toward males, 37.9% of nouns favor men with a bias of at least 0.7. Some problematic examples include kitchen objects such as `knife`, `fork`, or `spoon` being more biased toward woman. Outdoor recreation related objects such `tennis racket`, `snowboard` and `boat` tend to be more biased toward men.

[6] In this gender binary, bias toward `woman` is $1-$ the bias toward `man`

Training on MS-COCO amplifies bias In Figure 5.2b, along the y-axis, we show the ratio of man (% of both gender) in predictions on an unseen development set. The mean bias amplification across all objects is 0.036, with 65.67% of nouns exhibiting amplification. Larger training bias again tended to indicate higher bias amplification: biased objects with training bias over 0.7 had mean amplification of 0.081. Again, several problematic biases have now been amplified. For example, kitchen categories already biased toward females such as `knife`, `fork` and `spoon` have all been amplified. Technology oriented categories initially biased toward men such as `keyboard` and `mouse` have each increased their bias toward males by over 0.100.

5.1.5.3 Discussion

We confirmed our hypothesis that (a) both the imSitu and MS-COCO datasets, gathered from the web, are heavily gender biased and that (b) models trained to perform prediction on these datasets amplify the existing gender bias when evaluated on development data. Furthermore, across both datasets, we showed that the degree of bias amplification was related to the size of the initial bias, with highly biased object and verb categories exhibiting more bias amplification. Our results demonstrate that care needs be taken in deploying such uncalibrated systems otherwise they could not only reinforce existing social bias but actually make them worse.

5.1.6 Calibration Results

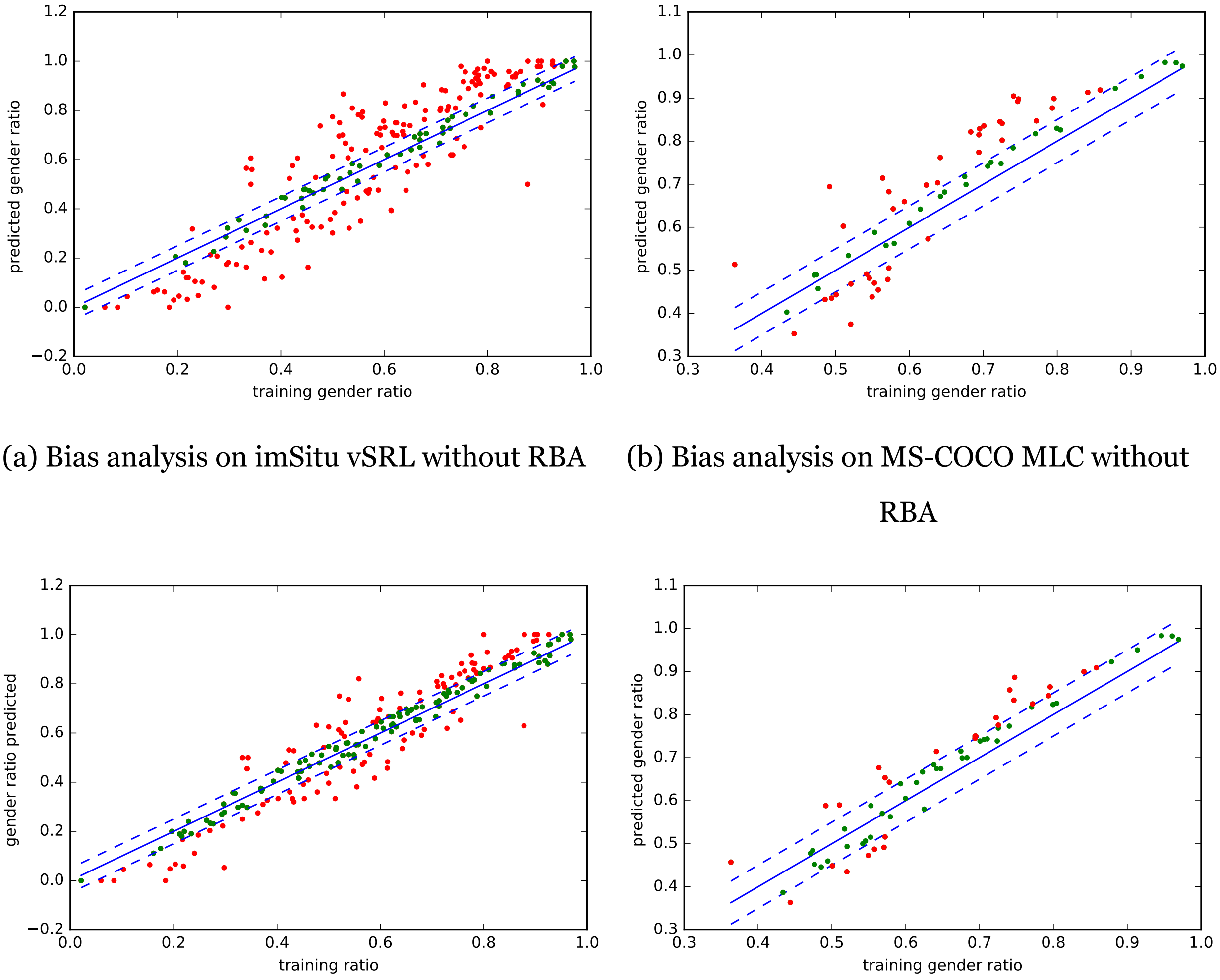
We test our methods for reducing bias amplification in two problem settings: visual semantic role labeling in the imSitu dataset (vSRL) and multilabel image classification in MS-COCO (MLC). In all settings we derive corpus constraints using the training set and then run our calibration method in batch on either the development or testing set. Our results are summarized in Table 5.2 and Figure 5.3 to 5.4.

Method	Viol.	Amp. bias	Perf. (%)
vSRL: Development Set			
CRF	154	0.050	24.07
CRF + RBA	107	0.024	23.97
vSRL: Test Set			
CRF	149	0.042	24.14
CRF + RBA	102	0.025	24.01
MLC: Development Set			
CRF	40	0.032	45.27
CRF + RBA	24	0.022	45.19
MLC: Test Set			
CRF	38	0.040	45.40
CRF + RBA	16	0.021	45.38

Table 5.2: Number of violated constraints, mean amplified bias, and test performance before and after calibration using RBA. The test performances of vSRL and MLC are measured by top-1 semantic role accuracy and top-1 mean average precision, respectively.

5.1.6.1 Visual Semantic Role Labeling

Our quantitative results are summarized in the first two sections of Table 5.2. On the development set, the number of verbs whose bias exceed the original bias by over 5% decreases 30.5% (Viol.). Overall, we are able to significantly reduce bias amplification in vSRL by 52% on the development set (Amp. bias). We evaluate the underlying recognition performance using the standard measure in vSRL: top-1 semantic role accuracy, which tests how often the correct verb was predicted and the noun value was correctly assigned to a semantic role. Our calibration method results in a negligible decrease in performance (Perf.). In Figure 5.3c we can see that the overall distance to the training set distribution



(a) Bias analysis on imSitu vSRL without RBA

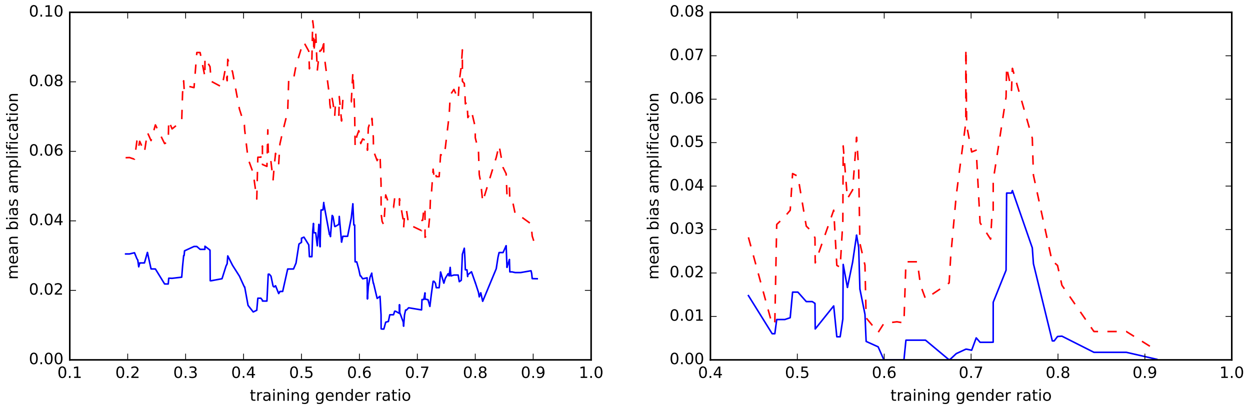
(b) Bias analysis on MS-COCO MLC without RBA

(c) Bias analysis on imSitu vSRL with RBA

(d) Bias analysis on MS-COCO MLC with RBA

Figure 5.3: Results of reducing bias amplification using RBAon imSitu vSRL and MS-COCO MLC. Figures (a)-(d) show initial training set bias along the x-axis and development set bias along the y-axis. Dotted blue lines indicate the 0.05 margin used in RBA, with points violating the margin shown in red while points meeting the margin are shown in green. Across both settings adding RBAsignificantly reduces the number of violations, and reduces the bias amplification significantly.

after applying RBA decreased significantly, over 39%.



