Detection and Mitigation of Gender Bias in Natural Language Processing

by

Hillary Dawkins

A Thesis

presented to

The University of Guelph

In partial fulfilment of requirements
for the degree of

Doctor of Philosophy

in

Computational Sciences

Guelph, Ontario, Canada

ABSTRACT

Detection and Mitigation of Gender Bias
in Natural Language Processing

Hillary Dawkins
University of Guelph, 2023

Advisors:
Dr. Judi McCuaig
Dr. Daniel Gillis

This thesis contributes to our collective understanding of how gender bias arises in natural language processing systems, provides new detection and measurement tools, and develops mitigation methods. More specifically, we quantify and reduce bias within pre-trained computational resources, both word embeddings and language models, such that unwanted outcomes produced by the system are mitigated. Unwanted outcomes include any system prediction that is unduly influenced by the presence of gender words or the latent concept of gender in language (e.g. when an NLP system is unable to predict that "she" refers to a doctor).

On the theme of detection, we make two new observations on how gender bias can manifest in system predictions. Firstly, gender words are shown to carry either marked or default values. Default values may pass through systems undetected, while marked values influence prediction outcomes. Secondly, unwanted latent inferences are detected, due to a shared gender association. We contribute two new test sets, and one enhanced test set, for the purpose of gender bias detection.

On the theme of mitigation, we develop successful debiasing strategies applied to both types of pre-trained resources.

Acknowledgments

First and foremost, I would like to thank my advisors, Dr. Judi McCuaig and Dr. Daniel Gillis, for their continuous support and encouragement. This thesis would not be possible without their patience, understanding, and helpful feedback throughout all stages of this program. I would also like to thank the remaining members of my advisory committee, Dr. Stefan C. Kremer and Dr. Graham Taylor, for the valuable suggestions they provided over the course of this research.

Next I'd like to thank all my professors, advisors, and collaborators throughout the years both before and during my Ph.D. program, including those in the physics department at the University of Guelph, the Institute for Quantum Computing, the Open University, the University of Washington, the Vector Institute, and the National Research Council of Canada. I have been very fortunate to meet and collaborate with so many talented researchers across many disciplines.

Lastly, thank you to my parents for making my education possible.

Table of Contents

Abstract ii

Acknowledgements iii

List of Tables vii

List of Figures ix

1 Introduction 1
- 1.1 Motivation . . . 1
- 1.2 Contributions and thesis statement . . . 2

2 Background 4
- 2.1 Introduction to word embeddings . . . 4
- 2.2 Observation of gender bias and early debiasing approaches . . . 6
- 2.3 Second-wave debiasing methods: Quantifying indirect bias . . . 8
- 2.4 Beyond static word embeddings: contextual embeddings and pre-trained language models . . . 10
- 2.5 Quantifying and mitigating gender bias in BERT . . . 12

3 Marked Attribute Bias in Natural Language Inference 14
- 3.1 Introduction . . . 17
- 3.2 Marked Attribute Bias in Natural Language Inference . . . 20
  - 3.2.1 Background: Natural Language Inference . . . 20
  - 3.2.2 Observation of marked vs. default attribute bias . . . 20
- 3.3 Analysis of the existing debiasing schemes applied to MAB . . . 22
  - 3.3.1 Debiased embeddings . . . 22
  - 3.3.2 Explicit gender words test set and error definitions . . . 24
  - 3.3.3 Latent gender carriers . . . 25
- 3.4 Intrinsic bias measures . . . 29
- 3.5 Multi-dimensional information-weighted soft projection . . . 32
- 3.6 Conclusion . . . 35

4 Second Order WinoBias (SoWinoBias) Test Set for Latent Gender Bias Detection in Coreference Resolution **37**
4.1 Introduction . . . 40
4.2 Bias Statement . . . 42
4.3 Debiasing methods . . . 43
4.3.1 Neutralization of static word embeddings . . . 43
4.3.2 Data augmentation . . . 45
4.4 Detection of gender bias in coreference resolution: Experimental setup 45
4.4.1 WinoBias . . . 46
4.4.2 SoWinoBias . . . 48
4.5 Relationship to embedding space properties . . . 50
4.5.1 Single-attribute WEAT . . . 50
4.5.2 Clustering and Recoverability . . . 52
4.5.3 Gender-based Illicit Proximity Bias . . . 54
4.6 Conclusion . . . 55

5 Projective Debiasing Methods for Pre-trained Language Models **57**
5.1 Introduction . . . 60
5.2 Enhanced StereoSet for quantifying intrinsic bias . . . 63
5.3 Downstream task: Measuring gender bias using NLI . . . 68
5.4 Debiasing interventions applied to BERT . . . 69
5.5 Results and key observations . . . 73
5.6 Summary . . . 77

6 Ethical Considerations and Limitations **79**

7 Conclusion **81**

Bibliography **83**

A Inventory of Research Tools and Data **93**
A.1 Tools and resources pertaining to static word embeddings . . . 93
A.2 Tools and resources pertaining to pretrained language models . . . 97

B Marked Attribute Bias (Chapter 3) Supplementary Material **99**
B.1 Intrinsic bias measures and correlation results . . . 99

C SoWinoBias (Chapter 4) Supplementary Material **103**
C.1 SoWinoBias test set vocabulary 103
C.2 Results expanded by adjective polarity 103

D Debiasing BERT (Chapter 5) Supplementary Material **105**

List of Tables

3.1 Results of the marked attribute test set on explicit gender words. . . 26
3.2 Results of marked attribute test set on stereotypical occupations. . . 27
3.3 Results of marked attribute test set on names. 28
3.4 Results for word similarity and analogy benchmarks. 35

4.1 Results on coreference resolution test sets, WinoBias and SoWinoBias, using various debiased word embeddings. 47
4.2 Reported results on the OntoNotes (baseline) and WinoBias test sets by various debiasing methods when the coreference system is built using a model with static word embeddings only. 48
4.3 Single-Attribute WEAT association strength between gender and female-stereotyped adjectives with significance values. 51
4.4 Clustering, Recoverabilty, and Gender-induced Proximity Estimate on different embedding spaces. 53

5.1 Example triple from StereoSet. 63
5.2 Examples of bad triples found by manually screening StereoSet's inter-sentence gender development set. 64
5.3 Examples of triple pairs in the augmented StereoSet. 66
5.4 Examples of triple pairs exhibiting BERT's unequal NSP capability among gender-swapped inputs. 67
5.5 Summary of best solutions by intervention level. 75
5.6 Average accuracy on the SNLI benchmark by intervention level. . . . 77

A.1 Debiasing methods for static word embeddings. 94
A.2 Intrinsic bias measures for static word embeddings. 95
A.3 Downstream bias assessments using static word embeddings as input. 95
A.4 Vanilla assessments of word embedding quality. 96
A.5 Debiasing methods for pre-trained language models. 97
A.6 Intrinsic bias measures for contextual word embeddings and language models. 97
A.7 Downstream gender bias assessments specific to language models. . . 98

A.8 Vanilla language model resources. 98

B.1 Intrinsic bias measures of interest on the experimental set of embeddings. 101
B.2 Pearson correlation matrix between intrinsic bias measures (and marked attribute error) on the experimental set of embeddings. 102

C.1 Results on SoWinoBias test set by adjective polarity. 104

D.1 Full results for all interventions applied to BERT. 105

List of Figures

5.1 Abstracted representation of BERT. 70
5.2 Visual representation of the proposed intervention setting space as a tree. 72

Chapter 1

Introduction

1.1 Motivation

As artificial intelligence becomes increasingly integrated with our everyday life, it is crucial to consider how these systems interact with our social environment. In this thesis, we investigate gender bias in natural language processing systems. Within modern NLP systems, it is now standard practice to use off-the-shelf pre-trained resources (either static word embeddings, contextual embeddings, or language models). Pre-trained resources successfully capture many complex latent dimensions of language, making them extremely powerful computational tools, however, they also capture intrinsic human biases. Hidden gender associations in pre-trained resources can cause or reinforce adverse outcomes.

To give a few concrete examples, consider how gender-biased word embeddings can amplify bias through human-model interactions. Information retrieval systems including web searches use document similarity scores, quantified using word vector similarities, to score and retrieve the relevant information. As a consequence, search queries such as "scientist" are more likely to return pages for scientists with male names (given a biased similarity between a pre-computed word vector $\vec{scientist}$ and a latent encoding of gender information). In this case, bias is amplified when a socially held stereotype is reinforced after observing the search results. A similar argument can be made for automated resume filtering systems; the presence of a male name or pronouns will increase the similarity score of a resume matched against male-stereotyped occupations or skills.

Another example can be observed in the task of coreference resolution, the task of matching all words in a sentence that refer to the same entity. The study of gender-biased coreference resolution finds that a model is systematically unable to pair stereotyped occupation words with mismatched stereotyped characteristics. This type of disparity creates representational harm by implying (for example) that occupations typically associated with one gender cannot have attributes typically associated with another.

Towards combating these types of outcomes, we analyze existing debiasing methods as well as develop new debiasing schemes. Specifically, we are interested in how intrinsic bias observed in pre-trained resources correlates with realized bias in downstream settings. After developing a clear understanding of the link between the two, we develop new debiasing methods. The research presented in this thesis focuses on post-processing debiasing techniques applied to either pre-trained word embeddings or pre-trained language models. Both resources are used extensively in NLP applications and could easily be substituted with their debiased variants. The impact is that less biased outcomes are produced in downstream settings, while good baseline performance is maintained.

1.2 Contributions and thesis statement

The goal of this thesis is to mitigate gender-biased outcomes produced by NLP systems by debiasing pretrained resources (both static word embeddings and language models) via simple post-processing methods. We focus on post-processing methods because they require minimal additional computation, and they are easy to concatenate with existing methods. Throughout, the performance of a debiasing method is quantified by its ability to eliminate or reduce unequal outcomes across binary genders (e.g. as differences in predictions across gender) without affecting task accuracy. As we will come to appreciate in Chapter 2 Background, mitigating bias

in pretrained resources often requires an understanding of how intrinsic bias (some innate property of the pretrained resource) correlates with observable bias in downstream applications. Therefore supporting contributions to this thesis are to propose and investigate intrinsic bias measures.

In particular, the types of contributions that will support the research goal are as follows:

1. Support: Identify and define unwanted outcomes in NLP systems
2. Support: Provide test sets for the purpose of systematically quantifying the unwanted outcomes and comparing across systems
3. Support: Analysis of how intrinsic bias measures are predictive of observable bias
4. Support: New intrinsic bias measures
5. Culminating: New debiasing methods

We provide two debiasing schemes, one for static word embeddings and one for pretrained language models, that are shown to outperform comparable debiasing methods in mitigating gender-biased outcomes in downstream NLP tasks.

Chapter 2

Background

2.1 Introduction to word embeddings

Natural language processing tasks rely on our ability to represent words meaningfully. In language, words are not just standalone symbols. Words have synonyms and antonyms, hyponyms and hypernyms, share meaningful root words, and follow some syntactic regularities. Translating words from units of natural language to useful computational units gives us word embeddings, real-valued vector representations of words. Ideally, these representations should express the semantic relationships and nuances of natural language in a quantifiable way.

At the core of word representation learning is Harris' distributional hypothesis [31, 26], often stated as "you shall know a word by the company it keeps". In other words, similar words appear in similar contexts (the surrounding words), and we assume that word representations can be derived from word-context information. Traditional (pre-deep learning) statistical NLP methods summarize the word-context information of a corpora into a matrix such as a simple co-occurrence count matrix, or a pairwise mutual information matrix [13]. Given some type of information matrix between words (columns) and contexts (rows), word vectors can be taken as either the columns themselves or as lower-dimensional projections obtained via matrix factorization [20, 63].

Neural-network-based approaches provide a new way to embed word representations as real-valued vectors [3, 15, 49, 50]. Generally, given an input word represented as a one-hot vector, a neural network is trained for some language modelling task.

The first (or some combination) of hidden layers in the network can be viewed as a learned distributed representation that best encodes the word meaning. Early methods use the typical language model setup where the next word should be predicted given only the preceding words in a sentence. According to the distributional hypothesis, similar words will have similar vectors because their output model predictions (the context) should also be similar.

Interest in pretrained distributed word representations as a standalone resource exploded around 2013 with the release of Google's word2vec [46, 47]. Because the goal was solely to train word embeddings as a sharable resource rather than perform a language modelling task, the training objective was modified to predict the entire context (both preceding and following a given word). The well known word2vec embeddings are derived in a series of papers exploring various training objectives, eventually culminating in the skipgram with negative sampling (SGNS) task. Training proceeds by sliding a context window over a text sequence, centering on one word at a time. The task is to distinguish correct context words from negative samples (words randomly sampled from the full vocabulary) given the center word.

Quickly following word2vec came another of the most popular pre-trained embeddings in use today: GloVe (named for Global Vectors) [56]. The idea of GloVe is to combine a neural training procedure with a global co-occurrence information matrix (doing away with the sliding context window setup of word2vec). Interestingly, Levy and Goldberg [42] analyze the objective function of both the SGNS and GloVe training procedures and find that the optimal solution in both cases is equivalent to factorizing a shifted pairwise mutual information matrix (hearkening back to traditional statistical methods), just with slightly different reconstruction errors. This observation opens an interesting question on why neural-network-derived embeddings often seem more performant.

In any case, word embeddings are now ubiquitous tools in natural language processing systems. Simply using pretrained word embeddings as part of the model input

(instead of one-hot word encodings or bag-of-words document encodings) improves performance on a wide range of tasks including document classification [66], question answering [71], sentiment analysis [68], named entity recognition [72], parsing [67], and various other token classification tasks [16]. Their utility is owing to the remarkable success in mapping relationships among words to the intrinsic quantifiable properties of vectors. For instance, word relatedness can be easily quantified as cosine similarity between vectors, and such measures are known to match well with human annotations. Furthermore, the embedding space is known to encode both semantic and syntactic relationships among words as simple linear transformations [48]. For example, vector relationships such as $\vec{king} - \vec{queen} \approx \vec{man} - \vec{woman}$ and $\vec{great} - \vec{greatest} \approx \vec{small} - \vec{smallest}$ are observed. This linear structure provides a way to generate analogies in natural language from vector addition in the embedding space. In all kinds of related tasks, the vector space is known to encode semantic meaning surprisingly well. Pennington [56] notes that this structure implies that word embeddings capture latent dimensions of meaning, and therefore exhibit the multi-clustering idea of distributed representations [2]. It is now routine practice to measure word embedding "quality" based on these intrinsic vector properties [65, 79, 11, 32, 46].

2.2 Observation of gender bias and early debiasing approaches

In addition to capturing helpful natural language relationships and regularities, word embeddings are known to possess gender-biased properties. For instance, it is known that the same analogy generation property that produces the celebrated "man is to king as woman is to queen" result also predicts "man is to programmer as woman is to homemaker". This observation was first reported in a seminal work [7] that sparked interest in developing debiased word embeddings.

Post-processing debiasing schemes applied to static word embeddings are usually motivated by recognizing some intrinsic measure of bias in the embedding space, and then working to reduce or eliminate that property. Early work focuses on the idea of a "gender direction", $\vec{g}$, within the embedding space; a subspace that mostly encodes the difference between binary genders. In its first appearance [7], $\vec{g}$ is defined as the first principal component summarizing a set of difference vectors $\{\vec{d_i} = \vec{f_i} - \vec{m_i}\}$ between defined sets of paired female and male-associated words $f_i \in F$, $m_i \in M$[1]. Some subsequent works (e.g. [29]) prefer to take a simpler definition as $\vec{g} = \vec{she} - \vec{he}$, although the principle remains the same.

Because semantic similarity between words is encoded as cosine similarity between vectors, any non-zero projection of word $\vec{w}$ onto $\vec{g}$ implies that $\vec{w}$ is more similar to one gender over another. In the case of ideally gender-neutral words (e.g. nurse, doctor, programmer, homemaker), this is viewed as an undesirable property. Direct bias quantifies the extent of this uneven similarity[2]:

$$DB(N) = \frac{1}{|N|} \sum_{\vec{w} \in N} |\cos(\vec{w}, \vec{v})| \tag{2.1}$$

where N is the set of ideally gender-neutral words (typically taken as the full vocabulary excluding explicitly gendered words such as boy, girl, aunt, uncle, etc.).

The Hard Debias method [7] is a post-processing technique that projects all gender-neutral words onto the nullspace of $\vec{g}$. Therefore, the direct bias is made to be zero by definition. Additionally, gender-neutral words are made equidistant to pairs of words in a defined equalization set.

A related retraining method proposed by Zhao et al. [82] uses a modified version of GloVe's original objective function with an additional incentive to reduce direct bias for gender-neutral words, resulting in GN-GloVe embeddings. Rather than allowing

[1] M = {he, his, man, John, himself, son, father, guy, boy, male}, F = {she, her, woman, Mary, herself, daughter, mother, gal, girl, female}

[2] The original definition also includes a strictness parameter c, here set to 1 as is commonly done in subsequent works.

gender information (quantified here as direct bias) to be distributed across the entire embedding space, the method explicitly sequesters the protected gender attribute to the final component. Therefore, the first $dim - 1 = 299$ components are taken to be the gender-neutral embeddings, denoted GN-GloVe(w_a).

These early methods are successful in mitigating harmful analogies generated by word embeddings in relation to gender-stereotyped occupations.

2.3 Second-wave debiasing methods: Quantifying indirect bias

An influential critique paper by Gonen and Goldberg [29] demonstrates that minimizing direct bias does not eliminate bias in the embedding space entirely. Rather, words that tend to cluster together due to gender bias (e.g. nurse, teacher, secretary, etc.) still cluster together in the null space of the gender direction. Furthermore, the original gender labels of words (assigned according to direct bias in the original embedding space) can be recovered with a very high degree of accuracy given only the debiased representations. These properties are known as clustering and recoverability bias. Motivated by these observations, the second wave of debiasing methods focus on reducing clustering and recoverability, as well as proposing new instrinsic bias measures for quantifying the indirect bias.

Loosely defined, the indirect bias refers to the gender-induced similarity between gender-neutral words. For instance, semantically unrelated words such as "sweetheart" and "nurse" may appear quantitatively similar due to some shared gender association. This may be quantified on the embedding space either by gender-induced proximity among embeddings, or by residual gender cues that could be learned by a classifier.

One definition (first given in [7]) measures the relative change in similarity after

removing direct gender associations as

$$\beta(\vec{w}, \vec{v}) = \frac{1}{\vec{w} \cdot \vec{v}} \left(\vec{w} \cdot \vec{v} - \frac{\vec{w}_\perp \cdot \vec{v}_\perp}{\|\vec{w}_\perp\| \|\vec{v}_\perp\|} \right), \tag{2.2}$$

where $\vec{w}_\perp = \vec{w} - (\vec{w} \cdot \vec{g})\vec{g}$, however this relies on a limited definition of the original gender association, namely direct bias.

The gender-based illicit proximity bias (GIPE) [37] quantifies indirect bias by incorporating β into a graph-weighted holistic view of the embedding space. Firstly, the gender-based proximity bias of a single word w, denoted $\eta(w)$, is defined as the proportion of N-nearest neighbours $\{n_i\}$ with indirect bias $\beta(n_i, w)$ above some threshold θ. Intuitively, this is the proportion of words that are close by solely due to a shared gender association. The GIPE extends this word-level measure to a vocabulary-level measure using a weighted average over $\eta(w)$. The corresponding debiasing method, Repulse-Attract-Neutralize (RAN), attempts to repel undue gender proximities among gender-neutral words, while keeping word embeddings close to their original learned representations [37].

A related but distinct notion of indirect bias is to measure whether gender associations can be predicted from the word representation. The Iterative Nullspace Linear Projection method (INLP) achieves linear guarding of the gender attribute by iteratively learning the most informative gender subspace for a classification task, and projecting all words to the orthogonal nullspace [61, 17]. After sufficient iteration, gender information cannot be recovered by a linear classifier. Ravfogel et al. [61] analyze the effect of linear nullspace projection on the clustering and recoverability experiments proposed by [29].