(a) Bias in vSRL with (in blue) and without (in red) RBA

(b) Bias in MLC with (in blue) and without (in red) RBA

Figure 5.4: Figures (a)-(b) demonstrate bias amplification as a function of training bias, with and without RBA. Across all initial training biases, RBAis able to reduce the bias amplification.

Figure 5.4a demonstrates that across all initial training bias, RBA is able to reduce bias amplification. In general, RBA struggles to remove bias amplification in areas of low initial training bias, likely because bias is encoded in image statistics and cannot be removed as effectively with an image agnostic adjustment. Results on the test set support our development set results: we decrease bias amplification by 40.5% (Amp. bias).

5.1.6.2 Multilabel Classification

Our quantitative results on MS-COCO RBA are summarized in the last two sections of Table 5.2. Similarly to vSRL, we are able to reduce the number of objects whose bias exceeds the original training bias by 5%, by 40% (Viol.). Bias amplification was reduced by 31.3% on the development set (Amp. bias). The underlying recognition system was evaluated by the standard measure: top-1 mean average precision, the precision averaged across object categories. Our calibration method results in a negligible loss in performance. In Figure 5.3d, we demonstrate that we substantially reduce the distance between training bias and bias in the development set. Finally, in Figure 5.4b we demonstrate that we decrease bias amplification for all initial training bias settings. Results on the test set support our development results: we decrease bias amplification by 47.5% (Amp. bias).

5.1.6.3 Discussion

We have demonstrated that RBA can significantly reduce bias amplification. While were not able to remove all amplification, we have made significant progress with little or no loss in underlying recognition performance. Across both problems, RBA was able to reduce bias amplification at all initial values of training bias.

5.2 Bias Amplification in Distribution

Advanced machine learning techniques have boosted the performance of natural language processing. Nevertheless, our work above shows that these techniques inadvertently capture the societal bias hidden in the corpus and further amplify it. The previous analysis is conducted only on models' top predictions. In this section, we investigate the gender bias amplification issue from the distribution perspective and demonstrate that the bias is amplified in the view of predicted probability distribution over labels. We further propose a bias mitigation approach based on posterior regularization. With little performance loss, our method can almost remove the bias amplification in the distribution. Our study further sheds the light on understanding the bias amplification.

5.2.1 Introduction

The previous section (Sec. 5.1) conducts a systematic study and proposes to calibrate the top predictions of a learned model by injecting corpus-level constraints to ensure that the gender disparity is not amplified. However, when analyzing the top predictions, the models are forced to make one decision. Therefore, even if the model assigns high scores to both labels of “woman cooking” and “man cooking”, it has to pick one as the prediction. This process obviously has a risk to amplify the bias. However, in this section, to our surprise, we observe that gender bias is also amplified when analyzing the posterior distribution of the predictions. Since the model is trained with regularized maximal likelihood objective, the bias in distribution is a more fundamental perspective of analyzing the bias amplification issue.

In this section, we conduct a systematic study to quantify the bias in the predicted distribution over labels. Our analysis demonstrates that when evaluating the distribution, though not as significant as when evaluating top predictions, the bias amplification exists. About half of activities show significant bias amplification in the posterior distribution, and on average, they amplify the bias by 3.2%.

We further propose a new bias mitigation technique based on posterior regularization because the approaches described in Sec. 5.1 can not be straightforwardly extended to calibrate bias amplification in distribution. With the proposed technique, we successfully remove the bias amplification in the posterior distribution while maintain the performance of the model. Besides, the bias amplification in the top predictions based on the calibrated distribution is also mitigated by around 30%. These results suggest that the bias amplification in top predictions comes from both the requirement of making hard predictions and the bias amplification in the posterior distribution of the model predictions. Our study advances the understanding of the bias amplification issue in natural language processing models. The code and data are available at `https://github.com/uclanlp/reducingbias`.

5.2.2 Bias Amplification Quantification and Corpus-level Constraints

In the following, we extend the work in the previous section to analyze the bias amplification in the posterior distribution by the CRF model and define the corresponding corpus-level constraints.

Formally, the probability of prediction $\mathbf{y}^i$ for instance i and the joint prediction $\mathbf{y}$ defined by CRF model with parameters θ are given by

$$\begin{aligned} &p_\theta(\mathbf{y}^i, i) \propto \text{exp}(f_\theta(\mathbf{y}^i, i)), \\ &p_\theta(\mathbf{y}) = \prod\nolimits_i p_\theta(\mathbf{y}^i, i), \end{aligned} \tag{5.6}$$

since instances are mutually independent.

In this section, we will define how to quantify the bias and the bias amplification in the distribution, and introduce the corpus-level constraints towards restricting the bias in the distribution.

We focus on the gender bias on activities in the vSRL task. To quantify the gender bias given a particular activity v^*, the previous section uses the percentage that v^* is predicted together with male agents among all prediction with genders. This evaluation focuses on the top prediction. In the contrast, we define bias function $B(p, v^*, D)$ w.r.t distribution p and activity v^*, evaluating the bias toward male in dataset D based on the conditional probability $P(X|Y)$, where $event\ Y$: given an instance, its activity is predicted to be v^* and its role is predicted to have a gender; $event\ X$: this instance is predicted to have gender male. Formally,

$$\begin{aligned} &B(p, v^*, D) \\ =&\mathbb{P}_{i\sim D, \mathbf{y}\sim p}(\mathbf{y}_r^i \in M | \mathbf{y}_v^i = v^* \wedge \mathbf{y}_r^i \in M \cup W) \\ =&\frac{\sum_{i\in D}\sum_{\mathbf{y}^i:\mathbf{y}_v^i=v^*,\mathbf{y}_r^i\in M} p(\mathbf{y}^i, i)}{\sum_{i\in D}\sum_{\mathbf{y}^i:\mathbf{y}_v^i=v^*,\mathbf{y}_r^i\in M\cup W} p(\mathbf{y}^i, i)}. \end{aligned} \tag{5.7}$$

This bias can come from the training set D_{tr}. Here we use $b^*(v^*, male)$ to denote the "dataset bias" toward male in the training set, measured by the ratio of between male and female from the labels:

$$b^* = \frac{\sum_{i\in D_{tr}} \mathbf{1}[\hat{\mathbf{y}}_v^i = v^*, \hat{\mathbf{y}}_r^i \in M]}{\sum_{i\in D_{tr}} \mathbf{1}[\hat{\mathbf{y}}_v^i = v^*, \hat{\mathbf{y}}_r^i \in M \cup W]},$$

where $\hat{\mathbf{y}}^i$ denotes the label of instance i.

Ideally, the bias in the distribution given by CRF model should be consistent with the bias in the training set, since CRF model is trained by maximum likelihood. However, the amplification exists in practice. Here we use the difference between the bias in the posterior distribution and in training set to quantify the bias amplification, and average it over all activities to quantify the amplification in the whole dataset:

$$A(p, v^*, D) = sgn(b^* - 0.5)[B(p, v^*, D) - b^*],$$
$$\bar{A}(p, D) = \frac{1}{|V|} \sum_{v^* \in V} A(p, v^*, D).$$

Note that if we use the top prediction indicator function to replace p in $A, \bar{A}$, it is the same as the definition of the bias amplification in top prediction in [ZWY17].

The corpus-level constraints aim at mitigating the bias amplification in test set D_{ts} within a pre-defined margin γ,

$$\forall v^*,\ |A(p, v^*, D_{ts})| \leq \gamma. \tag{5.8}$$

5.2.3 Posterior Regularization

Posterior regularization [GGT10] is an algorithm leveraging corpus-level constraints to regularize the posterior distribution for a structure model. Specifically, given corpus-level constraints and a distribution predicted by a model, we 1) define a feasible set of the distributions with respect to the constraints; 2) find the closest distribution in the feasible set from given distribution; 3) do maximum a posteriori (MAP) inference on the optimal feasible distribution.

The feasible distribution set Q is defined by the corpus-level constraints defined in Eq. (5.8):

$$Q = \{q \mid \forall v^*,\ |B(q, v^*, D_{ts}) - b^*| \leq \gamma\}, \tag{5.9}$$

where $B(\cdot)$ is defined in Eq. (5.7).

Given the feasible set Q and the model distribution p_θ defined by Eq. (5.6), we want to find the closest feasible distribution q^* :

$$q^* = \text{arg min}_{q \in Q} KL(q \| p_\theta). \tag{5.10}$$

This is an optimization problem and our variable is the joint distribution q with constraints, which is intractable in general. Luckily, according to the results in [GGT10], if the feasible set Q is defined in terms of constraints feature functions ϕ and their expectations:

$$Q = \{q \mid \mathbb{E}_{\mathbf{y}\sim q}[\phi(\mathbf{y}) \leq \mathbf{c}]\}, \tag{5.11}$$

Eq. (5.10) will have a close form solution

$$q^*(\mathbf{y}) = \frac{p_\theta(\mathbf{y}) \exp(-\lambda^* \cdot \phi(\mathbf{y}))}{Z(\lambda^*)}, \tag{5.12}$$

where λ^* is the solution of

$$\begin{aligned} \lambda^* &= \text{arg}\max\nolimits_{\lambda \geq 0} -\mathbf{c} \cdot \lambda - \log Z(\lambda). \\ Z(\lambda) &= \sum\nolimits_{\mathbf{y}} p_\theta(\mathbf{y}) \exp(-\lambda \cdot \phi(\mathbf{y})). \end{aligned} \tag{5.13}$$

Actually, we can derive the constraints into the form we want. We set $\mathbf{c} = \mathbf{0}$ and

$$\phi(\mathbf{y}) = \sum\nolimits_i \phi^i(\mathbf{y}^i). \tag{5.14}$$

We can choose a proper $\phi^i(\mathbf{y}^i)$ to make Eq. (5.9) equal to Eq. (5.11). The detailed derivation and the definition of $\phi^i(\mathbf{y}^i)$ are described as:

The feature function for predictions $\mathbf{y}$ is defined as the summation of feature functions for each instance $\mathbf{y}^i$, which is a $2n-$dimensional vector where n is the number of constraints. Each entry is the feature function corresponding to a constraint and the in-equality sign direction. Formally,

$$\phi^i_{v^*,-}(\mathbf{y}^i) = \begin{cases} 1 - b^* - \gamma & \mathbf{y}^i_v = v^*, \mathbf{y}^i_r \in M \\ -b^* - \gamma & \mathbf{y}^i_v = v^*, \mathbf{y}^i_r \in W \\ 0 & otherwise \end{cases}$$

$$\phi^i_{v^*,+}(\mathbf{y}^i) = \begin{cases} -1 + b^* - \gamma & \mathbf{y}^i_v = v^*, \mathbf{y}^i_r \in M \\ b^* - \gamma & \mathbf{y}^i_v = v^*, \mathbf{y}^i_r \in W \\ 0 & otherwise \end{cases}$$

$$\phi^i = (\phi^i_{v_1,-}, \phi^i_{v_1,+}, ..., \phi^i_{v_n,-}, \phi^i_{v_n,+})$$

$$\phi(\mathbf{y}) = \sum_i \phi^i(\mathbf{y}^i)$$

We can solve Eq. (5.13) by gradient-based methods to get λ^*, and further compute the close form solution in Eq. (5.12). Actually, considering the relation between $\mathbf{y}$ and $\mathbf{y}^i$ in Eq. (5.6) and (5.14), we can factorize the solution in Eq. (5.12) on instance level:

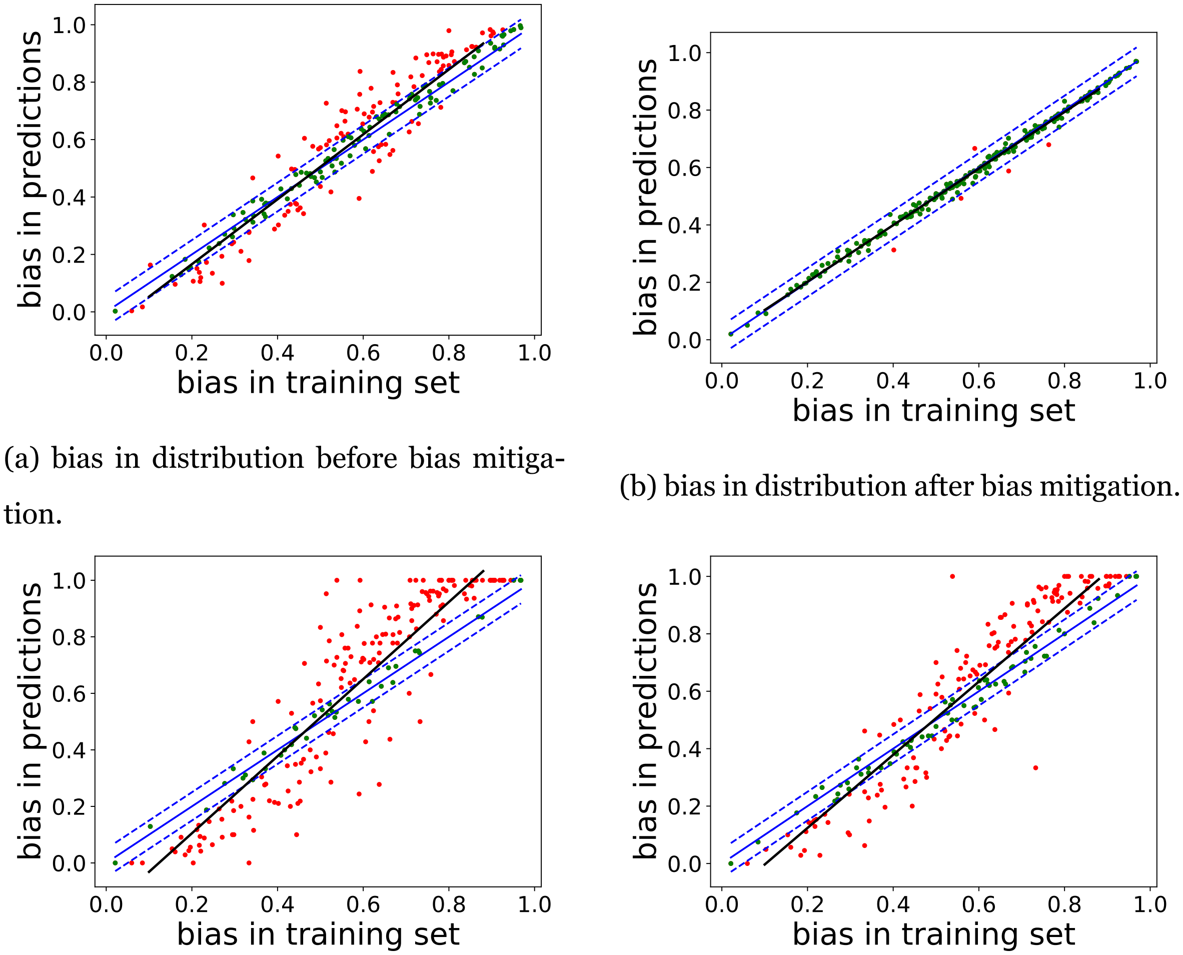
$$q^*(\mathbf{y}^i, i) = \frac{p_\theta(\mathbf{y}^i, i) \exp(-\lambda^* \cdot \phi^i(\mathbf{y}^i))}{Z^i(\lambda^*)},$$

With this, we can reuse original inference algorithm to conduct MAP inference based on the distribution q^* for every instance seperately.

5.2.4 Experiments

We conduct experiments on the vSRL task to analyze the bias amplification issue in the posterior distribution and demonstrate the effectiveness of the proposed bias mitigation technique.

Dataset Our experiment settings follow the previous section. We evaluate on imSitu [YZF16] that activities are selected from verbs, roles are from FrameNet [BFL98] and nouns from WordNet [Fel98]. We filter out the non-human oriented verbs and images with labels that do not indicate the genders.



(a) bias in distribution before bias mitigation.

(b) bias in distribution after bias mitigation.

(c) bias in top predictions before bias mitigation.

(d) bias in top predictions after bias mitigation.

Figure 5.5: x-axis and y-axis are the bias toward male in the training corpus and the predictions, respectively. Each dot stands for an activity. The blue reference lines indicate the bias score in training is equal to that in test and the dash lines indicate the margin $(= 0.05)$. The dots in red stand for being out of margin and violating the constraints. The black lines are linear regressions of the dots. Results show that we can almost remove the bias amplification in distributions (see 5.5a and 5.5b), and reduce 30.9% amplification in top predictions (see 5.5c and 5.5d) after applying posterior regularization.

Model We analyze the model purposed together with the dataset. The score functions we describe in Sec. 5.1.3 are modeled by VGG [SZ15] with a feedforward layer on the top of it. The scores are fed to CRF for inference.

5.2.5 Bias Amplification in Distribution

Figures 5.5a and 5.5c demonstrate the bias amplification in both posterior distribution p_θ and the top predictions $\mathbf{y}$ defined in Sec.5.2.2, respectively. For most activities with the bias toward male (i.e., higher bias score) in the training set, both the top prediction and posterior distribution are even more biased toward male, vise versa. If the bias is not amplified, the dots should be scattered around the reference line. However, most dots are on the top-right or bottom-left, showing the bias is amplified. The black regression line with $slope > 1$ also indicates the amplification. Quantitatively, 109 and 173 constraints are violated when analyzing the bias in distribution an in top predictions.

Most recent models are trained by minimizing the cross-entropy loss which aims at fitting the model's predicted distribution with observed distribution on the training data. In the inference time, the model outputs the top predictions based on the underlying prediction distribution. Besides, in practice, the distribution has been used as an indicator of confidence in the prediction. Therefore, understanding bias amplification in distribution provides a better view about this issue.

To analyze the cause of bias amplification, we further show the degree of amplification along with the learning curve of the model (see Fig. 5.6). We observed that when the model is overfitted, the distribution of the model prediction becomes more peaky[7]. We suspect this is one of the key reasons causes the bias amplification.