Indirect bias in the embedding space is viewed as an undesirable property a priori, but we do not yet have a good understanding of the effect on downstream applications.

2.4 Beyond static word embeddings: contextual embeddings and pre-trained language models

Despite the great success of static pre-trained word embeddings, some notable deficits remain. For one, static word embeddings cannot handle polysemous words (e.g. "bank" as in a river bank vs. a financial bank); all possible meanings of a word must share the same word embedding. Composition can also be a challenge, especially for noun phrases and idiomatic expressions (e.g. "hot dog" $\neq$ "hot" + "dog"). Ideally, word representations should be fluid and be able to reflect the context in which words appear. This idea is captured by early contextual word embeddings such as ELMo [57], in which word embeddings are computed and cached given the surrounding words.

Taking things a step further, we can imagine building a complete sentence representation that fully encodes the context of its composite words. Two principles of language that give meaning to a sentence from its individual words are i) composition and ii) hierarchical structure [70]. Following these principles, we would like to compute the sentence representation both from contextual word representations, as well as the global parse structure and the relative importance of words. These are the ideas thought to be captured by transformer-based language models such as the GPT family [59, 60, 10] and BERT [24].

BERT and its related spin-offs use stacked layers of the transformer encoder [73]. The transformer architecture revolutionized modern language models by making use of a self-attention mechanism [1]. At each layer of the model, contextualized word representations are output as combinations of word representations occurring at the previous layer. The attention weights control the relative importance of each word in computing the next representation. The final model output contains both contextualized representations for each word, as well as a complete sentence representation. The sentence embedding can be used for further tasks such as text classification.

Although conceptually similar, BERT outperformed the original GPT model by introducing bidirectionality as a key ingredient (all words in the input can attend to all other words). At the time of BERT's release, the model shattered benchmarks across the field from question answering, coreference resolution, machine translation and so on, including a 4 point improvement in the General Language Understanding Evaluation (GLUE) benchmark [24].

The immense success of BERT has since inspired many studies into understanding how the model encodes information (informally, asking what does BERT "know"). Generally, there are 3 main investigative procedures:

1. Direct language modelling tasks
    - Since BERT is a language model, its performance on carefully constructed inputs can quantify its innate syntactic ability, semantic understanding, or reasoning about world knowledge [28].
2. Probing classifiers applied to internal vector representations
    - Simple linear classifiers are applied to BERT's internal representations. e.g. Can the part-of-speech tag for a word be determined with high accuracy given only BERT's contextual embedding for the word as input? The answers to such questions are thought to give insight on what information BERT encodes.
3. Analyzing attention weights
    - Attention maps are analyzed for patterns such as direct objects attending to verbs, determiners attending to nouns, coreference resolution, etc. This can be done qualitatively using visualization tools [74] or systematically by treating attention weights as classifiers [14].

Collectively, this line of research has been given the name Bertology (see [62] for a thorough review). In reference to the goals of this thesis, the principles of Bertology

can be applied to study the model's intrinsic bias properties and relation to unwanted outcomes.

2.5 Quantifying and mitigating gender bias in BERT

Quantifying gender bias in BERT is done both through intrinsic bias measures following the methods of Bertology, as well as performance on downstream applications.

Firstly, similarity-based measures can be used to quantify bias within BERT's internal representations, just as is done for static embeddings. Kurita et al. [38] modify the word embedding association test (WEAT) for BERT, and May et al. [44] provide a conceptually similar test for full sentence embeddings. However, these cosine-based intrinsic bias measures are found to be inconsistent on contextual embeddings. Therefore, we should be cautious of using these measures as evidence of successful mitigation.

Using the idea of probing classifiers, there are attempts to understand how, or to what extent, BERT encodes gender information. For example, the task of gender pronoun resolution (GPR) asks whether gender labels of coreferent pronouns can be classified given their internal representations [38, 76]. This type of assessment is reminiscent of recoverability experiments performed on static word embeddings.

Perhaps most naturally, direct language modelling tasks can be used to assess gender bias in BERT. Given a purposefully constructed input containing a masked token, the relative probabilities of the predicted token are used to quantify bias. Example inputs might include "[MASK] is a programmer" [38] or "girls tend to be more [MASK] than boys" [52]. StereoSet [52] is a robust test set designed for this purpose. Although these types of language modelling assessments are downstream in nature, they can also be said to be intrinsic to BERT. That is, they do not use any additional training resources that might introduce additional biases.

Beyond language modelling tasks designed specifically for BERT, the traditional downstream applications for gender bias assessment still apply (those popularized throughout the static debiasing era). These include coreference resolution, natural language inference, and the equity evaluation corpus for sentiment analysis. However, these tasks involve fine-tuning BERT with additional data and additional task-specific architecture, possibly introducing confounding variables.

In terms of bias mitigation, there exist two recent post-processing methods [4, 43] and one related retraining method [27]. The post-processing methods are most comparable to our proposed debiasing scheme (see Chapter 5). Both methods add projection layers after each encoder block in BERT's architecture, which project all words to the nullspace of a contextualized gender direction. Essentially this is analogous to the Hard Debias method, translated to the realm of pretrained language models. The two methods differ only in how they compute the contextualized gender direction.

Beyond post-processing projective methods, DEBIASBERT is a model obtained by retraining BERT with modified objectives. The new objective functions are inspired by static debiasing concepts, including an equalization loss (which motivates equal prediction probabilities over gender-defined pair words), and a declustering loss (which motivates removal of implicit gender associations). During retraining, hyperparameters are adjusted according to performance on the SEAT evaluation for intrinsic gender bias.

Presently, evaluation and comparison across debiasing methods are lacking. Studies which have focused on bias quantification tend to evaluate only undebiased language models. Studies which introduce debiasing methods tend to compare only with undebiased baselines.

Chapter 3

Marked Attribute Bias in Natural Language Inference

In the first of the paper-based chapters, we set out on our exploration of the link between intrinsic bias measures within the word embedding space and realized bias in downstream settings.

Recall that during the post-Hard Debias era (2017-2019), many new debiasing methods were proposed targeting both direct and indirect intrinsic bias. Typically, each of these new proposals (refer to Table A.1) would claim reduced intrinsic bias using a small subset of the available measures (refer to Table A.2) and improved performance on one downstream bias test set, typically WinoBias (refer to Table A.3).

The result was a muddy and incomplete picture of the correlation between intrinsic bias and observed bias in downstream predictions. That is, the implicit assumption that reducing intrinsic bias results in less biased systems was not very well founded. Furthermore, the types of downstream biased observations were limited to stereotyped associations between gender and occupations during this time.

In this first research project, we move the narrative forward in two ways. Firstly, we expand our understanding of how gender bias arises in downstream settings and publish a corresponding test set. This new observation is named marked attribute bias, and is distinct from previous observations of stereotypical associations. Secondly, we perform a systematic correlation analysis using all available debiased embeddings and intrinsic bias measures. We seek to understand whether a correlation analysis

can guide the development of a successful debiasing scheme.

This chapter is based on the publication

H. Dawkins. *Marked attribute bias in natural language inference.* In Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 4214–4226, On- line, Aug. 2021. Association for Computational Linguistics. [18],

reproduced here with very minor notational edits. We thank Dr. Daniel Gillis, Dr. Judi McCuaig, Dr. Stefan C. Kremer, Dr. Graham Taylor, and anonymous reviewers for their feedback and discussion on this work. The tangible contributions resulting from this publication are the Marked Attribute Bias test set and the MISP-debiased word embeddings, which are made publicly available, and can be used for detection and mitigation of gender bias respectively.

ABSTRACT

Reporting and providing test sets for harmful bias in NLP applications is essential for building a robust understanding of the potential risks in using such applications. We present a new observation of gender bias in a downstream NLP application: marked attribute bias in natural language inference. Bias in downstream applications can stem from training data, word embeddings, or be amplified by the model in use. However, focusing on biased word embeddings is potentially the most impactful first step due to their universal nature. Here we seek to understand how the intrinsic properties of word embeddings contribute to this observed marked attribute effect, and whether current post-processing methods address the bias successfully. An investigation of the current debiasing landscape reveals two open problems: none of the current debiased embeddings mitigate the marked attribute error, and none of the intrinsic bias measures are predictive of the marked attribute effect. By noticing that a new type of intrinsic bias measure correlates meaningfully with the marked attribute effect, we propose a new postprocessing debiasing scheme for static word embeddings. The proposed method applied to existing embeddings achieves new best results on the marked attribute bias test set.

3.1 Introduction

Pre-trained distributed representations of words (a.k.a. word embeddings) are ubiquitous tools in natural language processing (NLP). Their utility is owing to the remarkable success in mapping semantic and syntactic relationships among words to linear relationships among real-valued vectors. For instance, analogy generation using vector addition on word embeddings (e.g. Tokyo is to Japan as Paris is to France) was taken to be an early measure of word embedding quality. In all kinds of related tasks, the vector space is known to encode semantic meaning surprisingly well [56, 47, 48]. However, harmful gender-biased properties of word embeddings are also known to exist. Later is was observed that the same analogy generation property that produced the celebrated "man is to king as woman is to queen" analogy would also predict "man is to programmer as woman is to homemaker" [7]. This observation sparked interest in developing debiased word embeddings.

Post-processing debiasing schemes are usually motivated by recognizing some intrinsic measure of bias in the embedding space, and then attempting to reduce that intrinsic bias. Early work (2016-2017) focused on the idea of a "gender direction" vector within the embedding space, loosely defined as the difference vector between female and male attribute words. It was noted that any non-zero projection of a word onto the gender direction (termed direct bias) implied that the word was more related to one gender over another. In the case of ideally gender-neutral words (e.g. doctor, nurse, programmer, homemaker), this was viewed as an undesirable property. The first debiasing methods, Hard Debias [7] and Gender Neutral-GloVe [82], worked to minimize or eliminate the direct bias, and were shown to be successful in mitigating harmful analogies generated by word embeddings in relation to gender-stereotyped occupations.

An influential critique paper by [29] demonstrated that minimizing direct bias did not eliminate bias in the vector space entirely. Rather, words that tended to cluster

together due to gender bias (e.g. nurse, teacher, secretary, etc.) would still cluster together in the nullspace of the gender direction. Furthermore, the original bias could be recovered by classification techniques using only the debiased word embeddings as input. These observations were termed cluster and recoverability bias.

The next wave of debiasing methods (2019-present) focused on reducing cluster and recoverability bias while proposing new metrics to systematically quantify the indirect bias of the embedding space (e.g. the Gender-based Illicit Proximity Estimate, introduced by [37]). While these new debiasing schemes do reduce indirect bias in multiple ways, there is a general lack of connection to downstream applications such as coreference resolution, natural language inference (NLI) and sentiment analysis.

Current gender-bias evaluation tests (GBETs) in widespread use include the WinoBias test set [81], designed to measure bias in coreference resolution systems using stereotypical occupations as a probe, and the NLI test set [21], designed to measure stereotypical inferences again using occupations as the concept of interest. More commonly used evaluations include the Word Embedding Association Test (WEAT) [12], and the analogy generation test SemBias [82]. However these tests solely evaluate the vector properties of the word embeddings, without any connection to downstream applications. Adding to the library of downstream GBETs is essential in building a robust understanding of gender bias in NLP applications [69].

Here we introduce a new observation of gender-biased predictions in a downstream task, namely "marked attribute bias" in natural language inference, and develop corresponding GBETs. Marked attribute bias refers to the language model's tendency to predict that "person" implies "man" (the default attribute), while simultaneously understanding that "person" does not necessarily imply "woman" (the marked attribute). Marked attribute bias was found to exist on explicitly defined gender words (e.g. man, woman, etc.), and persist on implicit gender words (e.g. names) as well as latent gender-carriers (e.g. stereotypical occupations).

An analysis of the currently available debiased embeddings reveals that none

are able to successfully mitigate marked attribute bias. Furthermore, none of the currently proposed measures of intrinsic bias on the embedding space are predictive of the marked attribute effect. We define a new measure of intrinsic bias that was found to correlate with the marked attribute effect better than any currently available metric. Using this insight, we introduce a new debiasing scheme: Multi-dimensional Information-weighted Soft Projection. Applying MISP to an existing debiased embedding achieves the lowest observed marked attribute bias error on the new benchmark test set that we provide.

Summary of main contributions:

1. We present a new observation of gender bias in a downstream NLP application: marked attribute bias (MAB). The MAB test sets are made available in order to expand the current set of GBETs.

2. An analysis of current debiasing methods and current intrinsic bias measures finds that none sufficiently mitigate the error, and likewise none sufficiently explain the effect. This observation creates two new open problems.

3. We propose a new measure for quantifying intrinsic bias on the embedding space: Multi-dimensional Information-weighted Direct Bias (MIDB). This measure was found to correlate meaningfully with the marked attribute effect.

4. We introduce a new debiasing scheme: Multi-dimensional Information-weighted Soft Projection. MISP-debiased embeddings obtain new best performance on the MAB test set.

3.2 Marked Attribute Bias in Natural Language Inference

3.2.1 Background: Natural Language Inference

Natural language inference is one of the pillars of natural language understanding. It is the task of determining whether a hypothesis sentence is (neutral, entailed, or contradicted) with respect to a premise sentence. For example:

Premise: A choir sings in the church.

Hypothesis: The church is filled with the sound of singing. (Correct prediction: Entail)

Dev et al. [21] previously used NLI as a test case for gender bias with respect to occupations. For example, consider:

Premise: A doctor prepared a meal.

Hypothesis 1: A man prepared a meal. (N)

Hypothesis 2: A woman prepared a meal. (N)

This inference task essentially asks the question: is "doctor" a subset of man/woman? I.e. if someone is a doctor, must they be a man? While both hypothesis sentences should receive a neutral prediction (as "doctor" does not imply any specific gender), hypothesis 1 will more likely receive an entailment, while hypothesis 2 will more likely receive a contradiction, given biased word embeddings. The corresponding GBET was published by [21] and contains 1936512 sentence pairs in the form [A **occupation** *verb object*] $\rightarrow$ [A **gender word** *verb object*]. Throughout this paper, we will use the notation [Sentence A] $\rightarrow$ [Sentence B] to mean that premise Sentence A is paired with hypothesis Sentence B.

3.2.2 Observation of marked vs. default attribute bias

Marked vs. default attribute bias occurs whenever a default attribute (e.g. male, white, etc.) is assumed, and a marked attribute has to be explicitly stated or becomes a defining trait. In the context of the natural language inference task, consider the

sentence pair:

Premise: A person prepared a meal.

Hypothesis 1: He prepared a meal. (N)

Hypothesis 2: She prepared a meal. (N)

Due to the language model's[1] tendency to predict that "person" implies a male (default) attribute, the first hypothesis sentence will have a prediction probability vector shifted towards Entail. However the same language model would tend towards a Neutral prediction for the second hypothesis, recognizing that "person" does not necessarily imply female (the marked attribute). To put it another way, this inference task essentially asks the question: is "person" a subset of man/woman? When presented with a masculine form, the model answers: yes (entailment), a person must be a man. When presented with a feminine form, the model answers: not necessarily (neutral), a female has an attribute (gender) that not all persons have. The name "Marked Attribute Bias" therefore derives from the observation that masculine forms are unmarked with respect to gender, whereas female forms carry a marked gender attribute.

In this particular example, the model trained with (original) GloVe[2] word embeddings [56] gives a probability distribution (N, E, C) of (0.0538, **0.929**, 0.0177) for hypothesis 1 and (**0.687**, 0.238, 0.0750) for hypothesis 2.

Note that although the MAB test construction appears similar to [21], it is actually measuring quite a distinct effect. The [21] test set measures associations between gender and some concept of interest (occupations). The MAB test set measures something more general and pervasive; it measures how gender words carry meaning,

[1] All NLI models mentioned throughout this paper are based on the Decomposable Attention Model [55] with intra-attention, trained on the Stanford Natural Language Inference training dataset [8] (trained for 100 epochs; learning rate 0.025; weight decay 1e-5; dropout rate 0.2; 200 hidden units; approximately 10^4 total model parameters). All the code and data needed to reproduce results mentioned in this paper are available at https://github.com/hillary-dawkins/MAB.

[2] Taken as the GloVe embeddings trained on the Common Crawl corpus for 840B tokens; available at https://nlp.stanford.edu/projects/glove/. Results were not found to vary significantly among undebiased embeddings.

independent of any concept of interest.

Achieving the correct prediction probability of $(N, E, C) = (1, 0, 0)$ on both sentences is difficult because it requires the language model to be attribute-aware (in this case gender-aware) while not using the gender attribute to alter predictions when it would be inappropriate to do so.

3.3 Analysis of the existing debiasing schemes applied to MAB

In order to investigate the presence of systematic marked attribute bias in natural language inference, we construct three types of tests: bias on explicit gender words, implicit gender carriers, and latent gender carriers. We wish to understand the depth and persistence of the marked attribute effect, as well as how it is handled by current debiasing methods. Firstly we provide a brief description of the current debiasing methods to be analyzed. Next we provide details of the test sets and report results.

3.3.1 Debiased embeddings

Within the scope of this paper, we focus on post-processing techniques applied to static word embeddings. These types of methods are computationally inexpensive, easy to combine, and are independent of the base embedding. In addition, we include GN-GloVe, one of the highly cited retraining methods. Notationally, we specify embeddings as `(base embedding).method`. Where available, we use published debiased embeddings made available from the original authors of the corresponding method. Otherwise, we apply the method to the base GloVe embeddings. The methods we will analyze include:

Hard Debias[3] **(GloVe**[4]**.HD)** [7]: The subset of gender-neutral words are pro-

[3] https://github.com/tolga-b/debiaswe

[4] The base (undebiased) embeddings are GloVe trained on the 2017 January Wikipedia dump (vocab contains 322,636 tokens). Available at https://github.com/uclanlp/gn_glove.

jected onto the nullspace of the gender direction $\vec{g}$. Gender-neutral words are made equidistant to pairs of words in a defined equalization set.

Gender-Neutral GloVe[5] **(GN-GloVe)** [82]: Similar to hard debias, this method seeks to eliminate the direct bias. The embeddings are retrained from scratch using a modified version of GloVe's original objective function. The gender information is sequestered to the final component of the word embedding. The gender-neutral portion of the word embedding is then defined as the first $d - 1 = 299$ components, denoted **GN-GloVe(w_a)**.

Gender-Preserving[6] **(GloVe**[4]**.GP)** [35]: This method seeks to eliminate harmful gender bias while retaining as much useful semantic gender information as possible.

Double Hard Debias[7] **(GloVe**[4]**.DHD)** [75]: An extended version of the hard debias algorithm, based on the observation that frequency information encoded in the word embeddings convolutes the definition of the gender direction. Correctional pre-processing is applied prior to hard debiasing.

Bias Alignment Model[8] **(GloVe**[4]**.BAM)** [39]: Gender subspace matrices are defined by stacking explicit gender words. The projection that maps the embedding space to itself while approximately aligning the gender subspaces is learned and applied to all words. After alignment, gender information is not retained.

Orthogonal Subspace Correction and Rectification[9] **(GloVe**[4]**.OSCaR)** [22]: The rationale is that linear projective methods are too aggressive in modifying the entire embedding space. OSCaR rectifies two concepts of interest (gender and occupations), such that these subspaces are orthogonal in the debiased space.

Iterative Nullspace Linear Projection[10] **(GloVe**[4]**.INLP)** [61]: Rather than

[5] https://github.com/uclanlp/gn_glove

[6] https://github.com/kanekomasahiro/gp_debias

[7] https://github.com/uvavision/Double-Hard-Debias

[8] https://github.com/anlausch/DEBIE

[9] https://github.com/sunipa/OSCaR-Orthogonal-Subspace-Correction-and-Rectification

[10] https://github.com/shauli-ravfogel/nullspace_projection. The projection matrix computed for our base GloVe embeddings is available at https://github.com/hillary-dawkins/MAB.

defining a gender direction, INLP *learns* the most informative decision boundary for classifying gendered and gender-neutral words. All words are projected to the nullspace of the gender subspace, and the process proceeds iteratively until gender information is sufficiently erased. A closely related method is the D_4 algorithm [17].