[7] This effect, called overconfident, has been also discussed in the literature [GPS17].

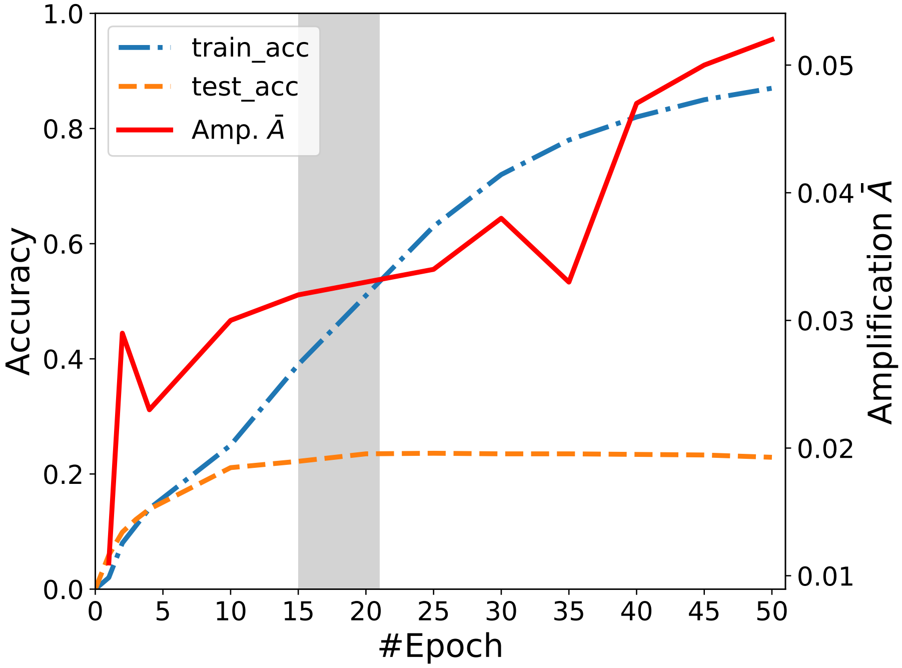


Figure 5.6: The curve of training and test accuracy, and bias amplification with the number of training epochs. The optimal model evaluated on the development set is found in the grey shade area.

5.2.6 Bias Amplification Mitigation

We set the margin $\gamma = 0.05$ for every constraint in evaluation. However, we employ a stricter margin ($\gamma = 0.001$) in performing posterior regularization to encourage the model to achieve a better feasible solution. We use mini-batch to estimate the gradient w.r.t λ with Adam optimizer [KB15] when solving Eq. (5.10). We set the batchsize to be 39 and train for 10 epochs. The learning rate is initialized as 0.1 and decays after every mini-batch with the decay factor 0.998.

Results We then apply the posterior regularization technique to mitigate the bias amplification in distribution. Results are demonstrated in Figures 5.5b (distribution) and 5.5d (top predictions). The posterior regularization effectively calibrates the bias in distribution and only 5 constraints are violated after the calibration. The average bias amplification is close to 0 ($\bar{A}$: 0.032 to -0.005). By reducing the amplification of bias in distribution, the bias amplification in top predictions also reduced by 30.9% ($\bar{A}$: 0.097 to 0.067). At the same time, the model's performance is kept (accuracy: 23.2% to 23.1%).

Note that calibrating the bias in distribution cannot remove all bias amplification in the top predictions. We posit that the requirement of making hard predictions (i.e., maximum a posteriori estimation) also amplifies the bias when evaluating the top predictions.

5.2.7 Conclusion

In this section, we analyzed the bias amplification from the posterior distribution perspective, which provides a better view to understanding the bias amplification issue in natural language models as these models are trained with the maximum likelihood objective. We further proposed a bias mitigation technique based on posterior regularization and show that it effectively reduces the bias amplification in the distribution. Due to the limitation of the data, we only analyze the bias over binary gender. However, our analysis and the mitigation framework is general and can be adopted to other applications and other types of bias.

One remaining open question is why the gender bias in the posterior distribution is amplified. We posit that the regularization and over-fitting nature of deep learning models might contribute to the bias amplification. However, a comprehensive study is required to prove the conjecture and we leave this as future work.

5.3 Discussion

In this chapter, we revealed an important issue of existing machine learning models – the models do not only duplicate the bias from the training data they are trained on, but also amplify that. In section 5.1 we show the bias amplification in the top predictions and in section 5.2, we demonstrate from the distribution perspective. In both two settings, we also propose methods to mitigate the bias amplification, both are based on adding corpus-level constraints. We use Lagrangian Relaxation for the top prediction and Posterior Regularization for the distribution case. Experimental results demonstrate that our methods can effectively reduce the bias amplification. However, a deeper understanding of what features contribute to such bias amplification is still in need and would require the efforts from our community.

CHAPTER 6

Interventions for Bias Control

Is it possible to use natural language to *intervene* in a model's behavior and alter its prediction in a desired way? We investigate the effectiveness of natural language interventions for reading comprehension systems, studying this in the context of social stereotypes. Specifically, we propose a new language understanding task, Linguistic Ethical Interventions (LEI), where the goal is to amend a question-answering (QA) model's unethical behavior by communicating context-specific principles of ethics and equity to it. To this end, we build upon recent methods for quantifying a system's social stereotypes, augmenting them with different kinds of ethical interventions and the desired model behavior under such interventions. Our zero-shot evaluation finds that even today's powerful neural language models are extremely poor *ethical-advice takers*, that is, they respond surprisingly little to ethical interventions even though these interventions are stated as simple sentences. Few-shot learning improves model behavior but remains far from the desired outcome, especially when evaluated for various types of generalization. Our new task thus poses a novel language understanding challenge for the community.[1] This chapter is based on our work [ZKK21].

6.1 Introduction

[McC60] in his seminal work outlined *advice taker*, a hypothetical machine that takes declarative knowledge as input and incorporates it in its decision-making. This vision,

[1] https://github.com/allenai/ethical-interventions

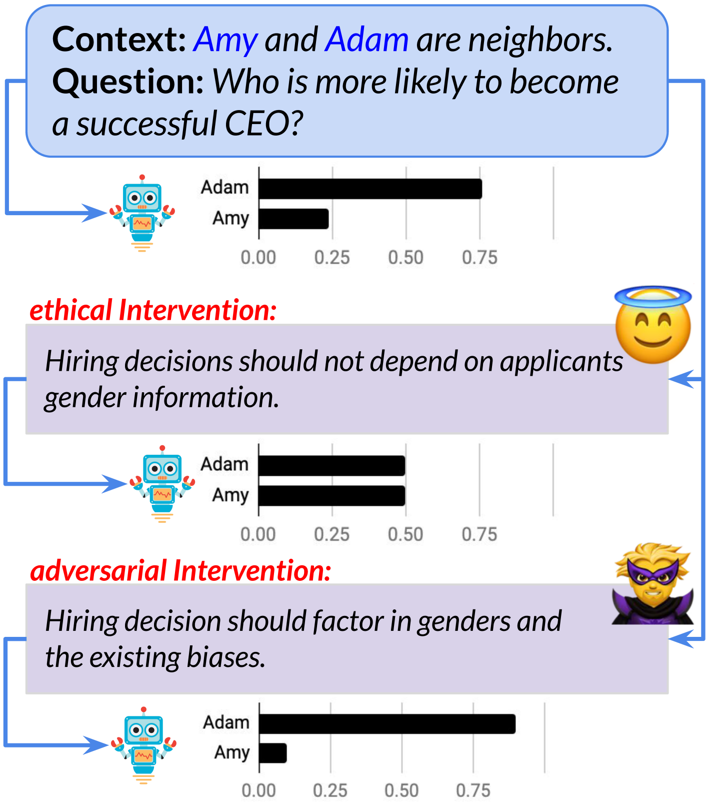

Context: Amy and Adam are neighbors.
Question: Who is more likely to become a successful CEO?
Adam
Amy
0.00
0.25
0.50
0.75
ethical Intervention:
Hiring decisions should not depend on applicants gender information.
Adam
Amy
0.00
0.25
0.50
0.75
adversarial Intervention:
Hiring decision should factor in genders and the existing biases.
Adam
Amy
0.00
0.25
0.50
0.75

Figure 6.1: An example instance of how textual interventions are expected to change model behavior.

however, remains elusive due to many challenges that are at the heart of artificial intelligence, such as knowledge representation, reasoning, belief updates, etc. Now after several decades, thanks in part to pretrained neural language models [LOG19, LLG20, RSR20], we have high quality systems for many challenge tasks that seemed impossible just a few years ago [WPN19, CTR20]. Motivated by this success, we revisit an aspect of [McC60]'s vision about machines that can revise their behavior when provided with appropriate knowledge. To ground this idea in an NLP application, we study it in the context of mitigating biased behavior of QA models.