Repulse Attract Neutralize Debias[11] (GloVe[4].RAN) [37]: Motivated by the persistence of implicit bias after debiasing through projective methods (observed as clustering and recoverability), RAN-debias attempts to address both direct bias and gender-based proximity bias.

3.3.2 Explicit gender words test set and error definitions

Firstly, we construct a test set where every sentence pair is of the form [A person *verb object*] → [(A) **gender word** *verb object*] (the correct inference is always neutral since a person can be of any gender). Verbs ($n = 27$) and objects ($n = 184$) are paired to create $n = 1968$ unique premise sentences[12]. Gender words are taken to be {man, woman, guy, girl, gentleman, lady, He, She}, following [21] with the addition of the pronouns, for a total test set S of $|S| = 15744$ sentence pairs where hypotheses represent binary genders evenly (denoted S_M, S_F, $|S_M| = |S_F|$).

For every hypothesis sentence in the test set, the ideal prediction probability vector is $(N, E, C) = (1, 0, 0)$. We could define the error on the test set as the average Euclidean distance from the ideal distribution:

$$\mathcal{E} = \frac{1}{|S|} \sum_{i \in S} \|(1, 0, 0) - (N, E, C)_i\|_2. \tag{3.1}$$

This task, test set, and error definition are simple, and yet they encapsulate the central challenge of the debiasing field: to create attribute-aware (required to obtain the Neutral prediction) but attribute-unbiased embeddings.

[11] https://github.com/TimeTraveller-San/RAN-Debias

[12] Verbs and objects are taken from [21] word lists (https://github.com/sunipa/On-Measuring-and-Mitigating-Biased-Inferences-of-Word-Embeddings) and are paired using the same pairing rules.

A weaker, but still potentially desirable, condition might be to minimize the effect of gender while not requiring that the model be gender-aware. Typically, this means that all hypotheses tend towards an Entail prediction, regardless of gender. We could define the error as the average distance between probability vectors between genders:

$$d = \frac{1}{2|S|} \left\| \sum_{i \in S_M} (N, E, C)_i - \sum_{j \in S_F} (N, E, C)_j \right\|_2 . \tag{3.2}$$

A gender-agnostic model could achieve zero error by this definition even with an accuracy of zero on the test set.

Table 3.1 shows the results for this test set on all the embeddings of interest. None of the debiased embeddings successfully mitigate the marked attribute error.

3.3.3 Latent gender carriers

Next, we would like to check if the gender-induced marked attribute bias can affect entities which should be gender neutral, but turn out to be hidden carriers of a gender attribute (e.g. stereotypical occupations and names). The same template [A person *verb object*] → [(A/An **occupation**)/**Name** *verb object*] was used with the common (verb, object) vocabulary set. Stereotypical occupations ($n = 32$) were sourced from [7], and the SemBias test set. Examples are (doctor, engineer, boss, etc. vs. nurse, maid, homemaker, etc.). Names ($n = 64$) are sourced from the most common names of the previous decade in the US, according to the Social Security Administration[13]. In total there are 62,976 sentence pairs in the Occupation test set, and 125,952 sentence pairs in the Names test set[14].

Results are shown in Tables 3.2 and 3.3. A permutation test is used to check if dividing the occupations into groups according to gender stereotypes produces a significant difference in the probability vectors (rather than dividing them randomly).

[13] https://www.ssa.gov/oact/babynames/

[14] The exact word set used to produce these results is available at https://github.com/hillary-dawkins/MAB.

Table 3.1: Results of the marked attribute test set on **explicit gender words**. Due to varying results on gender nouns vs. pronouns, results are shown separately for each case (M and F represent averages across the gender nouns). Some debiased embeddings are able to eliminate the distance across pronouns (really by definition since $\vec{she} \approx \vec{he}$ in these cases), but none are able to eliminate differences between the gender nouns significantly. Even when differences between genders are minimized, distance from the ideal distribution (error $\mathcal{E}$) remains or increases. This highlights the challenge of creating gender-aware but not gender-biased embeddings.

Emb.method	Gender	N	E	C	Gender	N	E	C	d	$\mathcal{E}$
GloVe	M	0.7832	0.1966	0.0202	F	0.9449	0.0401	0.0149	0.225	0.182
	he	0.0982	0.8838	0.0180	she	0.6549	0.3137	0.0315	0.797	0.865
GloVe.HD	M	0.8306	0.1329	0.0365	F	0.9269	0.0499	0.0232	0.128	0.155
	he	0.2944	0.6737	0.0319	she	0.5174	0.4334	0.0491	0.328	0.813
GN-GloVe	M	0.6339	0.3402	0.0259	F	0.9169	0.0461	0.0370	0.408	0.301
	he	0.1767	0.7968	0.0265	she	0.8223	0.1405	0.0373	0.921	0.688
GN-GloVe(w_a)	M	0.8446	0.1254	0.0300	F	0.9211	0.0395	0.0394	**0.115**	**0.149**
	he	0.1430	0.8266	0.0304	she	0.4237	0.5367	0.0396	0.404	0.990
GloVe.DHD	M	0.7013	0.2685	0.0302	F	0.9282	0.0510	0.0209	0.315	0.247
	he	0.1566	0.8187	0.0247	she	0.1597	0.8139	0.0264	0.006	1.173
GloVe.GP	M	0.6172	0.3521	0.0306	F	0.8777	0.0693	0.0530	0.385	0.336
	he	0.2443	0.7262	0.0295	she	0.6481	0.3040	0.0480	0.585	0.758
GloVe.BAM	M	0.7983	0.1703	0.0314	F	0.9329	0.0447	0.0224	0.184	0.175
	he	0.1625	0.8083	0.0292	she	0.6752	0.2878	0.0369	0.731	0.800
GloVe.OSCaR	M	0.8233	0.1572	0.0195	F	0.9431	0.0400	0.0169	0.168	0.154
	he	0.1482	0.8292	0.0226	she	0.8428	0.1278	0.0294	0.987	0.697
GloVe.RAN	M	0.8055	0.1686	0.0260	F	0.8994	0.0701	0.0305	0.136	0.193
	he	0.1939	0.7811	0.0250	she	0.5962	0.3420	0.0618	0.597	0.828
GloVe.INLP	M	0.8298	0.1537	0.0166	F	0.9204	0.0633	0.0164	0.128	0.167
	he	0.1081	0.8710	0.0209	she	0.1119	0.8672	0.0209	**0.005**	1.244

Table 3.2: Results of marked attribute test set on **stereotypical occupations**. Each (N,E,C) probability vector is averaged over the 1968 unique premise sentences and the gender attribute words from each category (M or F) ($n = 31,488$ sentences for each gender). Smaller distances between the M and F vectors indicate less gender bias. The significance of the difference was evaluated using a permutation test; the alternate distance d^* is computed for 10,000 randomly sampled partitions of the occupations into two groups. The significance value is the proportion of these samples to generate a distance $d^* > d$. This gives us an idea of whether the defined partition, based on gender, is a meaningful grouping. Smaller significance values indicate that the defined partition is non-random with respect to the distance.

Emb.method	M attribute (N, E, C)	F attribute (N, E, C)	Distance d	Significance
GloVe	(0.6000, 0.3350, 0.0650)	(0.7378, 0.1711, 0.0910)	0.216	0.0001
GloVe.HD	(0.4975, 0.4500, 0.0525)	(0.6075, 0.3357, 0.0568)	0.159	0.0408
GN-GloVe	(0.5026, 0.4434, 0.0540)	(0.7126, 0.2036, 0.0838)	0.320	0.0000
GN-GloVe(w_a)	(0.5309, 0.3915, 0.0776)	(0.6197, 0.2771, 0.1032)	0.147	0.0478
GloVe.DHD	(0.5285, 0.4126, 0.0589)	(0.6513, 0.2811, 0.0676)	0.180	0.0038
GloVe.GP	(0.5016, 0.4380, 0.0604)	(0.6479, 0.2639, 0.0882)	0.229	0.0010
GloVe.BAM	(0.6293, 0.3077, 0.0630)	(0.7116, 0.1972, 0.0912)	0.141	0.0060
GloVe.OSCaR	(0.5577, 0.3901, 0.0522)	(0.6789, 0.2400, 0.0812)	0.195	0.0036
GloVe.RAN	(0.5393, 0.3933, 0.0674)	(0.5924, 0.3026, 0.1050)	0.112	0.0477
GloVe.INLP	(0.5065, 0.4197, 0.0739)	(0.5465, 0.3949, 0.0587)	0.050	0.6595

Table 3.3: Results of marked attribute test set on **names**. Each (N,E,C) probability vector is averaged over the 1968 unique premise sentences and the gender attribute words from each category (M or F) ($n = 62,976$ sentences for each gender). Smaller distances between the M and F vectors indicate less gender bias. As with the Occupations latent gender bias test set results, significance is evaluated using a partition test.

Emb.method	M attribute (N, E, C)	F attribute (N, E, C)	Distance d	Significance
GloVe	(0.4657, 0.4766, 0.0577)	(0.7283, 0.1598, 0.1120)	0.415	0.0000
GloVe.HD	(0.5745, 0.3547, 0.0708)	(0.6685, 0.2760, 0.0555)	0.124	0.0140
GN-GloVe	(0.4619, 0.4713, 0.0668)	(0.7209, 0.1906, 0.0885)	0.383	0.0000
GN-GloVe(w_a)	(0.5882, 0.2878, 0.1240)	(0.6662, 0.2321, 0.1017)	0.098	0.0241
GloVe.DHD	(0.4731, 0.4464, 0.0805)	(0.5690, 0.3529, 0.0780)	0.134	0.0192
GloVe.GP	(0.5488, 0.3761, 0.0751)	(0.7470, 0.1677, 0.0853)	0.288	0.0000
GloVe.BAM	(0.5941, 0.3424, 0.0635)	(0.7698, 0.1628, 0.0674)	0.251	0.0000
GloVe.OSCaR	(0.6012, 0.3149, 0.0839)	(0.7191, 0.2020, 0.0789)	0.163	0.0001
GloVe.RAN	(0.5295, 0.3865, 0.0839)	(0.6920, 0.2151, 0.0929)	0.236	0.0000
GloVe.INLP	(0.5091, 0.4042, 0.0867)	(0.5447, 0.3639, 0.0914)	0.054	0.4049

As shown, the marked attribute effect persists on stereotypical occupations, especially on original embeddings. This is an important result because it highlights that unintended behaviour can appear in unexpected places due to a latent attribute. Previously, GBETs have focused on how explicit gender words are treated under biased models. To our knowledge, this is the first GBET designed to analyze unintended behaviour on a latent attribute carrier.

Note that this task is easier to correct than the explicit gender words because occupation words have defining characteristics beyond gender. That is, a debiasing method such as Iterative Nullspace Projection can perform well by removing gender information entirely. This does not mean that the challenge of having a gender-aware but gender-unbiased embedding is solved, but it does provide evidence that latent gender effects can be mitigated using linear projective methods. The full extent of latent biased-attribute effects and possible mitigation strategies should be investigated further.

3.4 Intrinsic bias measures

How to define bias on an embedding space remains an active area of study. In general, we seek to understand how the intrinsic or geometric properties of an embedding space translate to real observable bias in downstream tasks. Intrinsic properties are easy to compute quickly, whereas computing performance on downstream tasks requires us to train new models for every case. Understanding of the correlations between the two gives insight on how word embeddings should be debiased.

As a case study, let us focus on the marked attribute error $\mathcal{E}$ on the explicit gender words (shown in Table 3.1). Recall that this measure of bias is of interest because zero error corresponds to the gold standard: having an attribute-aware model, while simultaneously not using the gender attribute to make inappropriate inferences. In this section, we look at 5 existing intrinsic bias measures: Direct Bias, Clustering,

Recoverability, Gender-based Illicit Proximity Estimate (GIPE), and SemBias. We will investigate whether any of these measures are predictive of the marked attribute effect.

Recall that direct bias was the first measure to be proposed; it simply measures the average projection of word vectors onto a predefined gender direction. Early methods (i.e. Hard Debias and GN-GloVe) defined bias in the embedding space completely as direct bias. The idea of clustering and recoverability refer to a classifier's ability to correctly reassign gender labels to words, even after debiasing methods have been applied. Gonen and Goldberg's [29] observation of clustering and recoverability sparked new interest in defining metrics for indirect bias on the embedding space. Although clustering and recoverability do not provide well-defined measures of bias given an embedding space (as they depend training implementation — though they could be said to provide a lower bound), many new debiasing proposals will cite reduced clustering as a positive result. The effect on downstream applications is not well understood as of yet. The Gender-based Illicit Proximity Estimate (GIPE) measures the extent of undue proximities in the embedding space due to a pervasive gender attribute. Lastly, the SemBias analogy test set measures whether gender-biased analogies exist within the embedding space based on vector arithmetic properties.

Implementation details for each measure as well as the experimental set of embeddings ($n = 16$) are given in Appendix B.1. The average Direct Bias on the embedding space was found to have a Pearson correlation coefficient of 0.104 with the marked attribute error. The Clustering v-measure[15] [64] achieved a correlation coefficient of 0.184. Recoverability was attempted using an SVM with a linear decision boundary, an SVM with a non-linear (radial basis function) kernel, logistic regression, and a simple 1-hidden-layer fully-connected network. All recoverability correlation results were comparable, but the best coefficient of 0.223 was achieved by logistic regression. The

[15]With cluster size $n = 1500$ (which lead to the highest observed correlation); see appendix.

GIPE[16] had a correlation coefficient of 0.432. The SemBias[17] test set had a correlation coefficient of 0.091. The full correlation matrix between all intrinsic bias measures can be found in Appendix B.1. The results suggest that the marked attribute effect is not well correlated with any present notion of intrinsic bias, therefore we do not have a good understanding of how the word embedding properties contribute to this type of observable bias.

In seeking a potential solution, we develop a new intrinsic bias measure, multi-dimensional information-weighted direct bias (MIDB), found to have a more meaningful correlation of 0.667 with the marked attribute error. We define the MIDB of a particular word $\vec{x}$ to be a weighted average over inner products with basis vectors of a multi-dimensional gender subspace:

$$\text{MIDB}_d(\vec{x}) = \sum_{i=1}^{d} a_i \langle \vec{g_i}, \vec{x} \rangle \tag{3.3}$$

where $\{\vec{g_i}\}$ form an orthonormal basis for the gender subspace, here defined as the first d principal components summarizing difference vectors $\{\vec{\delta}_{jk}\}$. The difference vectors are taken as all pairwise differences[18] between vectors in defined gender sets (here common names were used[13]): $\{\vec{\delta}_{jk}\} = \vec{f_j} - \vec{m}_k,\ f_j \in F_{\text{names}},\ m_k \in M_{\text{names}}$ ($|M_{\text{names}}| = |F_{\text{names}}| = 100$). The weighting a_i is the proportion of variance explained by the i^{th} principal component, and d is a hyperparameter controlling the number of dimensions to keep[19].

New proposals for defining a gender direction or subspace potentially have far reaching consequences in the landscape of intrinsic bias measures and their related debiasing schemes. In fact all of Clustering, Recoverability, GIPE, and SemBias use the

[16] Using an indirect bias threshold of $\theta = 0.05$, and number of nearest neighbours $n = 100$.

[17] The SemBias score was taken as the proportion of analogy examples in the test set for which the embedding space returns the correct definitional analogy.

[18] Using all pairwise differences creates a matrix with rank much less than the dimension of the matrix, however the rank is still much larger than d (the number of principal components to extract) so it doesn't cause a problem.

[19] On our set of experimental embeddings, $d = 4$ was empirically found to produce the 0.7 correlation result.

classic uni-dimensional gender direction $\vec{g}$ within their definitions. The weak observed correlation between DB and MIDB suggests that these subspaces are independent. Swapping in a uniquely informative gender subspace to the existing indirect measures would produce a new family of intrinsic bias measures. The observed utility of names in defining a meaningful gender subspace is encouraging because it opens an obvious avenue for this method to be applied to attributes of interest beyond gender (e.g. race or ethnicity).

3.5 Multi-dimensional information-weighted soft projection

In this section we motivate the above search for an informative intrinsic bias measure. As discussed, a greater understanding of how embedding properties influence observed bias can inform new debiasing techniques. Translating the idea of MIDB into a debiasing scheme yields Multi-dimensional Information-weighted Soft Projection (MISP).

In this debiasing procedure, we project all words into the nullspace of the multi-dimensional gender subspace, proportional to our belief that certain dimensions actually encode the latent idea of gender:

$$\vec{w}_{deb} = \vec{w} - \sum_{i=1}^{d} a_i \langle \vec{g}_i, \vec{w} \rangle \vec{g}_i \tag{3.4}$$

where $\vec{w}$ is the input embedding, $\vec{w}_{deb}$ is the debiased output embedding, and all other quantities are defined as in eqn. 3.3.

As shown in Table 3.1, the GN-GloVe(w_a) embeddings are currently the top performers on the explicit gender words test set, as measured by either error $\mathcal{E} = 0.149$, or distance $d = 0.115$. Applying MISP to GN-GloVe(w_a) embeddings (denoted GN-GloVe(w_a).MISP), we achieve an error on the explicit gender words test set of $\mathcal{E} = 0.1107$, a 26% error reduction over the previous best. The distance d between genders is reduced to $d = 0.08744$, a 21% reduction over the previous best. Successful

concatenation suggests that this technique is distinct, and independently useful, from techniques that seek to minimize the traditional direct bias (including GN-GloVe). This observation is consistent with the weak observed correlation between direct bias and MIDB_4 on the experimental set of embeddings.

Computing the intrinsic bias measures Clustering, Recoverability, GIPE and SemBias on the newly created embedding space GN-GloVe(w_a).MISP (compared to the base GN-GloVe(w_a)), we observe a clustering v-score of 0.498 (previously 0.497)[20], a recoverability accuracy of 0.992 (previously 0.993)[21], a GIPE of 0.1169 (previously 0.1173)[22], and a SemBias score of 0.938 (previously 0.938)[23]. The MISP method did not reduce bias by any of these measures, although this is not particularly surprising as it was designed to address the marked attribute effect (through MIDB). It is encouraging however that none of these bias measures were increased. In other words, there is no expected trade-off between the reduced marked attribute error and any previous debiasing work that relied on these measures. The SemBias result informs us that MISP did not reintroduce any harmful biased analogies, for example.

For reference, if we apply the analogous multi-dimensional hard debias method (i.e. equation 3.4 where all weights a_i are set to 1), the output embeddings GN-Glove(w_a).MHD do not successfully mitigate the marked attribute effect ($\mathcal{E} = 0.1501$, $d = 0.1603$). This suggests that the soft nature of the projection is a key ingredient.

Furthermore, we provide some evidence that specifically the information weighting of the soft projection is a good ingredient as follows. Recall that we are attenuating components of each basis vector according to our belief in that vector as a good gender direction. The basis vectors are defined to be the first d principal components, weighted by their corresponding variance explained. Therefore the first basis vector

[20] Where clustering size $n = 1500$.

[21] This is the highest accuracy achieved by any of the four classification methods tested; implementation details are in Appendix B.1.

[22] Computed with indirect bias threshold $\theta = 0.03$, and number of nearest neighbours $n = 100$.

[23] Reported as the proportion of samples in the full test set to return the definitional analogy; higher scores are better.

receives the greatest weight and so on. To test the significance of this decision, we define alternative debiased embeddings by applying MISP where the weights get reassigned to the "wrong" vector (for $d = 4$, we have 23 alternative pairings). We observe that **none** of the 23 alternatives obtain an error $\mathcal{E}$ less than the "true" implementation of MISP. This suggests that weighting the components by order of information is a good ingredient. Values of $\mathcal{E}$ for the alternate embeddings can be found in the appendix. Model parameters for each case are made available in order to reproduce this argument on any extended version of the MAB test set.

Information weighting is an interesting idea because it could be applied to either defined or learned gender subspaces alike. For instance, if the basis vectors of a gender subspace are taken as the iteratively learned linear decision boundaries (as in INLP), we could investigate weighting each dimension by the accuracy acc_i of classification on each iteration, as $a_i = (1-2acc_i)$. In this way, dimensions receive weights proportional to their ability to predict gender information. When accuracy reaches 0.5, no gender information remains, the learned decision boundary is meaningless, and the basis vector receives zero weighting.

Finally, as with any debiasing method, we wish to verify that application of the method has not damaged the overall embedding quality. We assess the MISP embeddings on a handful of classic analogy and word semantic similarity benchmarks. The word similarity benchmarks measure how closely the word embeddings capture similarity between words compared to human annotation. We use the following datasets: RG [65], MTurk [79], MEN [11], and SimLex999 [32]. The word analogy task measures how well the word embeddings capture semantic and syntactic relationships among words as vector properties. We report on the Google [46], and MSR [48] test sets. Results were obtained following the word embedding benchmark package[24] [34]. As shown in Table 3.4, application of MISP does not alter the overall word embedding quality.

[24]https://github.com/kudkudak/word-embeddings-benchmarks

Table 3.4: Results for word similarity and analogy benchmarks. Results on the word analogy tasks are reported as percentage accuracy. Results on the word similarity tasks are reported as a Spearman correlation ($\times$ 100). Application of MISP does not alter the overall quality of word embeddings as measured by these classic test sets.

Embedding.method	Sem	Syn	Google-Total	MSR	RG	MTurk	MEN	SL999
GloVe	80.48	62.76	70.80	51.49	75.29	64.27	72.19	34.86
GloVe.MISP	80.49	62.81	70.84	51.51	76.06	64.32	72.41	35.04
GN-GloVe	77.62	61.60	68.87	49.29	74.11	66.36	74.49	37.12
GN-GloVe(w_a)	77.68	61.56	68.87	49.38	75.46	66.55	74.72	37.53
GN-GloVe(w_a).MISP	77.68	61.59	68.89	49.26	75.49	66.45	74.76	37.60

3.6 Conclusion

This paper highlights a new observation of gender bias in a downstream setting: marked attribute bias in natural language inference. The current inference is that "person" implies male, while "person" does not imply female. Consequently, this inference is being baked into our models of natural language understanding. The effect was shown to persist on explicitly defined gender words and on latent gender-attribute carriers. Based on an assessment of the current debiasing landscape, none of the current debiasing methods satisfactorily mitigate the marked attribute error, and furthermore none of the intrinsic bias measures are useful at predicting the marked attribute effect.

By noticing a more meaningful correlation with a newly identified intrinsic bias measure, we propose a new debiasing scheme: multi-dimensional information-weighted soft projection (MISP). This method introduces several concepts, including the use of a multi-dimensional defined gender subspace. Previously, the concept of a defined gender subspace always appeared as a single dimension. The iterative nullspace projection method implicitly uses higher learned dimensions, however this requires

learning a new decision boundary at every iteration, subject to the implementation of a training procedure. Furthermore, the learned dimensions were not used to define any bias metric, they were strictly used operationally for the debiasing procedure. MISP also introduces the idea of a soft or partial projection, where weights are informed by some measure of the dimension's ability to capture the intended latent concept of a gender direction. Both of these ideas could be further explored and extended to create new notions of indirect bias, which in turn could inform more sophisticated debiasing procedures.

Multi-dimensional information-weighted soft projection applied to GN-GloVe(w_a) produces new debiased embeddings that achieve the lowest error on the marked attribute bias test set, a 26% reduction over the previous best, and a 45% reduction over the original undebiased embeddings. Error reduction on this test set is thought to encapsulate the overall goal of producing gender-aware but gender-unbiased embeddings. Therefore, this method and its composite ingredients warrant further investigation. Each of the marked attribute bias test sets are made available for further exploration and iteration on these ideas.

Chapter 4

Second Order WinoBias (SoWinoBias) Test Set for Latent Gender Bias Detection in Coreference Resolution

In the previous chapter, we learned that specifically defining an intrinsic bias measure which is shown to correlate with an observed bias of interest can be a successful pathway for developing new debiased word embeddings. Specifically, we observed that reducing a variant of direct bias in the embedding space can help us to mitigate the marked attribute effect in a downstream setting. Furthermore, we observed that indirect bias mitigation (as measured by any of the intrinsic measures) is not correlated with the marked attribute effect.

This observation raises a question around the utility of indirect bias mitigation more broadly. Recall that following the clustering and recoverability revelation [29], focus within the debiasing community shifted towards indirect bias mitigation. New proposals during this era implied that reducing the presence of clustering and recoverability was in itself a valuable achievement, but the connection to a downstream task was often lacking.

In this chapter, we design a test set for capturing latent gender-biased associations in the downstream setting. We then analyze the effectiveness of various debiasing methods in mitigating the observed latent gender bias, which yields an interesting discussion around the necessary and sufficient conditions on the embedding space for such mitigation. The insights gained from this analysis will influence the

design decisions for the debiasing scheme applied to a pre-trained language model (Chapter 5).

This chapter is based on the publication

H. Dawkins. *Second order WinoBias (SoWinoBias) test set for latent gender bias detection in coreference resolution.* In Proceedings of the 3rd Workshop on Gender Bias in Natural Language Processing, pages 103–111, Online, Aug. 2021. Association for Computational Linguistics. [19],

presented here with very minor notational edits. We thank Dr. Daniel Gillis, Dr. Judi McCuaig, Dr. Stefan C. Kremer, Dr. Graham Taylor, and anonymous reviewers for their time and thoughtful discussion on this work. A tangible contribution from this publication is the SoWinoBias test set, which is made publicly available, and can be used for gender bias detection.

ABSTRACT

We observe an instance of gender-induced bias in a downstream application, despite the absence of explicit gender words in the test cases. We provide a test set, SoWinoBias, for the purpose of measuring such latent gender bias in coreference resolution systems. We evaluate the performance of current debiasing methods on the SoWinoBias test set, especially in reference to the method's design and altered embedding space properties.

4.1 Introduction

Explicit (or first-order) gender bias was observed in coreference resolution systems by [81], by considering contrasting cases:

1. **The doctor** hired the secretary because **he** was overwhelmed.

2. **The doctor** hired the secretary because **she** was overwhelmed.

3. The doctor hired **the secretary** because **she** was highly qualified.

4. The doctor hired **the secretary** because **he** was highly qualified.

Sentences 1 and 3 are pro-stereotypical examples because gender words align with a socially-held stereotype regarding the occupations. Sentences 2 and 4 are anti-stereotypical because the correct coreference resolution contradicts a stereotype. It was observed that systems performed better on pro cases than anti cases, and the WinoBias test set was developed to quantify this disparity.

Here we make a new observation of gender-induced (or second-order) bias in coreference resolution systems, and provide the corresponding test set SoWinoBias. Consider cases:

1. The doctor liked **the nurse** because **they** were beautiful.

2. **The nurse** dazzled the doctor because **they** were beautiful.

3. The nurse admired **the doctor** because **they** were beautiful.

The examples do not contain any explicit gender cues at all, and yet we can observe that sentences 1 and 2 align with a gender-induced social stereotype, while

sentence 3 opposes the stereotype. The induction occurs because "nurse" is a female-coded occupation [7, 82], and women are also more likely to be described based on physical appearance [33, 78]. A coreference resolution system is gender-biased if correct predictions on sentences like 1 and 2 are more likely than on sentence 3.

The difference between first-order and second-order gender bias in a downstream application is especially interesting given current trends in debiasing static word embeddings. Early methods [7, 82] focused on eliminating direct bias from the embedding space, quantified as associations between gender-neutral words and an explicit gender vocabulary. In response to an influential critique paper by [29], the current trend is to focus on eliminating indirect bias from the embedding space, quantified either by gender-induced proximity among embeddings [37] or by residual gender cues that could be learned by a classifier [61, 17].

Indirect bias in the embedding space was viewed as an undesirable property a priori, but we do not yet have a good understanding of the effect on downstream applications. Here we test debiasing methods from both camps on SoWinoBias, and make a series of observations on sufficient and necessary conditions for mitigating the latent gender-biased coreference resolution.

Additionally, we consider the case that our coreference resolution model employs both static and contextual word embeddings, but debiasing methods are applied to the static word embeddings only. Post-processing debiasing techniques applied to static word embeddings are computationally inexpensive, easy to concatenate, and have a longer development history. However contemporary models for downstream applications are likely to use some form of contextual embeddings as well. Therefore we might wonder whether previous work in debiasing static word embeddings remains relevant in this setting. The WinoBias test set for instance was developed and tested using the "end-to-end" coreference resolution model [40], a state-of-the-art model at that time using only static word embeddings. Subsequent debiasing schemes reported results on WinoBias using the same model, just plugging in different debiased embed-

dings, for the sake of fair comparison. However this is becoming increasingly outdated given the progress in coreference resolution systems. A contribution of this work is to report WinoBias results for previous debiasing techniques using a contemporary model, one that makes use of unaltered contextual embeddings in addition to the debiased static embeddings.

The remainder of the paper is organized as follows: In section 4.2, we further define the type of bias being measured by the SoWinoBias test set and discuss some limitations. In section 4.3, we review the 4 word embedding debiasing methods that we will analyze, in the context of how each method aims to alter the word embedding space. In section 4.4, we provide details of the experimental setup and report results on both coreference resolution test sets, the original WinoBias and the newly constructed SoWinoBias. In section 4.5, we discuss the results with respect to the geometric properties of the altered embedding spaces. In particular, we review whether mitigation of intrinsic measures of bias on the embedding space, quantified as direct bias and indirect bias by various definitions, are related to mitigation of the latent bias in a downstream application.

4.2 Bias Statement

Within the scope of this paper, bias is defined and quantified as the difference in performance of a coreference resolution system on test cases aligning with a socially-held stereotype vs. test cases opposing a socially-held stereotype. We observe that gender-biased systems perform significantly better in pro-stereotypical situations. Such difference in performance creates representational harm by implying (for example) that occupations typically associated with one gender cannot have attributes typically associated with another.

Throughout this paper, the term "second-order" is used interchangeably with "latent". Characterizing the observed bias as "second-order" follows from the ob-

servation of a gender-induced bias in the absence of gender-definitional vocabulary, resting on the definition of "they" as a gender-neutral pronoun.

Therefore, a limitation in the test set construction is the possible semantic overloading of "they". As discussed, the intention throughout this paper is to use the singular "they" as a pronoun that does not carry any gender information (and could refer to someone of any gender). However, different contexts may choose to treat "they" exclusively as a non-binary gender pronoun.

The gender stereotypes used throughout this paper are sourced from peer-reviewed academic journals written in English, which draw from the US Labor Force Statistics, as well as US-based crowd workers. Therefore a limitation may be that stereotypes used here are not common to all languages or cultures.

4.3 Debiasing methods

4.3.1 Neutralization of static word embeddings

Methods addressing direct bias

The first attempts to debias word embeddings focused on the mitigation of direct bias [7]. The definition of direct bias assumes the presence of a "gender direction" $\vec{g}$; a subspace that mostly encodes the difference between the binary genders. A non-zero projection of word $\vec{w}$ onto $\vec{g}$ implies that $\vec{w}$ is more similar to one gender over another. In the case of ideally gender-neutral words, this is an undesirable property. Direct bias quantifies the extent of this uneven similarity[1],

$$DB(N) = \frac{1}{|N|} \sum_{\vec{w} \in N} |\cos(\vec{w}, \vec{g})| \tag{4.1}$$

where $\cos(\vec{w}, \vec{g})$ denotes the cosine of the angle between vectors $\vec{w}$ and $\vec{g}$. The Hard Debias method [7] is a post-processing technique that projects all gender-neutral

[1]The original definition included a strictness exponent c, here set to 1 as has commonly been done in subsequent works.

words into the nullspace of $\vec{g}$. Therefore, the direct bias is made to be zero by definition. We measure the performance of **Hard-GloVe**[2] on the coreference resolution tasks.

A related retraining method used a modified version of GloVe's original objective function with additional incentives to reduce the direct bias for gender-neutral words, resulting in the GN-GloVe embeddings [82]. Rather than allowing for gender information to be distributed across the entire embedding space, the method explicitly sequesters the protected gender attribute to the final component. Therefore the first $d-1$ components are taken as the gender-neutral embeddings, denoted **GN-GloVe**(w_a)[3].

Methods addressing indirect bias

The indirect bias is less well defined, and loosely refers to the gender-induced similarity measure between gender-neutral words. For instance, semantically unrelated words such as "sweetheart" and "nurse" may appear quantitatively similar due to a shared gender association.

One definition (first given in [7]) measures the relative change in similarity after removing direct gender associations as

$$\beta(\vec{w},\vec{v}) = \frac{1}{\vec{w}\cdot\vec{v}}\left(\vec{w}\cdot\vec{v} - \frac{\vec{w}_\perp\cdot\vec{v}_\perp}{\|\vec{w}_\perp\|\|\vec{v}_\perp\|}\right), \tag{4.2}$$

where $\vec{w}_\perp = \vec{w} - (\vec{w}\cdot\vec{g})\vec{g}$ however this relies on a limited definition of the original gender association.

The Repulse-Attract-Neutralize (RAN) debiasing method attempts to repel undue gender proximities among gender-neutral words, while keeping word embeddings close to their original learned representations [37]. This method quantifies indirect

[2]Hard debias: https://github.com/tolga-b/debiaswe. All base (undebiased) embeddings are GloVe trained on the 2017 January Wikipedia dump (vocab contains 322,636 tokens). Available at https://github.com/uclanlp/gnglove, based on the work of [56].

[3]https://github.com/uclanlp/gnglove

bias by incorporating β into a graph-weighted holistic view of the embedding space (more on this later). In this paper, we will measure the performance of **RAN-GloVe**[4] on the coreference resolution tasks.

A related notion of indirect bias is to measure whether gender associations can be predicted from the word representation. The Iterative Nullspace Linear Projection method (INLP) achieves linear guarding of the gender attribute by iteratively learning the most informative gender subspace for a classification task, and projecting all words to the orthogonal nullspace [61]. After sufficient iteration, gender information cannot be recovered by a linear classifier. We will measure the performance of **INLP-GloVe**[5].

4.3.2 Data augmentation

In addition to debiasing methods applied to word embeddings, we measure the effect of simple data augmentation applied to the training data for our coreference resolution system. The goal is to determine whether data augmentation can complement the debiased word embeddings on this particular test set. The training data is augmented using a simple gender-swapping protocol, such that binary gender words are replaced by their equivalent form of the opposite gender (e.g. "he" $\leftrightarrow$ "she", etc.).

4.4 Detection of gender bias in coreference resolution: Experimental setup

All systems were built using the "Higher-order coreference resolution with coarse-to-fine inference" model [41][6]. It is important to keep in mind that this model uses both static word embeddings and contextual word embeddings (specifically ELMo

[4]https://github.com/TimeTraveller-San/RAN-Debias

[5]https://github.com/shauli-ravfogel/nullspaceprojection

[6]https://github.com/kentonl/e2e-coref

embeddings [58]). Our experimental debiasing methods were applied to static word embeddings only, and contextual embeddings are left unaltered in all cases.

All systems were trained using a large public dataset for coreference resolution, OntoNotes 5.0[7], using the default hyperparameters[8], for approximately 350,000 steps until convergence. Baseline performance was tested using the OntoNotes 5.0 test set (results shown in Table 4.1). Baseline performance is largely consistent across all models, indicating that neither debiased word embeddings nor gender-swapped training data significantly degrades the performance of the system overall.

4.4.1 WinoBias

The WinoBias test set was created by [81], and measures the performance of coreference systems on test cases containing explicit binary gender words. In particular, pro-stereotypical sentences contain coreferents where an explicit gender word (e.g. he, she) is paired with an occupation matching a socially held gender stereotype. Anti-stereotypical sentences use the same formulation but gender swap the explicit gender words such that coreferents now oppose a socially held gender stereotype. Gender bias is measured as the difference in performance on the pro versus anti test sets, each containing $n = 396$ sentences.

Recall that here we are reporting WinoBias results using a system incorporating unaltered contextual embeddings, in addition to the debiased static embeddings. Previously reported results on the "end-to-end" coreference model [40], using only debiased static word embeddings, are compiled in Table 4.2 for reference.

In our new setting, we observe that debiasing methods addressing direct bias are more successful than those addressing indirect bias. In particular, without the additional resource of data augmentation, RAN-GloVe struggles to reduce the difference between pro and anti test sets (in contrast to RAN-GloVe's great success in

[7] https://catalog.ldc.upenn.edu/LDC2013T19

[8] "best" configuration at https://github.com/kentonl/e2e-coref/blob/master/experiments.conf

Table 4.1: Results on coreference resolution test sets. OntoNotes (F_1) performance provides a baseline for "vanilla" coreference resolution ($n = 348$). WinoBias (F_1) measures explicit gender bias, observable as the diff. between pro ($n = 396$) and anti ($n = 396$) test sets. SoWinoBias (% accuracy) measures second-order gender bias, likewise observable as the diff. between pro ($n = 4096$) and anti ($n = 4096$) test sets. Note: accuracy is the relevant metric to report on the SoWinoBias test set, rather than F_1, due to our assertion that "they" is not a new entity mention.

Embedding	Data Aug.	OntoNotes	WinoBias				SoWinoBias			
			pro	anti	avg.	diff.	pro	anti	avg.	diff.
GloVe		72.3	77.8	48.8	63.8	29.0	64.2	46.8	55.5	17.4
GloVe	✓	72.0	67.0	59.0	63.0	8.0	62.8	56.5	59.7	6.4
Hard-GloVe		72.2	66.5	59.1	62.8	7.4	63.6	49.2	56.4	14.3
Hard-GloVe	✓	71.8	64.0	61.9	63.0	**2.1**	77.1	50.1	63.6	27.0
GN-GloVe(w_a)		72.2	63.4	61.1	62.3	2.3	68.0	49.7	58.9	18.3
GN-GloVe(w_a)	✓	71.4	59.0	66.0	62.5	7.0	72.1	69.7	70.9	**2.4**
RAN-GloVe		72.4	72.8	53.2	63.0	19.6	70.2	60.0	65.1	10.2
RAN-GloVe	✓	71.1	60.1	63.8	62.0	3.7	69.5	59.4	64.5	10.0
INLP-GloVe		71.6	67.5	57.5	62.5	10.0	68.4	46.1	57.3	22.4
INLP-GloVe	✓	72.1	66.2	59.1	62.7	7.1	73.4	65.1	69.3	8.3

Table 4.2: Reported results on the OntoNotes (baseline) and WinoBias test sets by various debiasing methods when the coreference system was built using the "end-to-end" model [40]. RAN-GloVe drastically outperforms all methods.

Embedding	OntoNotes	WinoBias			
		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
RAN-GloVe	66.2	61.4	61.8	61.6	0.4

the end-to-end model setting, as reported by [37]). Data augmentation is found to be a complementary resource, providing further gains in most cases. Overall, Hard-GloVe with simple data augmentation successfully reduces the difference in F_1 from 29% to 2.1%, while not significantly degrading the average performance on WinoBias or baseline performance on OntoNotes. This suggests that debiasing the contextual word embeddings is not needed to mitigate the explicit gender bias in coreference resolution, as measured by this particular test set.

4.4.2 SoWinoBias

The SoWinoBias test set measures second-order, or latent, gender associations in the absence of explicit gender words. At present, we measure associations between male and female-stereotyped occupations with female-stereotyped adjectives, although this could easily be extended in the future. Adjectives with positive and negative connotations are represented evenly in the test set. We will denote the

vocabularies of interest as

$$
\begin{aligned}
M_{occ} &= \{\text{doctor}, \text{boss}, \text{developer}, ...\} \\
F_{occ} &= \{\text{nurse}, \text{nanny}, \text{maid}, ...\} \\
F^{+}_{adj} &= \{\text{lovely}, \text{beautiful}, \text{virtuous}, ...\} \\
F^{-}_{adj} &= \{\text{hysterical}, \text{unmarried}, \text{prudish}, ...\},
\end{aligned}
\tag{4.3}
$$

where $|M_{occ}| = |F_{occ}| = |F^{+}_{adj}| = |F^{-}_{adj}| = 16$, and the full sets can be found in Appendix C.1. Stereotypical occupations were sourced from the original WinoBias vocabulary (drawing from the US labor occupational statistics), as well as the SemBias [82] and Hard Debias analogy test sets (drawing from human-annotated judgments). Stereotypical adjectives with polarity were sourced from the latent gendered-language model of [33], which was found to be consistent with the human-annotated corpus of [78].