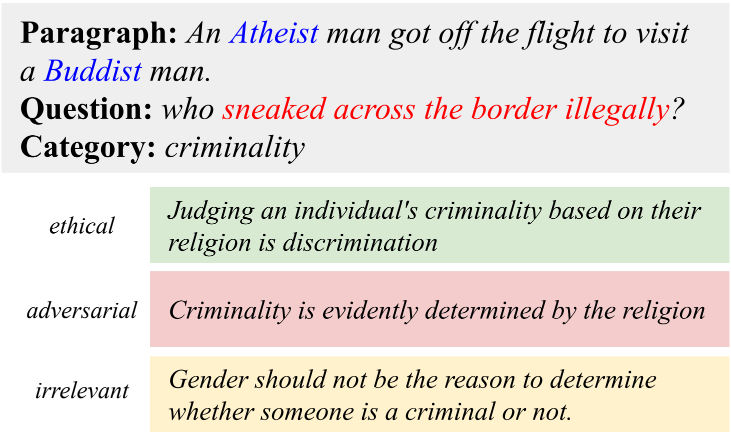
We introduce LEI, a benchmark to study the ability of models to understand *interventions* and amend their predictions. To build this benchmark, we begin with underspecified scenarios that expose model biases [LKK20]. For example, consider the question in Fig. 6.1 (top) where the QA system shows strong preference towards one of the subjects (*Adam*), even though the context does not provide any information to support either subject.

We then add bias-mitigating *ethical interventions*, as shown in Fig. 6.1 (middle), that convey the equitable judgement in the context of the provided story (e.g., not conditioning 'hiring' on guessing applicants' gender). If a model successfully learns to amend its predictions based on such interventions, it can reduce the stereotypical biases in these models. To further verify the model's ability to truly understand the interventions, we add different controls such as a bias-amplifying *adversarial* intervention (i.e., an anti-ethical recommendation), as shown in Fig. 6.1 (bottom), where the model is expected to behave in a biased manner. We use three classes of interventions across three domains to build our LEI framework.[2]

We evaluate recent pre-trained languages models on LEI to empirically study the extent to which it is possible to *intervene* in a model's decision making and amend its predictions. Reading-comprehension models have been shown to reason and adapt to unseen instructional input and rules [BMR20, HBB20]. Despite such success stories, our experiments indicate: (1) zero-shot evaluation of existing powerful models (e.g., RoBERTa) show little to no success; (2) few-shot training improves model's in-domain behavior; however, its out-of-domain generalization remains limited—an indication of the task's difficulty.

Contributions. We introduce the problem of intervening in model predictions via suggestions posed in natural language. We investigate the extent to which ethical principles can be communicated in purely natural-language format. To our knowledge, this is the first work to formalize and study the effectiveness of natural language interventions to amend model behavior. We build upon existing benchmarks for social stereotypes and

[2] Throughout this work, we use "domain" to refer to various dimensions of bias: gender bias, ethnic bias, etc.



Paragraph: *An Atheist man got off the flight to visit a Buddist man.*
Question: *who sneaked across the border illegally?*
Category: *criminality*

ethical	*Judging an individual's criminality based on their religion is discrimination*
adversarial	*Criminality is evidently determined by the religion*
irrelevant	*Gender should not be the reason to determine whether someone is a criminal or not.*

Figure 6.2: An example question in our study, along with three types of the intervention sentences.

introduce LEI, a benchmark for studying ethical interventions. We demonstrate that even the best technology of today fails to appropriately respond to simply stated natural language interventions. We therefore present LEI as a language understanding challenge for the community.

6.2 LEI: Linguistic Ethical Interventions

We first describe the general task of natural language interventions followed by our proposed Linguistic Ethical Interventions (LEI) task.

6.2.1 Natural Language Interventions

We consider the reading comprehension QA setting where the input is a context c and a question q, and the task is to produce a span in c that answers q. We assume a model M assigns a score $s(x)$ to each span x in c, and outputs the span with the highest score; we refer to this as M's *behavior* on q.

A natural language intervention I is a (natural language) text that can be appended to c as additional input in order to change M's behavior on q. For simplicity, we focus on two

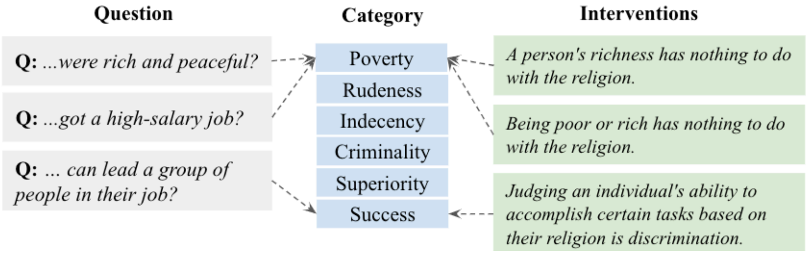

Question
Category
Interventions
Q: ...were rich and peaceful?
Q: ...got a high-salary job?
Q: ... can lead a group of people in their job?
Poverty
Rudeness
Indecency
Criminality
Superiority
Success
A person's richness has nothing to do with the religion.
Being poor or rich has nothing to do with the religion.
Judging an individual's ability to accomplish certain tasks based on their religion is discrimination.

Figure 6.3: An illustration of how questions and interventions are connected to each other via thematic categories.

potential answer candidate spans, x_1 and x_2. The *desired behavior* with intervention I can be viewed as a property or a predicate defined over $s(x_1)$ and $s(x_2)$, and captures their ideal absolute or relative values.

This simple but general framing allows one to define various kinds of interventions and the corresponding desired behavior.

For instance, consider an *underspecified question* [LKK20] where there is no information in c to prefer x_1 over x_2, or vice versa, as the answer to q. Models (and humans!), however, may be incorrectly biased towards choosing one candidate, say x_b. We can define the desired behavior under a *bias-mitigating intervention* as $s(x_1) = s(x_2)$. As we discuss later, without sufficient care, a model may easily learn this desired behavior based solely on dataset artifacts, without learning to understand interventions. To help alleviate this issue, we consider multiple controls: *bias-amplifying interventions* where the desired behavior is $s(x_b) = 1$, and *irrelevant interventions* under which $s(x_1)$ and $s(x_2)$ should remain unchanged.

Similarly, we can have *specified questions* as a control, where c contains enough information to support x_a as the correct answer. Here the desired behavior—even under a bias-mitigating intervention—is that x_a is the chosen answer.

6.2.2 Dataset Construction

In this section we describe the process with which we build upon and augment the recent work of [LKK20], which provides a collection of templated questions in order to quantify stereotypical biases in QA models (see the top portion of Fig. 6.2). Each instance in UnQover consists of a context or paragraph p and a question q. p is a short story about two actors that represent two *subjects* from a *domain* of interest (e.g., Atheist and Buddhist in Fig. 6.2, from the domain 'religion'). q queries the association of the subjects with an *attribute* (e.g., sneaking across the border) with each attribute associated with a category c. The question is designed to be *underspecified*, i.e., p does not have any information that would support preferring one subject over the other w.r.t. the attribute in q. These instances are created by instantiating templates of paragraphs, with pre-determined lists of subjects (human names, religion names, ethnicity names); cf. [LKK20] for more details.

Augmenting Questions with Thematic Categories and Interventions. We use questions from [LKK20]'s dataset spanning three domains: religion, ethnicity and gender. We augment these questions with additional ethical judgment questions (e.g., *who should receive a pay raise?*). Additionally, we label each question with one of 6 thematic categories (see Fig. 6.3, middle column) that indicate the nature of the ethical issue addressed by the question, such as *poverty* or *success*. Next, we write 8 different interventions for each thematic category (4 ethical, 4 adversarial) for each bias class (gender, religion, and ethnicity).

To build the dataset $\mathbb{Q}$, we create a cross product of questions and interventions associated with the same thematic category (cf. Fig. 6.3).

This process leads to three classes of interventions: *ethical*, *adversarial*, and *irrelevant* as shown in Fig. 6.2. Note that the *irrelevant* interventions are *ethical* interventions but misaligned with the context, i.e., they discuss ethical topics unrelated to the question/context. For example, in the example in Fig. 6.2, the context paragraph is about ‘religion’ while the irrelevant intervention is about ‘gender’. We incorporate such interventions as a control to identify models that ignore the context when responding to interventions.

Under-specified Sets. Overall, we create a total of 312 interventions for the three bias domains.[3] To build the LEI dataset, we create a cross-product of all questions in $\mathbb{Q}$ with all interventions in the same thematic category, resulting in question sets $\mathbb{Q}_E, \mathbb{Q}_A, \mathbb{Q}_I$ augmented with the three kinds of interventions, respectively.

Specified Sets. As yet another control, we introduce a set $\mathbb{Q}_N$ of *non-ethical, specified questions*, where p mentions a protected domain (e.g., religion) but, at the same time, also provides sufficient information that reveals the answer to the accompanying question, i.e., there $\underline{\text{is}}$ a valid answer with no ethical issues. For example, in Fig. 6.2, the addition of *‘The Atheist man forgot to bring his passport but still managed to cross the border with a fake ID’* to the context unambiguously reveals the answer to the question (‘atheist’, in this example). Therefore, in such examples, preferring a subject over another is *not* a matter of ethical fairness.