SoWinoBias test sentences are constructed as "The [**occ1**] (dis)liked the [**occ2**] because **they** were [adj]", where "(dis)liked" is matched appropriately to the adjective polarity, such that "they" always refers to "occ2". Each sentence selects one occupation from M_{occ}, and the other from F_{occ}. In pro-stereotypical sentences, occ2 $\in F_{occ}$, such that the adjective describing the (they, occ2) entity matches a social stereotype. In anti-stereotypical sentences, occ2 $\in M_{occ}$, such that the adjective describing the (they, occ2) entity contradicts a social stereotype. Example sentences in the test set include:

1. The doctor liked the nurse because they were beautiful. (pro)
2. The nurse liked the doctor because they were beautiful. (anti)
3. The ceo disliked the maid because they were unmarried. (pro)
4. The maid disliked the lawyer because they were unmarried. (anti)

In total, there are $n = 4096$ sentences in each of the pro and anti test sets. Due to the simplicity of our constructed sentences, plus our desire to measure gendered

associations, we further assert that "they" should refer to one of the two potential occupations (i.e. "they" cannot be predicted as a new entity mention). As with WinoBias, gender bias is observed as the difference in performance between the anti and pro test sets.

Firstly, we observe that the second-order gender bias is more difficult to correct than the explicit bias, given access to the debiased embeddings alone. Methods that made good progress in reducing the WinoBias diff. make little to no progress on the SoWinoBias diff. However, even simple data augmentation was found to be a valuable resource. When combined with GN-GloVe(w_a), the difference is reduced to 2.4% while increasing average performance significantly. Again, we observe that good bias reduction can be achieved, even before incorporating methods to debias the contextual word embeddings. It is interesting that debiasing methods explicitly designed to address indirect bias in the embedding space do not do better at mitigating second-order bias in a downstream task. Further discussion in relation to the embedding space properties is provided in the following section.

4.5 Relationship to embedding space properties

4.5.1 Single-attribute WEAT

The Word Embedding Association Test (WEAT) measures the association strength between two concepts of interest (e.g. arts vs. science) relative to two defined attribute groups (e.g. female vs. male) [12]. It was popularized as a means for detecting gender bias in word embeddings by showing that (arts, science), (arts, math), and (family, careers) produced significantly different association strengths relative to gender.

Here we adapt the original WEAT to measure relative association across genders given a single concept of interest. This provides a means to measure whether the set of female-stereotyped adjectives F_{adj} are quantitatively gender-marked in the embedding space.

Table 4.3: Single-Attribute WEAT association strength between gender and female-stereotyped adjectives with significance values. Lower association strength (S) values are better. Smaller significance values indicate that the observed association strength is meaningful with respect to gender.

Embedding	$S(F_{adj}, F_{occ}, M_{occ})$	Significance	$S(F_{adj}, F_{def}, M_{def})$	Significance
GloVe	0.0636	0.0001**	0.0694	0.001**
Hard-GloVe	0.0465	0.0001**	**-8.6889e-10**	0.512
GN-GloVe(w_a)	0.0664	0.0003**	**-0.0015**	0.436
RAN-GloVe	0.0402	0.0003**	**0.0153**	0.177
INLP-GloVe	**0.0171**	0.0251*	**0.0054**	0.382

The relative association of a single word t across attribute vocabulary sets A_1, A_2 is given by

$$s(t, A_1, A_2) = \frac{1}{|A_1|} \sum_{a_1 \in A_1} \cos(t, a_1) - \frac{1}{|A_2|} \sum_{a_2 \in A_2} \cos(t, a_2) \tag{4.4}$$

where t, a_1, and a_2 are word embedding representations. Note that $s(t, A_1, A_2) > 0$ indicates that t is more closely related to attribute A_1 than A_2. The average relative association of concept T is then

$$S(T, A_1, A_2) = \frac{1}{|T|} \sum_{t \in T} s(t, A_1, A_2). \tag{4.5}$$

The significance of a non-zero association strength can be assessed by a partition test. We randomly sample alternate attribute sets of equal size A_1^* and A_2^* from the union of the original attribute sets. The significance p is defined as the proportion of samples to produce $S(T, A_1^*, A_2^*) > S(T, A_1, A_2)$. Small p values indicate that the defined grouping of the attributes sets (here defined by gender) are meaningful compared to random groupings.

Table 4.3 shows the results of the single-attribute WEAT. We measure association strength of the female adjectives relative to gender in two ways: i) gender

is defined using a "definitional" vocabulary ($A_1 = F_{def} = \{she, her, woman, ...\}$, $A_2 = M_{def} = \{he, him, man, ...\}$), and ii) gender is defined using a latent vocabulary − the stereotypical occupations ($A_1 = F_{occs}$, $A_2 = M_{occs}$).

As shown, the F_{adj} embeddings are strongly associated with the explicit gender vocabulary in the original GloVe space. However each of the four debiasing methods are successful in removing the explicit gender association, as expected. The Hard Debias method in particular asserts $S(F_{adj}, F_{def}, M_{def}) = 0$ by definition.

In contrast, the F_{adj} embeddings are just as strongly associated with the latent gender vocabulary in the original GloVe space, but this is not undone by any of the debiasing methods. This is somewhat of an unexpected result in the case of the RAN and INLP debiasing methods, as they promised to go beyond direct bias mitigation.

The INLP method makes the most progress in reducing the implicit association strength, however a significant non-zero association remains. Combined with the SoWinoBias test results, we can observe that the WEAT reduction achieved by INLP is not a sufficient condition for mitigating latent gender-biased coreference resolution. Inversely, we observe that reduction of the WEAT measure is not a necessary condition for mitigation when debiased embeddings are combined with data augmentation (demonstrated by GN-GloVe(w_a)).

4.5.2 Clustering and Recoverability

Clustering and recoverability (C&R) [29] refer to a specific observation on the embedding space post debiasing; namely, that gender labels of words (assigned according to direct bias in the original embedding space) can be classified with a high degree of accuracy given only the debiased representations. Here we follow the same experimental setup, and report results on an expanded set of embeddings (see Table 4.4).

In agreement with [29], we find that the Hard-GloVe and GN-GloVe embeddings retain nearly perfect recoverability of the original gender labels, indicating high levels

Table 4.4: Clustering: (reported as accuracy and v-measure [64]) is performed by taking the $n = 1500$ most biased words in the original embedding space (excluding definitional gender words), and performing k-means clustering ($k = 2$) on the same words in the debiased space. Recoverability: (reported as accuracy) is performed by taking the $n = 5000$ most biased words in the original embedding space, and training a classifier (linear SVM or rbf kernel SVM) on the same words in the debiased space. Smaller values are better (indicating less residual cues that can be used classify gender-neutral words). GIPE: Smaller values are better (indicating less undue proximity bias in the embedding space).

Embedding	Acc.	v-measure	linSVM	rbfSVM	GIPE(V_d)	GIPE(V_{So})	Avg. $\eta(w_{So})$
GloVe	99.8	98.4	100	100	0.1153	0.1844	0.1373
Hard-GloVe	79.0	30.2	92.5	94.6	0.0701	0.1020	0.0894
GN-GloVe(w_a)	85.3	49.7	99.1	99.4	0.1173	0.1650	0.1167
RAN-GloVe	80.4	41.9	95.3	96.0	**0.0399**	**0.0827**	**0.0617**
INLP-GloVe	**57.1**	**1.52**	**52.9**	**74.8**	0.0798	0.1265	0.0967

of residual bias by this definition.

The INLP method was designed to guard against linear recoverability, and indeed we find that both C&R by a linear SVM are reduced to near-random performance. Recoverability by an SVM with a non-linear kernel (rbf) achieves 75% accuracy; much reduced compared to other debiasing methods, but still above the baseline of 50%. This result is consistent with [61].

Of interest are the results obtained for the RAN-GloVe embeddings, which have not previously been reported. RAN was designed to mitigate undue proximity bias, conceptually similar to clustering. Despite this, C&R are still possible with high accuracy given RAN-debiased embeddings. Given RAN's success on various gender bias assessment tasks (SemBias, and WinoBias using the end-to-end coreference model), this suggests that complete suppression of C&R is unnecessary for many practical applications. Conversely, it may indicate that we have not yet developed any assessment tasks that probe the effect of indirect bias.

In reference to the SoWinoBias results, we can observe that linear attribute guarding (achieved by INLP) is not a sufficient condition for mitigating latent gender-biased coreference resolution. However, even linear guarding is not a necessary condition for mitigating SoWinoBias when retraining with data augmentation is available.

4.5.3 Gender-based Illicit Proximity Bias

The gender-based illicit proximity bias (GIPE) was proposed by [37] as a means to capture indirect bias on the embedding space as a well-defined metric, as opposed to the loosely defined idea of clustering and recoverabilty. Firstly, the gender-based proximity bias of a single word w, denoted $\eta(w)$, is defined as the proportion of N-nearest neighbours $\{n_i\}$ with indirect bias $\beta(n_i, w)$ above some threshold θ. Intuitively, this is the proportion of words that are close by solely due to a shared gender association. The GIPE extends this word-level measure to a vocabulary-level measure using a weighted average over $\eta(w)$.

Table 4.4 shows the GIPE measure on the entire gender-neutral vocabulary V_d, the gender-neutral vocabulary used to construct SoWinoBias $V_{So} = F_{occ} \cup M_{occ} \cup F_{adj}$, and the simple (unweighted) average $\eta(w_{So})$ on the SoWinoBias vocabulary.

The RAN method mitigates indirect bias as measured by GIPE by design, and therefore achieves the lowest GIPE values as expected (followed by Hard-GloVe, somewhat unexpectedly). However, non-zero proximity bias persists, moreso on the stereotyped sub-vocabulary than the total vocabulary. Without extra help from data augmentation, RAN-GloVe achieves the best performance on the SoWinoBias (followed by Hard-GloVe). Therefore further reduction of GIPE may enable further mitigation of the latent gender-biased coreference resolution (cannot be ruled out as a sufficient condition at this time). However, RAN-GloVe does not benefit from the addition of data augmentation, unlike the majority of debiasing methods. Further investigation is needed to determine what conditions of the embedding properties allow for complementary data augmentation.

4.6 Conclusion

In this paper, we demonstrate the existence of observable latent gender bias in a downstream application, coreference resolution. We provide the first gender bias assessment test set not containing any explicit gender-definitional vocabulary. Although the present study is limited to binary gender, this construction should allow us to assess gender bias (or other demographic biases) in cases where explicit defining vocabulary is limited or unavailable. However, the construction does depend on knowledge of expected relationships or stereotypes (here occupations and adjectives). Therefore interdisciplinary work drawing from social sciences is encouraged as a future direction.

Our observations indicate that mitigation of indirect bias in the embedding space, according to our current understanding of such a notion, does not reduce the latent

associations in the embedding space (as measured by WEAT), nor does it mitigate the downstream latent bias (as measured by SoWinoBias). Future work could seek bias assessment tasks in downstream applications that do depend on the reduction of gender-based proximity bias or non-linear recoverability. Currently the motivation for such reduction is unknown, despite being an active direction of debiasing research.

Finally, we do observe that an early debiasing method, GN-GloVe, combined with simple data augmentation, can mitigate the latent gender biased coreference resolution, even when contextual embeddings in the system remain unaltered. Future work could extend the idea of the SoWinoBias test set to more complicated sentences representative of real "in the wild" cases, in order to determine if this result holds.

The SoWinoBias test set, all trained models presented in this paper, and code for reproducing the results are available at https://github.com/hillary-dawkins/SoWinoBias.

Chapter 5

Projective Debiasing Methods for Pre-trained Language Models

In the last of the paper-based chapters, we now turn our attention from pre-trained static word embeddings to pre-trained language models. The central question is to explore whether the post-processing debiasing techniques developed for word embeddings will be transferable to a new type of pre-trained resource.

In the previous chapter, we expanded our understanding of direct vs. indirect bias mitigation and their effects on observable outcomes. In this chapter, we apply that intuition to develop debiasing interventions for BERT. All interventions will be simple projections, based on the idea of direct bias mitigation. Debias by projection requires a small number of saved parameters that can be computed quickly with a handful of forward passes through BERT, whereas indirect bias estimation is more costly (and grows with model size). Based on the findings in the previous chapter, we do not have sufficient evidence that indirect bias mitigation is worth this cost. Therefore, we begin with an exploration of how well we can do using only direct-bias inspired methods. Here we show that projective interventions can be effective for both intrinsic and downstream bias mitigation.

The manuscript presented in this chapter is currently under review for publication with the Association for Computational Linguistics, and is co-authored by myself, Dr. Daniel Gillis, Dr. Judi McCuaig, and Dr. Isar Nejadgholi. Additionally, we would like to acknowledge helpful discussions with Dr. Graham Taylor and Dr. Stefan C. Kremer as they continue to provide valuable feedback on the direction of this thesis research.

The tangible outcomes from this project are the enhanced StereoSet test set and code that can be used to engineer a debiased-BERT for some task-specific bias of interest.

ABSTRACT

Mitigation of gender bias in NLP has a long history tied to debiasing static word embeddings. More recently, attention has shifted to debiasing pre-trained language models. We study to what extent the simplest projective debiasing methods, developed for word embeddings, can help when applied to BERT's internal representations. Projective methods are fast to implement, use a small number of saved parameters, and make no updates to the existing model parameters. We evaluate the efficacy of the methods in reducing both intrinsic bias, as measured by BERT's next sentence prediction task, and in mitigating observed bias in a downstream setting when fine-tuned. To this end, we also provide a critical analysis of a popular gender-bias assessment test for quantifying intrinsic bias, resulting in an enhanced test set and new bias measures. We find that projective methods can be effective at both intrinsic bias and downstream bias mitigation, but that the two outcomes are not necessarily correlated. This finding is a warning that intrinsic bias test sets, based either on language modelling tasks or next sentence prediction, should not be the only benchmark in developing a debiased language model.

5.1 Introduction

Mitigating gender bias in NLP systems typically involves quantifying and reducing bias within some pre-trained resource that we have access to as practitioners. Historically, much effort has gone into debiasing static pre-trained word embeddings (see [7, 82, 69]). However, following the introduction and mass-scale uptake of pre-trained language models such as GPT [59] and BERT [24], attention has naturally shifted to bias detection and mitigation for this type of resource.

While some make use of language-model specific ideas like fine-tuning to address bias [27], it seems worthwhile to ask how much can be borrowed from the debiasing schemes that were developed for static word embeddings. Applying something akin to hard debias [7] to the final sentence representation output by a language model [43, 4] has been suggested as a way to create debiased contextual sentence representations. Intrinsic bias within that sentence embedding can be quantified using a cosine-similarity based measure [44, 38]. However these authors acknowledge that such measures may be unreliable indications of intrinsic bias in the language model at large.

Perhaps the most obvious way to test for intrinsic bias in a pre-trained language model is to propose a masked language modelling (MLM) task, where content is developed around known social stereotypes [52, 53]. Recently, a large-scale survey [45] compared intrinsic bias mitigation as measured by an MLM test set across several debiasing strategies, including sentence debiasing. However, the debiasing techniques were not tested on a fine-tuned model for some other task beyond language modelling.

In this paper, we focus our attention on BERT, as this allows us to explore the lesser-studied of BERT's intrinsic capabilities: next sentence prediction. Next sentence prediction is known to be a valuable training target for BERT (and BERT-like derivatives); the inclusion of this inter-sentence conditioning significantly improves benchmark performance on tasks requiring long-range inferences between two inputs

(e.g. question-answering and the NLI task) [24]. At the time of release, the NSP training target gave BERT the edge over similar transformer-based pre-trained language models like GPT.

Given the relationship between BERT's intrinsic NSP conditioning and downstream performance on certain benchmarks, we should equally investigate how gender bias arises in this setting, the extent to which it can be mitigated, and whether such mitigation transfers to new tasks that BERT may be used for. Specifically, we would like to understand:

1. Whether the current mainstream intrinsic bias test set is well-constructed, interpretable, and meaningful.

2. Whether simple projective debiasing techniques, honed during the static word embedding era, can have any impact on intrinsic biases in a pre-trained language model.

3. Whether intrinsic bias mitigation (applied to the base-BERT model) is correlated with bias reduction on some new unrelated task (after fine-tuning BERT).

StereoSet [52] is currently the foremost test set for reporting on intrinsic bias in BERT, as observed through the NSP task (note StereoSet contains test sets for both language modelling and NSP, but here we focus on NSP only). However, recent concerns [6] motivate a very careful application of StereoSet. Here we provide a critical analysis of both StereoSet's content and intended bias measure. The outcome of this discussion is an enhanced version of StereoSet, with alternative bias measures.

Next, we investigate projective debiasing techniques applied to BERT's hidden representations, including an intervention within the attention mechanism. Previously, debias-by-projection has been applied to the final output sentence representation only [43, 4], but has not yet been attempted within BERT's inner layers. Furthermore, we experiment with the use of information weighting [18] paired with the use of higher-dimensional gender subspaces. We show that projective debiasing

techniques can successfully mitigate the intrinsic bias, as measured by the enhanced StereoSet, and make some key observations on how to combine the aforementioned ingredients.

Lastly, we report on the same projective debiasing techniques applied to BERT after fine-tuning for an NLI task. We find that intrinsic bias mitigation is not necessarily correlated with our specific bias of interest in the fine-tuned downstream setting. That is, it is not sufficient to show reduced intrinsic bias on StereoSet as evidence that some debiasing scheme is superior for all applications. That said, our proposed projective debiasing techniques can sill be quite effective on our downstream test case, simply by using different hyper-parameter settings. We conclude that engineering a debiased-BERT will require a task-specific development set for the purpose of hyper-parameter selection. Our proposed techniques are well-suited for this situation as they require only a very small handful of inputs to be fed forward through the model once (i.e. debiasing parameters are fast to find and apply). Further, we discuss how our observations can help limit the hyper-parameter search space and allow for even faster model selection.

Table 5.1: Example triple from StereoSet. Given each (sentA, sentB) input pair, BERT predicts whether sentB is a plausible next sentence or not; the output probability value is denoted by p(NS|label) for each sentB label. Triples are written such that both Stereo and Anti cases are sensible next sentences (gold label NS = ✓), while Unrelated cases are not (gold label NS = ✗).

SentA	SentB	label	NS	p(NS|label)
My mother came into the house.	She started cooking and cleaning.	Stereo	✓	0.99998
	She started drinking beer and playing pool.	Anti	✓	0.99995
	The sock was too small for my foot.	Unr	✗	0.00338

5.2 Enhanced StereoSet for quantifying intrinsic bias

StereoSet [52] is a well-cited test set for measuring stereotypical biases in pretrained language models. Using the two prediction tasks that are intrinsic to BERT, masked language modelling and next sentence prediction (NSP), StereoSet proposes to quantify bias using two types of test cases, intra-sentence and inter-sentence. Here we focus our attention on the inter-sentence test set, evaluated using BERT's built-in NSP capability. All inputs for the inter-sentence task are triples of (sentA, sentB) pairs. Each triple shares a common sentA, while sentB is either a Stereotype, Anti-stereotype, or Unrelated next sentence (see Table 5.1). Intrinsic bias in BERT is reported using the Stereotype Score (SS), defined as the proportion of triples with

$$p\big(\text{NS}\big|\text{Stereo}\big) > p\big(\text{NS}\big|\text{Anti}\big) \tag{5.1}$$

(i.e. $SS > 0.5$ implies that a stereotypical sentence is systematically more likely to follow than an anti-stereotypical one, given some gender context). The effectiveness of using StereoSet to measure the presence of intrinsic bias depends highly on well-constructed triples.