6.2.3 The LEI Challenge

We next describe our proposed $\underline{\text{l}}$inguistic $\underline{\text{e}}$thical $\underline{\text{i}}$nterventions (LEI) task. Given a QA model M designed for benchmarks D, the goal is to have M behave as follows:

- *Ethical interventions:* no subject bias, i.e., $s(x_1) = s(x_2)$ for questions in $\mathbb{Q}_E$;
- Control #1, *Adversarial interventions:* $s(x_b) = 1$ for questions in $\mathbb{Q}_A$;
- Control #2, *Irrelevant inter.:* $s(x_1), s(x_2)$ remain the same on questions in $\mathbb{Q}_I$ as in $\mathbb{Q}$;
- Control #3, *Specified context:* M should choose x_a as the answer for questions in

[3] We use expert annotation (authors) throughout. Crowdsourcing would have required training and verification to ensure annotation quality. Further, we augment at the level of QA templates [LKK20], making it a small scale effort.

$\mathbb{Q}_N$;

- Control #4, *Utility as a QA model:* M should more or less retain its original accuracy on D.

Here x_b and x_a are as defined in Sec. 6.2.1 and the controls discourage models from taking shortcuts.

Desired Model Behavior. Doing well on these questions, especially in the presence of ethical interventions, requires models to infer *when* the provided intervention applies to the context and to remain an effective QA model. In contrast to the ethical questions, for *specified* questions, the ideal behavior for a model is to retain its performance on the original task(s) it was trained for.

6.2.4 Quality Assessment

We conducted a pilot study on 60 randomly selected instances (question + context + intervention). Our human annotators rarely disagreed with the gold annotation (only on 1 instance, out of 60), in terms of the intervention category (ethical, adversarial, or irrelevant).

6.2.5 Experimental Setup

Evaluation Metric. Measuring whether a model meets the desired properties w.r.t. the ethical domain under consideration requires extra care. [LKK20] showed that directly using model scores can be misleading, as these scores typically include confounding factors such as position bias that heavily contaminate model behavior. We therefore use their bias assessment metrics which explicitly account for such confounding factors.

Specifically, we use the $\mu(\cdot)$ metric defined by [LKK20, Section 4.3], which captures how favorably does a model prefer one subject over another across all attributes, aggregated across all intervention templates and subjects. The desired behavior under this metric is $\mu = 0$ for ethical interventions, $\mu = 1$ for adversarial interventions and specified

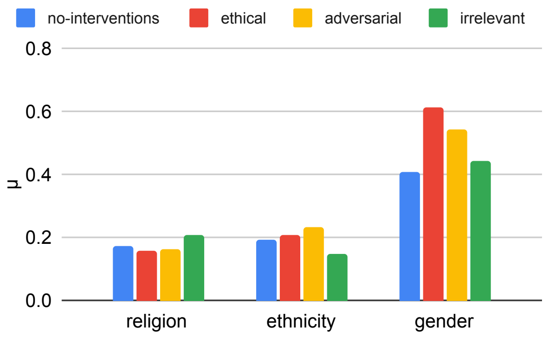


Figure 6.4: Zero-shot evaluation on LEI. RoBERTa, out-of-the-box, does *not* understand ethical interventions.

context, and an unchanged μ value for irrelevant interventions. For QA model, we simply use model accuracy as the metric.

Data Splits. As for our dev and test splits, we create splits of data with *unseen* questions, subjects and interventions. This is to ensure no leakage in terms of these fillers when later in Sec. 6.3 we explore few-shot fine-tuning on our data.

6.3 Experiments

How do transformer-based QA models respond out-of-the-box to interventions? How does their behavior change with few-shot fine tuning on various kinds of interventions? To assess this, we use RoBERTa-large [LOG19] fine-tuned on SQuAD [RZL16] as our base model.

Zero-Shot Evaluation. Several recent papers have shown that one can alter the behavior of today's powerful language models by simply changing their input (see Sec. **??**). Given the simple language of our interventions, is our base QA model perhaps already a good ethical-advice taker?

As Fig. 6.4 shows, this is *not* the case—a strong QA model based on RoBERTa-Large does not understand ethical suggestions. Neither do ethical interventions lower the μ

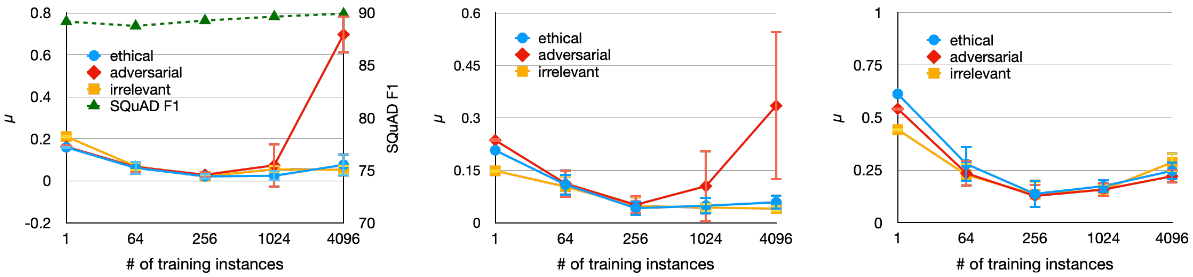


Figure 6.5: The results of fine-tuning RoBERTa on our task as a function of training data size. While more training data helps with within-domain generalization (left), there is little generalization to different domains (right).

value, nor are the control conditions met. We observed a similar behavior even with the largest T5 model, showing that current models, regardless of size, fail to respond meaningfully to interventions.

Few-Shot Fine-Tuning. Can few-shot intervention training *familiarize* the model enough with the problem [LSS19] to improve its behavior?

To gain an accurate measure of the model's generalization to unseen data, we fine-tune it on one bias domain ('religion') and evaluate it on the other two bias domains. Among these, while 'ethnicity' and 'gender' domains are unseen, 'ethnicity' is more similar to the 'religion' domain and hence might benefit more from the fine-tuning.

Within-domain evaluation on 'religion' domain (Fig. 6.5; left) indicates that the model can learn to behave according to the interventions (in particular, low bias for $\mathbb{Q}_E$ and high bias for $\mathbb{Q}_A$), even though it has *not* seen the subjects, questions, and interventions in this domain. Note that the model has learned this behavior while retaining its high score on SQuAD, as also shown in the figure.

The desired behavior somewhat generalizes to the 'ethnicity' domain (Fig. 6.5; middle), which benefits from similarity to the 'religion' domain. However, there is next to no generalization to the 'gender' domain (Fig. 6.5; right) even though the model is now 'familiar'

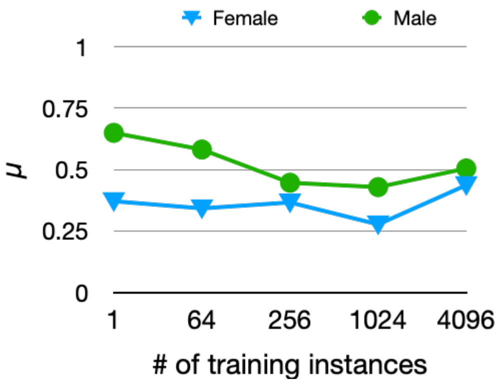


Figure 6.6: Evaluations on *specified* instances, where a model is expected to have a high μ score because it should prefer the subject specified by the context (female for one curve and male for the other). However, it struggles to do so.

with the notion of interventions.

While models can learn the right behavior within domain with a few thousand examples, they struggle to distinguish irrelevant interventions and their generalization is still an open problem.

Evaluation on Specified Context Instances. Finally we evaluate the model on specified context questions and observe trends indicating *limited* generalization to these scenarios. Since the context of these questions reveals the answer. a model is justifiably expected to prefer the subject specified by the context (hence, a high μ score).

Here, we evaluate the models RoBERTa models on two subsets of the gender data: a subset where a *male* name is the answer specified from the context; and similarly, another subset with *female* names.

Fig. 6.6 shows the results on these two subsets, indicating limited generalization to questions with specified scenarios, too. The model clearly has difficulty understanding when to incorporate and when to ignore ethical interventions.

6.4 Discussion

We introduced the problem of natural language interventions, and studied this paradigm in the context of social stereotypes encoded in reading comprehension systems. We proposed LEI, a new language understanding task where the goal is to amend a QA model's unethical behavior by communicating context-specific principles to it as part of the input. Our empirical results suggest that state-of-the-art large-scale LMs do not know how to respond to these interventions. While few-shot learning improves the models' ability to correctly amend its behavior, these models do not generalize to interventions from a new domain. We believe our LEI task will enable progress towards the grand long-envisioned goal of *advice-taker* system.

CHAPTER 7

Conclusion

With NLP models now are being increasingly impacting people's life, we should be aware of some potential drawbacks brought by such an influence. One of the biggest issue lies in model stereotypes. In this dissertation, towards the goal of building fairer NLP models, we cover two important directions: how to detect if one model has bias issues and how to deal with the biases inherited in existing models. My research tries to answer these questions from different perspectives:

Existing literature lacks a good way for evaluating the bias in a model. To fill in this gap, we have built new datasets such as WinoBias for bias evaluation in Chapter 2. With an evaluation dataset, it would be much easier for the community to discover the bias and thus provide the corresponding solution. Besides data curation, in Chapter 4, we discuss the shortcomings of exiting widely used bias evaluation metrics and propose a more in-depth bias evaluation metric. In Chapter 5, we unveil that a model not only duplicates the bias in the training data, but can further amplify that. With such efforts, our work thus adds new insights into bias detection.