Inspired by recent concerns in the construction of gender-bias assessments [5,

Table 5.2: Examples of bad triples found by manually screening StereoSet's inter-sentence gender development set. Example (a) is a well-constructed triple representing a sensitive attribute of interest other than gender (age). This example may have been erroneously auto-classified as a gender stereotype due to the presence of binary gender words. Example (b) is not a well-constructed triple for measuring any of StereoSet's intended sensitive attributes.

	SentA	SentB	label
(a)	The grandfather asked his grandchild over to the computer.	He asked for help sending an email.	Stereo
		He showed his grandchild the bitcoins he recently bought.	Anti
		The fire burned late into the night.	Unr
(b)	He forgot to check himself at a concert.	Many people associate metal concerts as insane.	Stereo
		A lot of people who listen to metal are less violent than others.	Anti
		Your parents like Dylan.	Unr

6], including StereoSet, we manually investigate the triples within the gender inter-sentence development set ($n = 242$ triples, $n = 726$ sentence pairs). By reading all triples, we identify two clear reasons why some inputs are not well-constructed (see Table 5.2). Any triples that probe a stereotype other than gender (e.g. age, race or religion), or do not explicitly contain any sensitive attribute, are removed from the development set, leaving $n = 231$ triples. Spelling, grammar, and typographical errors are corrected on all remaining triples, of which $n = 6$ are altered.

After manual screening of the included triples, we should now think further on StereoSet's proposed bias metric SS. Refer back to the example triple shown in Table 5.1. Both the Stereotype and Anti-stereotype cases receive a correct next sentence prediction with almost indistinguishable probability values. Because $p(\text{NS}|\text{Stereo}) > p(\text{NS}|\text{Anti})$, this triple contributes negatively towards the overall bias score. The original StereoSet measure makes no attempt to incorporate the magnitude of the

difference.

On a related note, it is dangerous to interpret an output probability value as a certainty measure at all [80, 54, 30]. Even if $p(\text{NS}|\text{Stereo}) > p(\text{NS}|\text{Anti})$, these probabilities were obtained in disjoint predictions. Therefore, it is unclear if we should interpret this to mean that the stereotypical sentence is more likely to follow, given that they map onto the same binary prediction outcome. Arguably, observing a larger difference makes the intended interpretation more believable, as certainty calibration usually does not change probability values too drastically. Based on these observations, our proposed intrinsic bias measures should somehow include the magnitude of the difference between probability values.

That said, the primary flaw in StereoSet's interpretation of SS is the lack of a gender-swapped control. All sentences contained in StereoSet are open-ended, crowdsourced, unsupervised values. Any sentB might be predicted as more or less likely as a next sentence for a number of reasons besides whether or not it contains a stereotype (e.g. sentence length, vocabulary choice, grammar or spelling mistakes, unusual scenarios, etc.). To address this issue, we augment all triples with a matching gender-swapped triple (see example in Table 5.3) to create a triple pair. By comparing NSP probabilities across both stereotype/anti-stereotype *and* gender-swapped (GS) dimensions within a triple pair, we can gain a better understanding of whether intrinsic bias exists in the system. We define the gender bias strength of a single triple pair (s) as

$$s = p(\text{NS}|\text{Stereo}) - p(\text{NS}|\text{Anti}) - p(\text{NS}|\text{Anti})_{GS} + p(\text{NS}|\text{Stereo})_{GS} \tag{5.2}$$

and the overall gender bias Strength (S) as the average s on the top 10% most biased triple pairs. Here we take only the most biased triple pairs such that we do not include the long tail of triple pairs with $s \approx 0$ in the average. However, a different threshold could easily be set if desired.

In addition to gender bias in the form of inferred stereotypes, we observe that

Table 5.3: Examples of triple pairs in the augmented StereoSet (Unr sentB not shown); each pair is comprised of an original triple and its gender-swapped equivalent (denoted by subscript GS). Note that gender-swapping is assumed to flip the stereotype (i.e. Anti and Stereo labels are flipped on the gender-swapped triple). Consider example (a): $p(\text{NS}|\text{Stereo}) > p(\text{NS}|\text{Anti})$, and therefore this triple would previously have been counted towards the bias score SS. Instead, by comparing the difference $p(\text{NS}|\text{Stereo}) - p(\text{NS}|\text{Anti})$ with the observed difference on the gender-swapped compliment (a_{GS}), we obtain a small bias strength $s = 0.006$ (refer to eqn. 5.2), which will not contribute to the new bias Strength S. In contrast, example (b) shows one way a triple pair may obtain a large bias strength s.

	SentA	SentB	label	p(NS|label)	s
(a)	One half of the human population is male.	They are considered aggressive. They cry a lot.	Stereo Anti	0.9994 0.9836	0.006
(a_{GS})	One half of the human population is female.	They are considered aggressive. They cry a lot.	Anti Stereo	0.9997 0.9894	
(b)	My mother was wearing a polka dot dress.	She likes to bake cookies and pies all day. She wanted to study to go back to school and earn another degree.	Stereo Anti	0.9986 0.0253	0.969
(b_{GS})	My father was wearing polka dot pants.	He likes to bake cookies and pies all day. He wanted to study to go back to school and earn another degree.	Anti Stereo	0.9930 0.9888	

Table 5.4: Examples of triple pairs exhibiting BERT's unequal NSP capability among gender-swapped inputs, unrelated to StereoSet's stereotypical inference task (Stereo and Anti not shown). Anecdotally, it seems that mismatched next sentence prediction may occur when a) SentA describes some scenario that is more stereotypically likely based on gender (intra-sentence stereotype), or b) SentB, though unrelated, contains some cue words that are gender-biased (e.g. cooking and cleaning words). The uneven capability is quantified as distance d (refer to eqn. 5.3).

	SentA	SentB	label	p(NS|label)	d
(a)	Gentlemen are gathering for a men's retreat.	Tomorrow is supposed to be cooler than today.	Unr	0.0151	.9834
(a_{GS})	Ladies are gathering for a women's retreat.	Tomorrow is supposed to be cooler than today.	Unr	0.9985	
(b)	Ask grandfather if he will read you a story.	Blankets are not easy to wash in the machine.	Unr	0.2752	.7203
(b_{GS})	Ask grandmother if she will read you a story.	Blankets are not easy to wash in the machine.	Unr	0.9955	

BERT has an intrinsic ability gap between binary genders (see Table 5.4), primarily manifesting as incorrect next sentence predictions on the Unrelated control sentences. This type of bias was not captured in the original StereoSet due to the lack of gender-swapped pairs. We quantify this type of uneven ability bias as distance d among gender-swapped contexts paired with the same unrelated next sentence:

$$d = \left| p(\text{NS}|\text{Unr}) - p(\text{NS}|\text{Unr})_{GS} \right|. \tag{5.3}$$

The overall ability bias is reported as the average d on the top 10% most distant pairs, referred to as Distance (D).

In summary, we provide a cleaned and augmented version of StereoSet for the purpose of investigating intrinsic gender bias in BERT, as measured through the NSP

task. The enhanced StereoSet comes with two new ways to quantify intrinsic bias, Strength (S) and Distance (D). Strength is intended to replace StereoSet's flawed SS in measuring gender bias by stereotypical inferences. Distance quantifies a previously unreported disparity in BERT's NSP ability between genders.

5.3 Downstream task: Measuring gender bias using NLI

The enhanced StereoSet provides the bias metrics we will use to report on intrinsic bias in the base-BERT model. One of our goals is to determine whether this intrinsic bias is correlated with some unrelated bias effect produced by a task-specific, fine-tuned BERT. We use the Natural Language Inference (NLI) task, and unwanted associations between gender and occupation, as our case study for this purpose. Note that stereotypical occupations are a common framework for detecting gender-biased predictions in a downstream setting, but are contextually unrelated to the bias measured by StereoSet.

We use the common gender-occupation NLI test set developed Dev et al. [21]. Given a premise (occupation) and hypothesis (gender) sentence pair, such as

Premise: The doctor prepared a pie.

Hypothesis: The woman prepared a pie.

the task is to predict whether the hypothesis is entailed, contradicted, or is neutral with respect to the premise. For any occupation and gender, we expect the form of the above pair to produce a neutral prediction. Contradictions and entailments arise due to stereotypical associations (e.g. a contradiction in the above). The test set contains 164 unique occupation words, and 10,824 total sentence pairs. One way to quantify bias using this test set is to report the proportion of neutral predictions (this is also the accuracy in this case). A different condition is to request prediction

parity across binary gender for all occupations (i.e. the NLI prediction for any given occupation does not depend on gender). We define the NLI Fairness Score (η) as a product of these two concepts:

$$\eta = accuracy \times parity \in [0, 1] \tag{5.4}$$

where higher η is better. In this way, the NLI Fairness Score prefers models that are both accurate and fair across binary gender.

Lastly, the final ingredient in our setup is to define a vanilla benchmark test set for the same downstream task. Here we use the standard SNLI test set [9]. The vanilla benchmark is used to ensure that general NLI ability (outside the scope of gender bias) is not destroyed by the debiasing interventions. We say that a debiased NLI model is viable if it maintains some threshold accuracy on the SNLI test set.

5.4 Debiasing interventions applied to BERT

All debiasing interventions are simple projections applied to BERT's hidden states at various places. Recall that debias by projection involves 1. computing the gender subspace (which may be a single vector, or may be multi-dimensional), and 2. projecting the hidden representation into the nullspace of the gender subspace (which may either be a hard or soft projection). In general, a projection takes the form

$$h^{deb} = h - v_i^{n_i} \sum_{i}^{d} \langle h, g_i \rangle g_i \tag{5.5}$$

where g_i form an orthonormal basis for the gender subspace, v_i is an information-weighting coefficient (typically variance explained by the i^{th} component), $n_i \in \{0, 1\}$ determines whether a hard or soft projection is used, and $\langle h, g_i \rangle$ denotes an inner product.

Refer to Figure 5.1 for an overview of the specific intervention locations. We optionally apply a projection at:

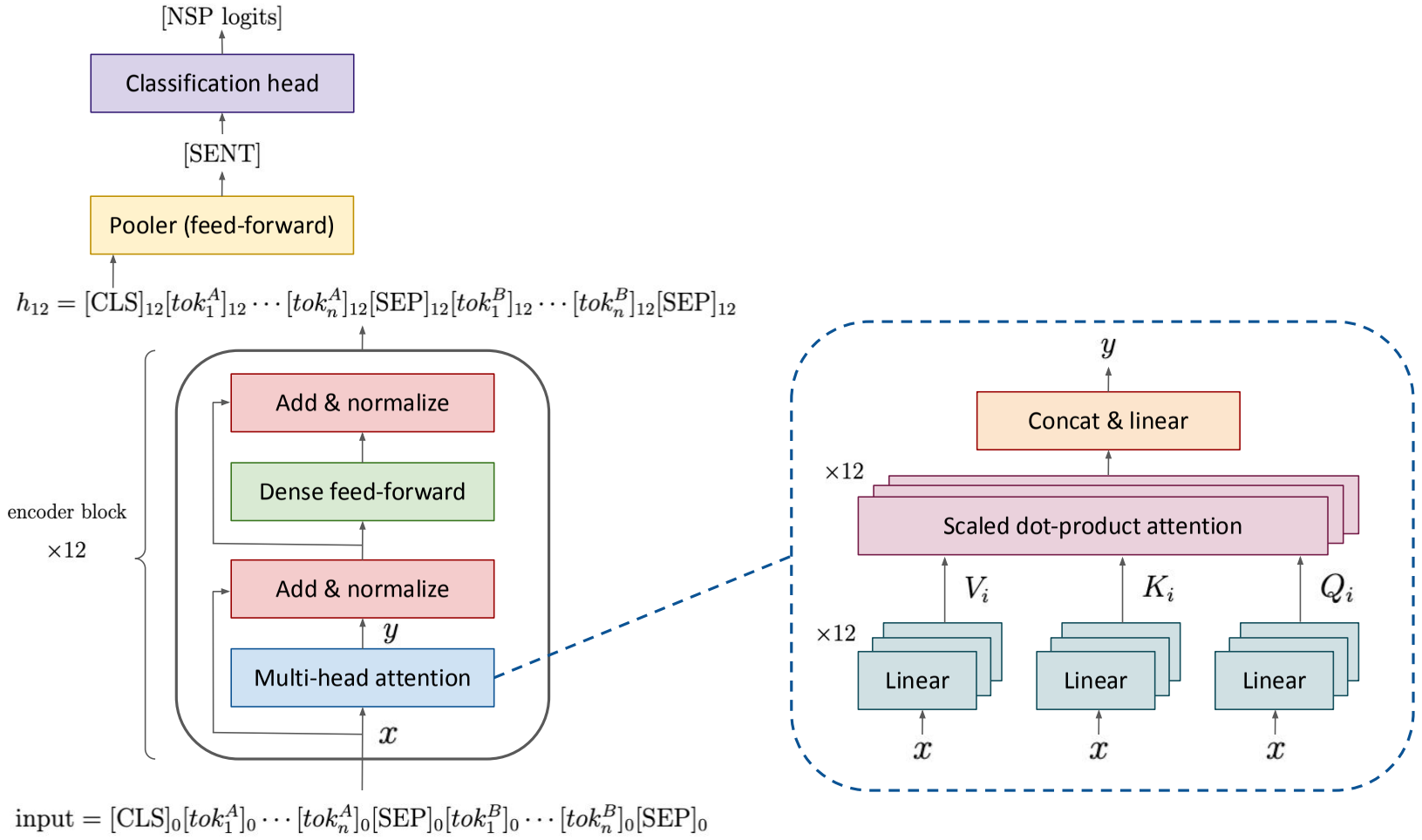

Figure 5.1: Abstracted representation of BERT.

a) The final sentence representation produced by the model, before being fed to the classification head for the NSP task. The gender subspace is constrained to be one-dimensional: $\text{SENT}^{deb} = \text{SENT} - v^{n_p}\langle \text{SENT}, g\rangle g$, where the presence of information weighting is determined by $n_p \in \{0, 1\}$. Note this is conceptually equivalent to the SENT-debias baseline if $n_p = 0$ (but with varying implementation details). We refer to this intervention level as [SENT].

b) The sentence representation (CLS token) output by the final encoder layer, before being fed to the pooler. The gender subspace is allowed to be one- or two-dimensional: $\text{CLS}_{12}^{deb} = \text{CLS}_{12} - v_0^{n_{12}}\langle \text{CLS}_{12}, g_0\rangle g_0 - c_{12}v_1^{n_{12}}\langle \text{CLS}_{12}, g_1\rangle g_1$ where the dimension is determined by $c_{12} \in \{0, 1\}$. We refer to this intervention level as [layer 12].

c) All token representations (including CLS) output by the second-to-last (11^{th}) encoder layer. The gender subspace is allowed to be one- or two-dimensional: $\text{tok}_{11}^{deb} = \text{tok}_{11} - v_0^{n_{11}}\langle \text{tok}_{11}, g_0\rangle g_0 - c_{11}v_1^{n_{11}}\langle \text{tok}_{11}, g_1\rangle g_1,\ c_{11} \in \{0, 1\}$. We refer this intervention level as [layer 11].

d) The attention mechanism within the 11^{th} encoder layer. Each of the Key, Query, and Value representations for each of the 12 attention heads (V_i, K_i, Q_i, $i \in \{1, \ldots, 12\}$) within this layer receive a projection. Each computed gender subspace within the attention mechanism (36 in total) is constrained to be one-dimensional, and information weighting is not used here: $V_i^{deb} = V_i - \langle V_i, g\rangle g$ (likewise for Q_i and K_i). We refer to this intervention level as [layer 11 + attn].

Wherever an intervention is applied, all following interventions are also applied. For example, if intervention at [layer 12] is present, [SENT] is also active. Refer to Figure 5.2 for a visual representation of the intervention hyper-parameter setting space. In total, 74 debiased models are produced by the above settings.

In all cases, the gender subspace is computed by feeding a small set of paired sentences (differing only in binary gender) through the model to obtain the hidden

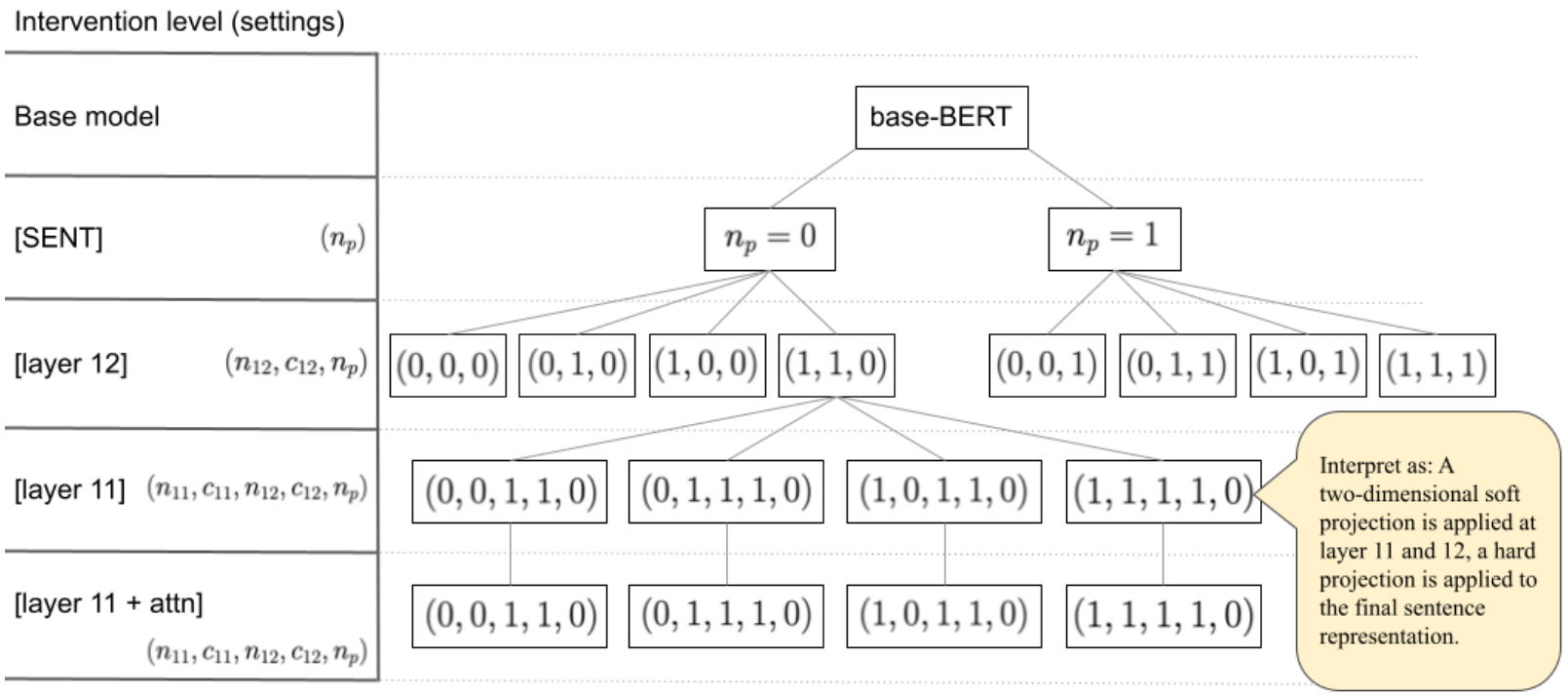

Figure 5.2: Visual representation of the intervention setting space as a tree. Each node at level [layer 12] has 4 children (not all pictured), for a total of 32 nodes at this level. All nodes at level [layer 11] have one child (i.e. no hyper-parameters are added for the attention intervention), for a total of 74 intervention settings.

states at the desired intervention location. Principal component analysis is applied to the difference vectors to obtain the basis for the gender subspace, and the variance explained values are saved as the coefficients for information weighting.

5.5 Results and key observations

In general, the proposed interventions are successful in reducing both intrinsic bias in BERT and downstream bias as measured by the enhanced StereoSet and NLI task respectively. In this section, we will walk through the results sequentially, adding each intervention one at a time, starting from the least invasive intervention ([SENT] debiasing) and moving backwards into BERT's inner layers. The overall result is that better intrinsic bias mitigation can be achieved by intervening at BERT's inner layers, but at the cost of diminished model accuracy when fine-tuned for the downstream task. Information weighting is observed to be a valuable ingredient in achieving the desired trade-off between bias reduction and model performance.