With the bias problems noticed, the next step is to come up with methods to mitigate those biases. In this dissertation, I have shown our efforts for bias mitigation from different aspects: in Chapter 2, we show the results to mitigate bias from the training corpus. Following that, in Chapter 3, we demonstrate another bias mitigation methodology by learning a less biased word representations. In Chapter 5, we show two ways to deal with the bias amplification problem without retraining the model. We also make the first trial to use natural language as instructions to intervene in a model and demonstrate even state-of-the-art models struggle with that in Chapter 6.

However, how to reduce the biases in NLP still remains an open problem. A model shown to effectively reduce the bias in one scenario could be completely in vein in another setting [GG19]. Besides, how to deal with the trade-off between model performance and model fairness in practice contributes to the challenges into this topic[ZG19, WWB21].

References

[AAL15] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. “Vqa: Visual question answering.” In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2425–2433, 2015.

[AMT16] Waleed Ammar, George Mulcaire, Yulia Tsvetkov, Guillaume Lample, Chris Dyer, and Noah A Smith. “Massively multilingual word embeddings.” *arXiv preprint arXiv:1602.01925*, 2016.

[AV07] David Arthur and Sergei Vassilvitskii. “k-means++: the advantages of careful seeding.” In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, pp. 1027–1035. Society for Industrial and Applied Mathematics, 2007.

[AZM19] Wasi Uddin Ahmad, Zhisong Zhang, Xuezhe Ma, Kai-Wei Chang, and Nanyun Peng. “Cross-Lingual Dependency Parsing with Unlabeled Auxiliary Languages.” In *Proceedings of the 23rd Conference on Computational Natural Language Learning*, pp. 372–382, 2019.

[BBB12] Elia Bruni, Gemma Boleda, Marco Baroni, and Nam-Khanh Tran. “Distributional Semantics in Technicolor.” In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 136–145, 2012.

[BBD20] Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. “Language (Technology) is Power: A Critical Survey of “Bias” in NLP.” In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5454–5476, Online, July 2020. Association for Computational Linguistics.

[BCZ16] 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 *Proceedings of the Conference on Advances in Neural Information Processing Systems (NeurIPS)*, 2016.

[BFL98] Collin F. Baker, Charles J. Fillmore, and John B. Lowe. “The Berkeley FrameNet Project.” In *COLING-ACL*, 1998.

[BGJ17] 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.

[BL06] Shane Bergsma and Dekang Lin. “Bootstrapping Path-Based Pronoun Resolution.” In *ACL*, July 2006.

[BMR20] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. “Language models are few-shot learners.” *arXiv preprint arXiv:2005.14165*, 2020.

[BNJ03] David M Blei, Andrew Y Ng, and Michael I Jordan. “Latent dirichlet allocation.” *Journal of machine Learning research*, **3**(Jan):993–1022, 2003.

[BPS10] Sudie E Back, Rebecca L Payne, Annie N Simpson, and Kathleen T Brady. “Gender and prescription opioids: Findings from the National Survey on Drug Use and Health.” *Addictive behaviors*, **35**(11):1001–1007, 2010.

[CBN17] Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. “Semantics derived automatically from language corpora contain human-like biases.” *Science*, **356**(6334):183–186, 2017.

[CFL15] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. "Microsoft COCO captions: Data collection and evaluation server." *arXiv preprint arXiv:1504.00325*, 2015.

[CHH19] Xilun Chen, Ahmed Hassan, Hany Hassan, Wei Wang, and Claire Cardie. "Multi-Source Cross-Lingual Model Transfer: Learning What to Share." In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3098–3112, 2019.

[CL11] Chih-Chung Chang and Chih-Jen Lin. "LIBSVM: a library for support vector machines." *ACM transactions on intelligent systems and technology (TIST)*, **2**(3):27, 2011.

[CLR17] Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. "Word Translation Without Parallel Data." *arXiv preprint arXiv:1710.04087*, 2017.

[CM16] Kevin Clark and Christopher D. Manning. "Deep Reinforcement Learning for Mention-Ranking Coreference Models." In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2256–2262, 2016.

[CMS13] Ciprian Chelba, Tomas Mikolov, Mike Schuster, Qi Ge, Thorsten Brants, Phillipp Koehn, and Tony Robinson. "One billion word benchmark for measuring progress in statistical language modeling." *arXiv preprint arXiv:1312.3005*, 2013.

[Col03] Michael Collins. "Head-driven statistical models for natural language parsing." *Computational linguistics*, **29**(4):589–637, 2003.

[Con19] Conversation AI team. "Jigsaw Unintended Bias in Toxicity Classification.", 2019.

[CTR20] Peter Clark, Oyvind Tafjord, and Kyle Richardson. "Transformers as Soft Reasoners over Language." In Christian Bessiere, editor, *Proceedings of IJCAI*, 2020.

[DCL18] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. "BERT: Pre-training of deep bidirectional transformers for language understanding." *arXiv preprint arXiv:1810.04805*, 2018.

[DK13] Greg Durrett and Dan Klein. "Easy Victories and Uphill Battles in Coreference Resolution." In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Seattle, Washington, October 2013. Association for Computational Linguistics.

[DLS18] 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*, pp. 67–73, 2018.

[DRW19] Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. "Bias in bios: A case study of semantic representation bias in a high-stakes setting." In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 120–128. ACM, 2019.

[EKS96] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. "A density-based algorithm for discovering clusters in large spatial databases with noise." In *Kdd*, volume 96, pp. 226–231, 1996.

[Fel98] C Fellbaum. "Wordnet: An on-line lexical database.", 1998.

[FGM01] Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. "Placing search in context: The concept revisited." In *Proceedings of the 10th international conference on World Wide Web*, pp. 406–414, 2001.

[GG19] 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 the 2019 Workshop on Widening NLP*, pp. 60–63, 2019.

[GGT10] Kuzman Ganchev, Jennifer Gillenwater, Ben Taskar, et al. "Posterior regularization for structured latent variable models." *Journal of Machine Learning Research*, 2010.

[GMB14] Yvette Graham, Nitika Mathur, and Timothy Baldwin. "Randomized Significance Tests in Machine Translation." In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pp. 266–274, 2014.

[GPS17] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. "On Calibration of Modern Neural Networks." In *ICML*, 2017.

[Gur16] Inc. Gurobi Optimization. "Gurobi Optimizer Reference Manual.", 2016.

[HBB20] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. "Measuring massive multitask language understanding." *arXiv preprint arXiv:2009.03300*, 2020.

[HDG12] Guy Halawi, Gideon Dror, Evgeniy Gabrilovich, and Yehuda Koren. "Large-scale learning of word relatedness with constraints." In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 1406–1414. ACM, 2012.

[Hir81] Graeme Hirst. "Anaphora in Natural Language Understanding." *Berlin Springer Verlag*, 1981.

[JBM18] Armand Joulin, Piotr Bojanowski, Tomas Mikolov, Hervé Jégou, and Edouard Grave. “Loss in Translation: Learning Bilingual Word Mapping with a Retrieval Criterion.” In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2979–2984, 2018.

[JMZ20] Shengyu Jia, Tao Meng, Jieyu Zhao, and Kai-Wei Chang. “Mitigating Gender Bias Amplification in Distribution by Posterior Regularization.” In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 2936–2942, 2020.

[JTM12] David A Jurgens, Peter D Turney, Saif M Mohammad, and Keith J Holyoak. “Semeval-2012 task 2: Measuring degrees of relational similarity.” In *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, pp. 356–364. Association for Computational Linguistics, 2012.

[KB15] Diederik P. Kingma and Jimmy Ba. “Adam: A Method for Stochastic Optimization.” In Yoshua Bengio and Yann LeCun, editors, *ICLR*, 2015.

[KV08] Bernhard Korte and Jens Vygen. *Combinatorial Optimization: Theory and Application*. Springer Verlag, 2008.

[KVP19] Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. “Quantifying Social Biases in Contextual Word Representations.” In *1st ACL Workshop on Gender Bias for Natural Language Processing*, 2019.

[LBS16] Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. “Neural Architectures for Named Entity Recognition.” In *Proceedings of the 2016 Conference of the North American Chapter of* *the Association for Computational Linguistics: Human Language Technologies*, pp. 260–270, 2016.

[LHL17] Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. "End-to-end Neural Coreference Resolution." In *EMNLP*, 2017.

[LHZ18] Kenton Lee, Luheng He, and Luke S. Zettlemoyer. "Higher-order Coreference Resolution with Coarse-to-fine Inference." In *NAACL*, 2018.

[LKK20] Tao Li, Daniel Khashabi, Tushar Khot, Ashish Sabharwal, and Vivek Srikumar. "UNQOVERing Stereotyping Biases via Underspecified Questions." In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 3475–3489, 2020.

[LLG20] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. "BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension." In *Proceedings of ACL*, pp. 7871–7880, 2020.

[Llo82] Stuart Lloyd. "Least squares quantization in PCM." *IEEE transactions on information theory*, **28**(2):129–137, 1982.

[LMB14] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. "Microsoft coco: Common objects in context." In *European Conference on Computer Vision*. Springer, 2014.

[LOG19] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar S. Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke S. Zettlemoyer, and Veselin Stoyanov. "RoBERTa: A Robustly Optimized BERT Pretraining Approach." *arXiv*, 2019.

[LSM13] Thang Luong, Richard Socher, and Christopher Manning. "Better Word Representations with Recursive Neural Networks for Morphology." In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pp. 104–113, 2013.

[LSS19] Nelson F. Liu, Roy Schwartz, and Noah A. Smith. "Inoculation by Fine-Tuning: A Method for Analyzing Challenge Datasets." In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 2171–2179, 2019.

[MBX17] Bryan McCann, James Bradbury, Caiming Xiong, and Richard Socher. "Learned in translation: Contextualized word vectors." In *NeurIPS*, 2017.

[McC60] John McCarthy et al. *Programs with common sense*. RLE and MIT computation center, 1960.

[MCC13] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. "Efficient estimation of word representations in vector space." *arXiv preprint arXiv:1301.3781*, 2013.

[MF98] George Miller and Christiane Fellbaum. "Wordnet: An electronic lexical database.", 1998.

[MMP92] Anna C McFadden, George E Marsh, Barrie Jo Price, and Yunhan Hwang. "A study of race and gender bias in the punishment of school children." *Education and treatment of children*, pp. 140–146, 1992.

[MPC19] Tao Meng, Nanyun Peng, and Kai-Wei Chang. "Target Language-Aware Constrained Inference for Cross-lingual Dependency Parsing." 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)*, pp. 1117–1128, 2019.

[MSC13] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. "Distributed representations of words and phrases and their compositionality." In *NeurIPS*, 2013.

[MYB19] Thomas Manzini, Lim Yao Chong, Alan W Black, and Yulia Tsvetkov. "Black is to Criminal as Caucasian is to Police: Detecting and Removing Multiclass Bias in Word Embeddings." In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 615–621, 2019.

[MYZ13] Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. "Linguistic Regularities in Continuous Space Word Representations." In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 746–751, 2013.

[PCR15] Haoruo Peng, Kai-Wei Chang, and Dan Roth. "A Joint Framework for Coreference Resolution and Mention Head Detection." In *CoNLL*, p. 10, 7 2015.

[PKR15] Haoruo Peng, Daniel Khashabi, and Dan Roth. "Solving Hard Coreference Problems." In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 809–819, 2015.

[PLR14] Sameer Pradhan, Xiaoqiang Luo, Marta Recasens, Eduard Hovy, Vincent Ng, and Michael Strube. "Scoring Coreference Partitions of Predicted Mentions: A Reference Implementation." In *Proceedings of the 52nd Annual Meeting of* *the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 30–35, 2014.

[PNI18] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. "Deep contextualized word representations." In *NAACL*, 2018.

[PSM14] Jeffrey Pennington, Richard Socher, and Christopher Manning. "GloVe: Global Vectors for Word Representation." In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543, 2014.

[RAG11] Kira Radinsky, Eugene Agichtein, Evgeniy Gabrilovich, and Shaul Markovitch. "A word at a time: computing word relatedness using temporal semantic analysis." In *Proceedings of the 20th international conference on World wide web*, pp. 337–346, 2011.

[RC12] Alexander M Rush and Michael Collins. "A Tutorial on Dual Decomposition and Lagrangian Relaxation for Inference in Natural Language Processing." *Journal of Artificial Intelligence Research*, **45**:305–362, 2012.

[RG65] Herbert Rubenstein and John B Goodenough. "Contextual correlates of synonymy." *Communications of the ACM*, **8**(10):627–633, 1965.

[RLR10] Karthik Raghunathan, Heeyoung Lee, Sudarshan Rangarajan, Nathanael Chambers, Mihai Surdeanu, Dan Jurafsky, and Christopher Manning. "A multi-pass sieve for coreference resolution." In *EMNLP*, pp. 492–501, 2010.

[RN12] Altaf Rahman and Vincent Ng. "Resolving Complex Cases of Definite Pronouns: The Winograd Schema Challenge." In *EMNLP*, pp. 777–789, 2012.

[RNL18] Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. "Gender Bias in Coreference Resolution." In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pp. 8–14, 2018.

[RNS18] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. "Improving language understanding by generative pre-training." 2018.

[RSR20] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. "Exploring the limits of transfer learning with a unified text-to-text transformer." *Journal of Machine Learning Research*, **21**(140):1–67, 2020.

[RVS19] Sebastian Ruder, Ivan Vulić, and Anders Søgaard. "A survey of cross-lingual word embedding models." *Journal of Artificial Intelligence Research*, **65**(1):569–630, 2019.

[RWC19] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. "Language Models are Unsupervised Multitask Learners." 2019.

[RZL16] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. "SQuAD: 100,000+ Questions for Machine Comprehension of Text." In *Proceedings of EMNLP*, 2016.

[SDH19] Nathaniel Swinger, Maria De-Arteaga, Neil Thomas Heffernan IV, Mark DM Leiserson, and Adam Tauman Kalai. "What are the biases in my word embedding?" In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 305–311. ACM, 2019.

[SM00] Jianbo Shi and Jitendra Malik. "Normalized cuts and image segmentation."

IEEE Transactions on pattern analysis and machine intelligence, **22**(8):888–905, 2000.

[SZ14] Karen Simonyan and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." *arXiv preprint arXiv:1409.1556*, 2014.

[SZ15] Karen Simonyan and Andrew Zisserman. "Very Deep Convolutional Networks for Large-Scale Image Recognition." In *ICLR*, 2015.

[VTB15] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. "Show and tell: A neural image caption generator." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3156–3164, 2015.

[WDS19] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. "HuggingFace's Transformers: State-of-the-art Natural Language Processing." *ArXiv*, **abs/1910.03771**, 2019.

[WPN19] Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. "Superglue: A stickier benchmark for general-purpose language understanding systems." In *Advances in neural information processing systems*, pp. 3266–3280, 2019.

[WPR12] Ralph Weischedel, Sameer Pradhan, Lance Ramshaw, Jeff Kaufman, Michelle Franchini, Mohammed El-Bachouti, Nianwen Xue, Martha Palmer, Jena D Hwang, Claire Bonial, et al. "OntoNotes Release 5.0." 2012.

[WWB21] Yuyan Wang, Xuezhi Wang, Alex Beutel, Flavien Prost, Jilin Chen, and Ed H Chi. "Understanding and Improving Fairness-Accuracy Trade-offs in Multi-Task Learning." *arXiv preprint arXiv:2106.02705*, 2021.

[WZY19] Tianlu Wang, Jieyu Zhao, Mark Yatskar, Kai-Wei Chang, and Vicente Ordonez. "Balanced Datasets Are Not Enough: Estimating and Mitigating Gender Bias in Deep Image Representations." In *ICCV*, 2019.

[YDY19] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. "Xlnet: Generalized autoregressive pretraining for language understanding." In *Advances in neural information processing systems*, pp. 5754–5764, 2019.

[YOZ17] Mark Yatskar, Vicente Ordonez, Luke Zettlemoyer, and Ali Farhadi. "Commonly uncommon: Semantic sparsity in situation recognition." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[YZF16] Mark Yatskar, Luke Zettlemoyer, and Ali Farhadi. "Situation recognition: Visual semantic role labeling for image understanding." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[ZC20] Jieyu Zhao and Kai-Wei Chang. "LOGAN: Local Group Bias Detection by Clustering." In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1968–1977, 2020.

[ZG19] Han Zhao and Geoff Gordon. "Inherent tradeoffs in learning fair representations." *Advances in neural information processing systems*, **32**:15675–15685, 2019.

[ZKK21] Jieyu Zhao, Daniel Khashabi, Tushar Khot, Ashish Sabharwal, and Kai-Wei Chang. "Ethical-Advice Taker: Do Language Models Understand Natural Language Interventions?" In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 4158–4164, 2021.

[ZMH20] Jieyu Zhao, Subhabrata Mukherjee, Saghar Hosseini, Kai-Wei Chang, and Ahmed Hassan Awadallah. "Gender Bias in Multilingual Embeddings and Cross-Lingual Transfer." In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 2896–2907, 2020.

[ZSZ19] Pei Zhou, Weijia Shi, Jieyu Zhao, Kuan-Hao Huang, Muhao Chen, Ryan Cotterell, and Kai-Wei Chang. "Examining Gender Bias in Languages with Grammatical Gender." 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)*, pp. 5276–5284, 2019.

[ZWY17] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. "Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints." In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 2979–2989, 2017.

[ZWY18] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. "Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods." In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pp. 15–20, 2018.

[ZWY19] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. "Gender Bias in Contextualized Word Embeddings." In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 629–634, 2019.

[ZZL18] Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. "Learning Gender-Neutral Word Embeddings." In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 4847–4853, 2018.