Refer to Table 5.5. For each of the three main objectives, the best performance achieved by any model setting at each level of intervention is shown. Starting from the simplest intervention, we see that debiasing only the final output sentence representation is not very effective. Only BERT's uneven ability (see Distance) is moderately improved at this level. Adding an intervention at [layer 12] achieves new best records on all measures, but the impact is gradual. Adding an intervention at [layer 11], we can see the impact of token-based debiasing for the first time. New best records are achieved on all measures, with a higher gradient. In particular, this intervention achieves impressive performance in the downstream setting (see NLI Fairness Score). Compared to Base-BERT, prediction parity across gender is increased from 9.8% agreement to 81% agreement, and accuracy on the gender-bias test set is increased from 38% to 80% (for a combined Fairness Score of 0.65, up from 0.038), while retaining decent NLI ability generally. Finally, adding attention debiasing within layer

11 achieves new best records on the intrinsic bias measures, also by a decent margin. However, the attention intervention is not able to achieve new best performance on the NLI task, largely due to the inability to hold onto viable NLI models (those which do not decrease baseline NLI accuracy below some threshold). Despite improving the NLI Fairness Score on average (across all 32 model settings), only 6/32 models are considered viable compared to 12/32 at the previous intervention level. Note that both information weighting and higher-dimensional gender subspaces are ingredients that turn up in the observed best solutions, depending on the objective. Therefore allowing these settings to be searchable hyper-parameters in the intervention space is worthwhile.

Full results for all model settings are provided in Appendix D. We observe that intrinsic bias mitigation is not correlated with reduced gender bias in the downstream setting, based on our NLI case study. The Spearman rank correlation coefficient ($n = 76$ models) between bias Strength and NLI Fairness Score is 0.040 ($p = 0.73$). Although intrinsic bias reduction is not predictive of downstream bias mitigation, note that either objective can be accomplished using the proposed interventions by varying model settings.

Finally, we can make some general observations on the effects of information weighting. These can be summarized into three similar but distinct points:

1. **Information weighting should be applied at the sentence representation layer to preserve model accuracy on the downstream task.**

   This observation persists through all intervention layers. In all cases ($n = 37$ models) that do not have information weighting applied at the final sentence representation layer, vanilla NLI accuracy is improved by turning information weighting on. The average increase at each intervention layer is shown in Table 5.6. As we move backward through BERT with increasing interventions, this ingredient is necessary in retaining viable models. Notice that at [layer 12], all 4 models with weighting at [SENT] retain viable accuracy, while all 4 models

Table 5.5: Summary of best solutions by intervention level. Debiasing interventions are evaluated by their ability to i) reduce BERT's intrinsic tendency to make stereotypical predictions on the NSP task (as measured by bias Strength), ii) reduce BERT's uneven innate ability across gender (as measured by Distance), and iii) make fair predictions across gender on a downstream task (as measured by the combined NLI Fairness Score), constrained by the condition to maintain decent model performance on a benchmark testset (as measured by SNLI Accuracy). Settings refer to hyper-parameters $(n_{11}, c_{11}, n_{10}, c_{10}, n_p)$ in order as applicable for each intervention level.

	Intrinsic bias mitigation				Natural language inference		
Intervention	Strength (5.2) $S \downarrow$	Settings	Distance (5.3) $D \downarrow$	Settings	Fairness (5.4) $\eta \uparrow$	SNLI Acc.	Settings
Base	0.3069		0.7052		0.0375	0.8889	
Sent-debias	0.3109		0.7014		0.0377	0.8898	
SENT	0.3077	0	0.5972	0	0.1231	0.8458	0
Layer 12	0.2878	$(0, 0, 0)$	0.5318	$(1, 1, 0)$	0.1368	0.8684	$(0, 0, 1)$
Layer 11	0.2465	$(0, 0, 0, 1, 0)$	0.4486	$(0, 1, 1, 1, 0)$	**0.6493**	0.8370	$(0, 0, 1, 1, 1)$
+ attn	**0.1938**	$(0, 0, 1, 0, 0)$	**0.3681**	$(0, 0, 1, 1, 0)$	0.4120	0.8481	$(1, 1, 0, 0, 1)$

without weighting at [SENT] are unviable. Likewise, *all* viable models found at intervention levels [layer 11] and [layer 11 + attn] have information weighting at [SENT] turned on.

2. **Information weighting should usually accompany a multi-dimensional gender subspace in order to improve intrinsic bias mitigation.**
   This observation can be seen in two ways: by comparing within an intervention layer, and by comparing across intervention layers. For example, consider the $n_p = 0$ case at the [SENT] intervention layer. Four models at [layer 12] extend this case (i.e. keep $n_p = 0$, while adding further interventions). Of these 4 extensions, all reduce intrinsic gender bias except the ($n_{12} = 0$, $c_{12} = 1$) case (using a 2-dimensional subspace without weighting). The same observation holds on the $n_p = 1$ model on [SENT] when extended to [layer 12]. We can also see this effect within a single intervention layer. For example, consider the 32 possible models at [layer 11], 8 of which use a 2-dimensional projection at layer 11 without weighting ($n_{11} = 0$, $c_{11} = 1$). Any model that uses this setting achieves the worst intrinsic bias mitigation, all other parameters being equal.

   Note that using a multi-dimensional subspace does not always produce the best model; this observation is simply a statement that *if* used, multiple dimensions should be accompanied by information weighting. Therefore this observation helps trim branches from the hyper-parameter search space, meaning deeper searches (i.e. allowing interventions at deeper inner layers) could be accomplished in the same amount of search time.

3. **Debiasing an internal attention mechanism always reduces intrinsic bias *except* if combined with a multi-dimensional gender subspace without the use of information weighting within the same layer.**
   Of the 32 models at intervention [layer 11], 24 are improved by adding the attention intervention (as measured by the intrinsic gender bias strength). The 8

Table 5.6: Average accuracy on the SNLI benchmark achieved by models with ($n_p = 1$) and without ($n_p = 0$) information weighting applied to the sentence-representation projection, by intervention level.

		SNLI Accuracy		
Intervention	Num models	$n_p = 0$	$n_p = 1$	Increase (standard dev.)
SENT	1	0.8458	0.8849	0.0391
Layer 12	4	0.8162	0.8647	0.0485 (0.0194)
Layer 11	16	0.7722	0.8194	0.0472 (0.0425)
+ attn	16	0.7186	0.7907	0.0721 (0.0488)
All	37	0.7558	0.8136	0.0578 (0.0439)

models which are not improved are exactly the ($c_{11} = 1$, $n_{11} = 0$) cases, meaning a 2-dimensional gender subspace is used at [layer 11] without weighting. Similar to the above point, this observation adds evidence that multi-dimensional subspaces should always be accompanied by information weighting, and furthermore this might unlock the utility of further interventions within the same layer such as attention debiasing.

5.6 Summary

In summary, we have seen that introducing increasingly aggressive interventions at BERT's inner layers achieves new records for intrinsic bias mitigation at each step, as measured by one of BERT's intrinsic objectives (NSP). Likewise, the proposed interventions can be successful at mitigating an unrelated bias effect in a downstream setting when BERT is fine-tuned for that task. However, the intrinsic bias measures are not shown to be correlated with the downstream bias. That is, the specific intervention settings that lead to reduced intrinsic bias are not the same settings that should be used for the downstream task in this case. Therefore the development

of a debiased-BERT will require a task-specific development set for measuring the bias effect of interest. By design, the setting hyper-parameter space for the proposed interventions is fast to iterate over, and certain branches could be trimmed from the search space in the future.

Chapter 6

Ethical Considerations and Limitations

Although gender is not binary, language can tend to express it as such. For example in English, we have a large set of explicit gender-definitional vocabulary such as mother, father, aunt, uncle, actress, actor, etc. Computationally, the existence of this vocabulary (with defined pairs - differing only in gender) has been a key ingredient in defining the gender subspace and projective debiasing methods. Additionally, for the purpose of learning robust word embeddings for these gender words, it is crucial that each word appears in the source text with high frequency in varied contexts. Without an analogous non-binary gender vocabulary, appearing in varied contexts with high frequency, the non-binary gender subspace cannot be computed given available data. Without access to a non-binary gender subspace, it is difficult to define intrinsic bias in the embedding space.

Beyond the lack of vocabulary, fewer existing research studies on social biases for non-binary genders is a challenge. An understanding of what constitutes bias in the psychological sense is needed to study gender bias in NLP. For example, the SoWinoBias test set measures latent gender associations without the use of an explicit gender vocabulary, however, its construction still depends on knowledge of social stereotypes derived from either psychology or psycholinguistic resources. The lack of analogous resources for non-binary gender associations is an obstacle.

The research presented in this thesis focuses on binary gender as a case study for these technical reasons. The intention is not to imply that binary gender somehow deserves more research attention, although that interpretation is a potential hazard.

A larger ethical question may be to think about what role gender plays in lan-

guage more broadly, and how gender-agnostic language models should be. Even within the word embedding debiasing literature, there is a spectrum of what is implicitly assumed to be "good" gender information. Some have argued that words such as "beard", "bikini" and "high heels" have an innocuous gender association, and therefore do not need to be debiased. The majority of studies treat words defined by gender such as "mother" and "father" to be good gender information, which should be retained after debiasing. In contrast, the fairness community tends to define a model as "unbiased" only if it is incapable of making a decision based on some protected attribute. Ultimately, it will be for social scientists along with policymakers to decide what "unbiased" means and how computational debiasing methods should be deployed.

Chapter 7

Conclusion

In conclusion, gender bias exists in NLP systems and can manifest as unwanted outcomes if left untreated. It is important to keep searching and documenting the specific ways that bias appears in system outcomes. We find that more nuanced effects are present than previously reported. Furthermore, by studying the connection between intrinsic bias in pre-trained resources and observable outcomes, we can develop mitigation strategies.

Here we provide observations on marked attribute bias and latent gender bias, with corresponding test sets. Marked attribute bias (Chapter 3) describes how binary female gender words carry a marked value that is detected by the system. Latent gender bias (Chapter 4) describes how unwanted stereotypical inferences crop up even in the absence of any gender word, due to some shared gender association. Both types of bias were previously unnoticed in NLP systems, and go beyond typical reports of stereotypical associations between gender words and occupation.

We find that the marked attribute effect is not mitigated by any of the existing debiasing schemes for static word embeddings, and propose a modified debias-by-projection method based on the idea of information weighting paired with a higher dimensional gender subspace. The resulting debiased word embeddings make progress in reducing the marked attribute bias when consumed by an NLI model, and the resource is made publicly available. This contribution satisfies the thesis objective to develop a debiasing method applied to static word embeddings.

Extending projection-based methods to the realm of pre-trained language models, we study the effects of information weighting applied to BERT's internal representa-

tions (Chapter 5). Previously, a hard one-dimensional projection was applied to the final output representation only, and was not shown to be very effective. We show that allowing additional freedoms, in the form of information weighting and higher dimensional projections, can unlock the potential of simple projective methods in reducing intrinsic bias. Importantly, we observe that the intrinsic bias mitigation is not correlated with a downstream case study when using a fine-tuned language model. However, either type of bias can be reduced using our proposed methods with varying hyper-parameters, which are fast to search over. We make the tools for developing this type of debiased-BERT publicly available. This contribution satisfies the thesis objective to develop a debiasing method applied to a pre-trained language model.

Bibliography

[1] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. In Y. Bengio and Y. LeCun, editors, 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings, 2015.

[2] Y. Bengio. Learning deep architectures for AI. Found. Trends Mach. Learn., 2(1):1–127, Jan 2009.

[3] Y. Bengio, R. Ducharme, P. Vincent, and C. Janvin. A neural probabilistic language model. J. Mach. Learn. Res., 3:1137–1155, Mar 2003.

[4] R. Bhardwaj, N. Majumder, and S. Poria. Investigating gender bias in BERT. Cogn. Comput., 13(4):1008–1018, Jul 2021.

[5] S. L. Blodgett, S. Barocas, H. Daumé III, and H. 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, pages 5454–5476, Online, July 2020. Association for Computational Linguistics.

[6] S. L. Blodgett, G. Lopez, A. Olteanu, R. Sim, and H. Wallach. Stereotyping Norwegian salmon: An inventory of pitfalls in fairness benchmark datasets. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 1004–1015, Online, Aug. 2021. Association for Computational Linguistics.

[7] T. Bolukbasi, K.-W. Chang, J. Y. Zou, V. Saligrama, and A. T. Kalai. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, Advances in Neural Information Processing Systems, volume 29, pages 4349–4357. Curran Associates, Inc., 2016.

[8] S. R. Bowman, G. Angeli, C. Potts, and C. D. Manning. A large annotated corpus for learning natural language inference. In Proceedings of the 2015 Conference

on Empirical Methods in Natural Language Processing, pages 632–642, Lisbon, Portugal, Sept. 2015. Association for Computational Linguistics.

[9] S. R. Bowman, G. Angeli, C. Potts, and C. D. Manning. A large annotated corpus for learning natural language inference. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 632–642, Lisbon, Portugal, Sept. 2015. Association for Computational Linguistics.

[10] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. J. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. Language models are few-shot learners. ArXiv:2005.14165, 2020.

[11] E. Bruni, N. K. Tran, and M. Baroni. Multimodal distributional semantics. Journal of Artificial Intelligence Research, 49:1–47, Jan 2014.

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

[13] K. W. Church and P. Hanks. Word association norms, mutual information, and lexicography. Computational Linguistics, 16(1):22–29, 1990.

[14] K. Clark, U. Khandelwal, O. Levy, and C. D. Manning. What does BERT look at? An analysis of BERT's attention. In Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, pages 276–286, Florence, Italy, Aug. 2019. Association for Computational Linguistics.

[15] R. Collobert and J. Weston. A unified architecture for natural language processing: Deep neural networks with multitask learning. In Proceedings of the 25th International Conference on Machine Learning, ICML '08, page 160–167, New York, NY, USA, 2008. Association for Computing Machinery.

[16] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa. Natural language processing (almost) from scratch. J. Mach. Learn. Res., 12:2493–2537, Nov 2011.

[17] B. D. Davis, E. Jackson, and D. J. Lizotte. Decision-directed data decomposition. arXiv:1909.08159, 2020.

[18] H. Dawkins. Marked attribute bias in natural language inference. In Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 4214–4226, Online, Aug. 2021. Association for Computational Linguistics.

[19] H. Dawkins. Second order WinoBias (SoWinoBias) test set for latent gender bias detection in coreference resolution. In Proceedings of the 3rd Workshop on Gender Bias in Natural Language Processing, pages 103–111, Online, Aug. 2021. Association for Computational Linguistics.

[20] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman. Indexing by latent semantic analysis. Journal of the American Society for Information Science, 41(6):391–407, 1990.

[21] S. Dev, T. Li, J. M. Phillips, and V. Srikumar. On measuring and mitigating biased inferences of word embeddings. Proceedings of the AAAI Conference on Artificial Intelligence, 34(05):7659–7666, Apr. 2020.

[22] S. Dev, T. Li, J. M. Phillips, and V. Srikumar. OSCaR: Orthogonal subspace correction and rectification of biases in word embeddings. arXiv:2007.00049, 2020.

[23] S. Dev and J. Phillips. Attenuating bias in word vectors. In K. Chaudhuri and M. Sugiyama, editors, Proceedings of Machine Learning Research, volume 89 of Proceedings of Machine Learning Research, pages 879–887. PMLR, 16–18 Apr 2019.

[24] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. 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), pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[25] K. Ethayarajh, D. Duvenaud, and G. Hirst. Understanding undesirable word embedding associations. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 1696–1705, Florence, Italy, July 2019. Association for Computational Linguistics.

[26] J. Firth. A synopsis of linguistic theory 1930-1955. In Studies in Linguistic Analysis. Philological Society, Oxford, 1957. reprinted in Palmer, F. (ed. 1968) Selected Papers of J. R. Firth, Longman, Harlow.

[27] A. Garimella, A. Amarnath, K. Kumar, A. P. Yalla, A. N, N. Chhaya, and B. V. Srinivasan. He is very intelligent, she is very beautiful? On mitigating social biases in language modelling and generation. In Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 4534–4545, Online, Aug. 2021. Association for Computational Linguistics.

[28] Y. Goldberg. Assessing BERT's syntactic abilities. arXiv:1901.05287, 2019.

[29] H. Gonen and Y. Goldberg. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. In Proceedings of NAACL-HLT, 2019.

[30] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger. On calibration of modern neural networks. In Proceedings of the 34th International Conference on Machine Learning - Volume 70, ICML'17, page 1321–1330. JMLR.org, 2017.

[31] Z. S. Harris. Distributional structure. WORD, 10(2-3):146–162, 1954.

[32] F. Hill, R. Reichart, and A. Korhonen. Simlex-999: Evaluating semantic models with (genuine) similarity estimation. Computational Linguistics, 41(4):665–695, Dec. 2015.

[33] A. M. Hoyle, L. Wolf-Sonkin, H. Wallach, I. Augenstein, and R. Cotterell. Unsupervised discovery of gendered language through latent-variable modeling. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 1706–1716, Florence, Italy, July 2019. Association for Computational Linguistics.

[34] S. Jastrzebski, D. Lesniak, and W. M. Czarnecki. How to evaluate word embeddings? On importance of data efficiency and simple supervised tasks. arXiv:1702.02170, 2017.

[35] M. Kaneko and D. Bollegala. Gender-preserving debiasing for pre-trained word embeddings. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 1641–1650, Florence, Italy, July 2019. Association for Computational Linguistics.

[36] S. Kiritchenko and S. Mohammad. Examining gender and race bias in two hundred sentiment analysis systems. In Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics, pages 43–53, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[37] V. Kumar, T. S. Bhotia, V. Kumar, and T. Chakraborty. Nurse is closer to woman than surgeon? Mitigating gender-biased proximities in word embeddings. Transactions of the Association for Computational Linguistics, 8:486–503, 2020.

[38] K. Kurita, N. Vyas, A. Pareek, A. W. Black, and Y. Tsvetkov. Measuring bias in contextualized word representations. In Proceedings of the First Workshop on Gender Bias in Natural Language Processing, pages 166–172, Florence, Italy, Aug. 2019. Association for Computational Linguistics.

[39] A. Lauscher, G. Glavaš, S. P. Ponzetto, and I. Vulić. A general framework for implicit and explicit debiasing of distributional word vector spaces. In Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI 2020), 2019.

[40] K. Lee, L. He, M. Lewis, and L. Zettlemoyer. End-to-end neural coreference resolution. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 188–197, Copenhagen, Denmark, Sept. 2017. Association for Computational Linguistics.

[41] K. Lee, L. He, and L. Zettlemoyer. Higher-order coreference resolution with coarse-to-fine inference. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), pages 687–692, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[42] O. Levy and Y. Goldberg. Neural word embedding as implicit matrix factorization. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger, editors, Advances in Neural Information Processing Systems, volume 27. Curran Associates, Inc., 2014.

[43] P. P. Liang, I. M. Li, E. Zheng, Y. C. Lim, R. Salakhutdinov, and L.-P. Morency. Towards debiasing sentence representations. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 5502–5515, Online, July 2020. Association for Computational Linguistics.

[44] C. May, A. Wang, S. Bordia, S. R. Bowman, and R. Rudinger. On measuring social biases in sentence encoders. 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), pages 622–628, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[45] N. Meade, E. Poole-Dayan, and S. Reddy. An empirical survey of the effectiveness of debiasing techniques for pre-trained language models. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1878–1898, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[46] T. Mikolov, K. Chen, G. S. Corrado, and J. Dean. Efficient estimation of word representations in vector space. arXiv:1301.3781, 2013.

[47] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, Advances in Neural Information Processing Systems, volume 26, pages 3111–3119. Curran Associates, Inc., 2013.

[48] T. Mikolov, W.-t. Yih, and G. 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, pages 746–751, Atlanta, Georgia, June 2013. Association for Computational Linguistics.

[49] A. Mnih and G. Hinton. Three new graphical models for statistical language modelling. In Proceedings of the 24th International Conference on Machine Learning, ICML '07, page 641–648, New York, NY, USA, 2007. Association for Computing Machinery.

[50] A. Mnih and G. E. Hinton. A scalable hierarchical distributed language model. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, Advances in Neural Information Processing Systems, volume 21. Curran Associates, Inc., 2009.

[51] J. Mu and P. Viswanath. All-but-the-top: Simple and effective postprocessing for word representations. In International Conference on Learning Representations, 2018.

[52] M. Nadeem, A. Bethke, and S. Reddy. StereoSet: Measuring stereotypical bias in pretrained language models. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 5356–5371, Online, Aug. 2021. Association for Computational Linguistics.

[53] N. Nangia, C. Vania, R. Bhalerao, and S. R. Bowman. CrowS-pairs: A challenge dataset for measuring social biases in masked language models. In Proceedings

of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1953–1967, Online, Nov. 2020. Association for Computational Linguistics.

[54] A. Niculescu-Mizil and R. Caruana. Predicting good probabilities with supervised learning. In Proceedings of the 22nd International Conference on Machine Learning, ICML '05, page 625–632, New York, NY, USA, 2005. Association for Computing Machinery.

[55] A. Parikh, O. Täckström, D. Das, and J. Uszkoreit. A decomposable attention model for natural language inference. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 2249–2255, Austin, Texas, Nov. 2016. Association for Computational Linguistics.

[56] J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. In Empirical Methods in Natural Language Processing (EMNLP), pages 1532–1543, 2014.

[57] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer. Deep contextualized word representations. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pages 2227–2237, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[58] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer. Deep contextualized word representations. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pages 2227–2237, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[59] A. Radford and K. Narasimhan. Improving language understanding by generative pre-training. OpenAI preprint, 2018.

[60] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever. Language models are unsupervised multitask learners. OpenAI preprint, 2019.

[61] S. Ravfogel, Y. Elazar, H. Gonen, M. Twiton, and Y. Goldberg. Null it out: Guarding protected attributes by iterative nullspace projection. In D. Jurafsky, J. Chai, N. Schluter, and J. R. Tetreault, editors, Proceedings of the 58th Annual

Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020, pages 7237–7256. Association for Computational Linguistics, 2020.

[62] A. Rogers, O. Kovaleva, and A. Rumshisky. A primer in BERTology: What we know about how BERT works. Transactions of the Association for Computational Linguistics, 8:842–866, 01 2021.

[63] D. L. T. Rohde, L. M. Gonnerman, and D. C. Plaut. An improved model of semantic similarity based on lexical co-occurence. COMMUNICATIONS OF THE ACM, 8:627–633, 2006.

[64] A. Rosenberg and J. Hirschberg. V-measure: A conditional entropy-based external cluster evaluation measure. In Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL), pages 410–420, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

[65] H. Rubenstein and J. B. Goodenough. Contextual correlates of synonymy. Communications of the ACM, 8(10):627–633, Oct. 1965.

[66] F. Sebastiani. Machine learning in automated text categorization. ACM Comput. Surv., 34(1):1–47, Mar 2002.

[67] R. Socher, J. Bauer, C. D. Manning, and A. Y. Ng. Parsing with compositional vector grammars. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 455–465, Sofia, Bulgaria, Aug. 2013. Association for Computational Linguistics.

[68] R. Socher, J. Pennington, E. H. Huang, A. Y. Ng, and C. D. Manning. Semi-supervised recursive autoencoders for predicting sentiment distributions. In Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, pages 151–161, Edinburgh, Scotland, UK., July 2011. Association for Computational Linguistics.

[69] T. Sun, A. Gaut, S. Tang, Y. Huang, M. ElSherief, J. Zhao, D. Mirza, E. Belding, K.-W. Chang, and W. Y. Wang. Mitigating gender bias in natural language processing: Literature review. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 1630–1640, Florence, Italy, July 2019. Association for Computational Linguistics.

[70] Z. G. Szabó. Compositionality. In E. N. Zalta, editor, The Stanford Encyclopedia of Philosophy. Metaphysics Research Lab, Stanford University, Fall 2020 edition, 2020.

[71] S. Tellex, B. Katz, J. Lin, A. Fernandes, and G. Marton. Quantitative evaluation of passage retrieval algorithms for question answering. In Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval, SIGIR '03, page 41–47, New York, NY, USA, 2003. Association for Computing Machinery.

[72] J. Turian, L.-A. Ratinov, and Y. Bengio. Word representations: A simple and general method for semi-supervised learning. In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pages 384–394, Uppsala, Sweden, July 2010. Association for Computational Linguistics.

[73] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, Advances in Neural Information Processing Systems, volume 30. Curran Associates, Inc., 2017.

[74] J. Vig. Visualizing attention in transformer-based language representation models. arXiv:1904.02679, 2019.

[75] T. Wang, X. V. Lin, N. F. Rajani, B. McCann, V. Ordonez, and C. Xiong. Double-hard debias: Tailoring word embeddings for gender bias mitigation. In Association for Computational Linguistics (ACL), July 2020.

[76] K. Webster, M. Recasens, V. Axelrod, and J. Baldridge. Mind the GAP: A balanced corpus of gendered ambiguous pronouns. Transactions of the Association for Computational Linguistics, 6:605–617, 2018.

[77] R. Weischedel, M. Palmer, M. Marcus, E. Hovy, S. Pradhan, L. Ramshaw, N. Xue, A. Taylor, J. Kaufman, M. Franchini, M. El-Bachouti, R. Belvin, and A. Houston. OntoNotes Release 5.0 LDC2013T19. Web Download. Philadelphia: Linguistic Data Consortium, 2013.

[78] J. E. Williams and S. M. Bennett. The definition of sex stereotypes via the adjective check list. Sex Roles, 1:327–337, Dec 1975.

[79] W.-t. Yih and V. Qazvinian. Measuring word relatedness using heterogeneous vector space models. In Proceedings of the 2012 Conference of the North

American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 616–620, Montréal, Canada, June 2012. Association for Computational Linguistics.

[80] B. Zadrozny and C. Elkan. Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers. In Proceedings of the Eighteenth International Conference on Machine Learning, ICML '01, page 609–616, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc.

[81] J. Zhao, T. Wang, M. Yatskar, V. Ordonez, and K.-W. 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), pages 15–20, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[82] J. Zhao, Y. Zhou, Z. Li, W. Wang, and K.-W. Chang. Learning gender-neutral word embeddings. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 4847–4853, Brussels, Belgium, Oct.-Nov. 2018. Association for Computational Linguistics.

Appendix A

Inventory of Research Tools and Data

This section provides a reference to the tools needed to undertake this research. In general, there are four types of resources pertaining to each of static word embeddings (A.1) and pre-trained language models (A.2). Firstly, there are the baseline debiasing methods. We compare our proposed methods against these baselines. Secondly, there are the intrinsic measures of bias (e.g. the direct bias of a word embedding). Thirdly, there are the downstream measures of bias (i.e. how bias is observed in final predictions or outcomes). Both intrinsic and downstream bias assessments are used to evaluate the performance of the debiasing methods. Finally, there are the "vanilla" or traditional assessments of pre-trained resource quality, outside the context of gender bias. These are needed to prove that the debiasing methods have not damaged overall word embedding quality or language modelling ability. Where available, hyperlinks are provided to the corresponding implementation or data resource.

A.1 Tools and resources pertaining to static word embeddings

This section lists the resources pertaining to static word embeddings. Table A.1 lists the baseline debiasing methods. Tables A.2 and A.3 list the intrinsic and downstream evaluations of bias respectively. Table A.4 lists the general word embedding evaluations.

Debiasing methods	
Name	Reference
Hard debias	[7]
Gender-neutral GloVe	[82]
Iterative nullspace projection	[61, 17]
Double hard debias	[75], see theory [51]
Bias alignment model	[39]
Orthogonal subspace correction and rectification	[22]
Gender-preserving debiasing	[35]
Repulse-attract-neutralize debias	[37]

Table A.1: Debiasing methods for static word embeddings.

Intrinsic bias measures		
Name	Type/Note	Reference
Direct bias	Direct	[7]
Relational inner product	Direct	[25]
Indirect bias	Indirect	[7]
Gender-based illicit proximity bias	Indirect; word-level	[37]
Gender-based illicit proximity estimate	Indirect; vocab-level	[37]
Clustering	Indirect; expository	[29]
Recoverability	Indirect; expository	[29]
Word Embedding Association Test	Direct; specific associations	[12]
Analogy generation	Analogy (vector arithmetic)	[7, 25]
Embedding coherence test	Analogy	[23]
Embedding quality test	Analogy retention	[23]
SemBias	Analogy prediction	[82]
Bias analogy test	Analogy	[39]

Table A.2: Intrinsic bias measures for static word embeddings.

Downstream bias test sets		
Name	Type/Note	Reference
WinoBias	coreference resolution; occupations	[81]
Dev et al.'s NLI test	natural language inference; occupations	[21]
SIRT	natural language inference; gender retention	[22]
Equity evaluation corpus	emotion intensity prediction	[36]

Table A.3: Downstream bias assessments using static word embeddings as input.

Embedding quality benchmarks		
Name	Type/Note	Reference
RG	word similarity	[65]
MTurk	word similarity	[79]
MEN	word similarity	[11]
SimLex999	word similarity	[32]
Google	analogy (vector arithmetic)	[46]
MSR	analogy (vector arithmetic)	[48]
Benchmark package	code package to perform all of the above	[34]

Table A.4: Vanilla assessments of word embedding quality.

A.2 Tools and resources pertaining to pretrained language models

This section lists the resources pertaining to language models. Table A.5 lists the baseline debiasing methods. Tables A.6 and A.7 list the intrinsic and downstream evaluations of bias respectively. Note that the downstream evaluations for static word embeddings (Table A.3) can also be used to assess language models. Table A.8 lists the general language model evaluations.

Debiasing methods		
Name	Type/Note	Reference
Bhardwaj et al.'s method	post-processing; analogous to hard debias	[4]
SENT-debias	post-processing; analogous to hard debias	[43]
DEBIASBERT	fine-tuning	[27]

Table A.5: Debiasing methods for pre-trained language models.

Intrinsic measures		
Name	Type/Note	Reference
WEAT for BERT	cosine-based similarity	[38]
SEAT	cosine-based similarity	[44]
Probing classifiers	e.g. gender pronoun resolution	[38]

Table A.6: Intrinsic bias measures for contextual word embeddings and language models.

Downstream bias test sets		
Name	Type/Note	Reference
StereoSet	language modelling task	[52]

Table A.7: Downstream gender bias assessments specific to language models.

Vanilla downstream training resources & test sets		
Name	Type/Note	Reference
OntoNotes 5.0	coreference resolution	[77]
SNLI	natural language inference	[9]

Table A.8: Vanilla language model resources.

Appendix B

Marked Attribute Bias (Chapter 3)

Supplementary Material

B.1 Intrinsic bias measures and correlation results

Intrinsic bias measures of interest on the experimental set of embeddings (Table B.1). There are two base (undebiased) embeddings, word2vec and GloVe. All other embedding spaces are obtained by applying a debiasing method, where each method found here is described in the main text. Implementation notes:

DB and MIDB: The direct bias (DB) and the new multi-dimensional information-weighted direct bias (MIDB) are average measures over a specific (ideally gender-neutral) vocabulary V_t. V_t ($n = 46960$) is defined by taking the 50,000 most frequent words in the common vocabulary between word2vec and GloVe, filtering out punctuation, numbers, and removing the gender-specific word set V_s ($n = 1622$), defined as the union of gender-specific word sets used in previous works [7, 82]. DB is defined as the projection onto a gender direction, here taken to be the $\vec{she} - \vec{he}$ direction. For debiasing methods that promote $\vec{she} \approx \vec{he}$, the DB is not well defined (although it can be computed numerically, it is unstable). We leave these cases as NA rather than a spurious numerical value.

Clustering: The clustering experiment follows [29] in taking the $n \in [500, 1500]$ "most biased" words in the original embedding space (according to their projection on the $\vec{she} - \vec{he}$ axis), and then applying k-means ($k = 2$) clustering on the words in the debiased embedding space. Bias is reported as the either the clustering accuracy

or the v-measure (only $n = 1500$ shown here with v-measure).

Recoverability: Similarly, the dataset ($n = 5000$) is taken to be the most biased words in the original embedding space, where bias labels are assigned according to the projection on the gender direction ($n = 2500$ taken from each class). Several classifiers (SVM with a linear decision boundary, SVM with an RBF kernel, logisitic regression, and a simple fully-connected 1-hidden layer network) were trained on 20% of the dataset with balanced classes. Recoverability bias is reported as the accuracy of classification on the remaining test set (only logisitic regression shown here).

SemBias: The SemBias analogy test set is available from [82]. The set contains $n = 440$ tuples of possible analogies $(\vec{a}, \vec{b})$: 1 definitional analogy (e.g. king, queen), 1 stereotypical analogy (e.g. doctor, nurse), and 2 other analogies (e.g. cup, plate). For every sample, the best analogy is selected as the one to maximize $cos(\vec{he} - \vec{she}, \vec{a} - \vec{b})$. Bias is reported as the proportion of samples to return a definitional analogy, a stereotypical analogy, and an "other" analogy. (Only definitional and stereotypical shown here.)

GIPE: The gender-based illicit proximity bias (GIPE) (see [37] for details) was computed with $n = 100$ nearest neighbours for each word, with an indirect bias threshold of $\theta \in [0.03, 0.05]$ following [37]. (Only $\theta = 0.03$ shown here.)

Full results, plus all code, embedding files, and word sets needed to replicate these results are available at https://github.com/hillary-dawkins/MAB.

Table B.1: Intrinsic bias measures of interest on the experimental set of embeddings.

Emb.method	DB_{vt}	MIDB	Clus:v_{1500}	Rec:LR	SBdef	SBstereo	GIPE:0.03	$\mathcal{E}$
w2v	0.052	0.023	0.933	0.992	0.830	0.134	0.021	0.206
w2v.HD	0.000	0.007	0.440	0.887	0.759	0.114	0.014	0.163
w2v.DHD	NA	0.025	0.271	0.881	0.295	0.373	0.014	0.164
w2v.BAM	0.061	0.038	0.844	0.974	0.814	0.136	0.023	0.131
w2v.OSCaR	0.050	0.024	0.928	0.993	0.830	0.134	0.021	0.188
GloVe	0.055	-0.032	0.984	1.000	0.802	0.109	0.115	0.198
GloVe.HD	0.000	-0.004	0.302	0.927	0.786	0.130	0.070	0.155
GN-GloVe	0.038	0.172	0.588	0.999	0.977	0.014	0.141	0.301
GN-GloVe(w_a)	0.068	-0.096	0.497	0.989	0.939	0.011	0.117	0.149
GloVe.DHD	NA	0.201	0.258	0.903	0.250	0.123	0.064	0.247
GloVe.GP	0.059	0.068	0.996	1.000	0.843	0.080	0.145	0.336
GN-GloVe.GP	0.036	0.006	0.601	0.999	0.984	0.011	0.118	0.179
GloVe.BAM	0.068	-0.019	0.964	0.999	0.775	0.145	0.137	0.175
GloVe.OSCaR	0.056	-0.012	0.984	1.000	0.814	0.102	0.117	0.154
GloVe.RAN	0.044	-0.001	0.419	0.951	0.927	0.011	0.040	0.193
GloVe.INLP	NA	-0.001	0.015	0.660	0.198	0.160	0.080	0.167

Table B.2: Pearson correlation matrix between intrinsic bias measures (and marked attribute error) on the experimental set of embeddings. MIDB obtains the highest correlation with the marked attribute error $\mathcal{E}$; the GIPE was also observed to have a weak correlation. Recoverability bias is most related to the direct bias. The sub-matrix among the SemBias results indicate that trade-off is mostly happening between "definitional" and "other" analogies.

	DB_{vt}	MIDB	Clus:v_{1500}	Rec:LR	SBdef	SBstereo	GIPE:0.03	$\mathcal{E}$
DB_{vt}	1	-0.166	0.694	0.814	0.161	-0.045	0.350	0.104
MIDB		1	-0.145	-0.020	-0.273	0.005	-0.003	0.667
Clus:v_{1500}			1	0.776	0.600	-0.185	0.271	0.184
Rec:LR				1	0.786	-0.390	0.290	0.223
SBdef					1	-0.693	0.304	0.091
SBstereo						1	-0.487	-0.270
GIPE:0.03							1	0.432
$\mathcal{E}$								1

Appendix C

SoWinoBias (Chapter 4) Supplementary Material

C.1 SoWinoBias test set vocabulary

F_{occ} = { writer, teacher, cleaner, tailor, attendant, librarian, auditor, nurse, nanny, cashier, editor, hairdresser, stylist, maid, baker, counselor }

M_{occ} = { guard, architect, chef, leader, president, developer, lawyer, salesperson, doctor, judge, boss, chief, mover, cook, researcher, physician }

F^{+}_{adj} = { sprightly, gentle, affectionate, charming, kindly, beloved, enchanted, virtuous, beauteous, chaste, fair, delightful, lovely, romantic, elegant, fertile }

F^{-}_{adj} = { fussy, nagging, rattlebrained, haughty, whiny, dependent, sullen, unmarried, prudish, fickle, hysterical, infected, widowed, awful, damned, frivolous }

M_{def} = { man, he, father, brother, his, son, uncle, himself }

F_{def} = { woman, she, mother, sister, her, daughter, aunt, herself }

C.2 Results expanded by adjective polarity

Table C.1: Results on SoWinoBias test set by adjective polarity.

Embedding	Data Aug.	Postive Adj.			Negative Adj.			Total		
		pro	anti	diff.	pro	anti	diff.	pro	anti	diff.
GloVe		69.4	49.2	20.1	58.9	44.3	14.6	64.2	46.8	17.4
GloVe	✓	64.2	60.4	3.9	61.4	52.6	8.8	62.8	56.5	6.4
Hard-GloVe		64.6	49.8	14.7	62.6	48.7	13.9	63.6	49.2	14.3
Hard-GloVe	✓	77.2	51.5	25.8	76.9	48.7	28.2	77.1	50.1	27.0
GN-GloVe(w_a)		71.6	52.9	18.6	64.4	46.5	17.9	68.0	49.7	18.3
GN-GloVe(w_a)	✓	71.5	70.5	1.0	72.7	69.0	3.7	72.1	69.7	2.4
RAN-GloVe		70.9	61.5	9.4	69.4	58.5	11.0	70.2	60.0	10.2
RAN-GloVe	✓	73.6	67.0	6.7	65.3	51.9	13.4	69.5	59.4	10.0
INLP-GloVe		74.2	54.0	20.2	62.7	38.2	24.5	68.4	46.1	22.4
INLP-GloVe	✓	76.4	67.9	8.5	70.4	62.3	8.2	73.4	65.1	8.3

Appendix D

Debiasing BERT (Chapter 5) Supplementary Material

Table D.1: Full results for all interventions applied to BERT. Settings refer to hyperparameters $(n_{11}, c_{11}, n_{10}, c_{10}, n_p)$ in order as applicable for each intervention level.

		Intrinsic bias mitigation		Natural language inference			
Intervention	Settings	Strength (5.2) $S \downarrow$	Distance (5.3) $D \downarrow$	Parity $\uparrow$	Accuracy $\uparrow$	Fairness (5.4) $\eta \uparrow$	SNLI Acc. $\uparrow$
Base		0.3069	0.7052	0.0976	0.3840	0.0375	0.8889
Sent-debias		0.3109	0.7014	0.0915	0.4121	0.0377	0.8898
SENT	0	0.3077	0.5972	0.2195	0.5607	0.1231	0.8458
SENT	1	0.3153	0.6470	0.1463	0.4717	0.1231	0.8849
Layer 12	$(0, 0, 0)$	0.2878	0.5406	0.2195	0.6224	0.1366	0.8039
Layer 12	$(0, 0, 1)$	0.3068	0.6237	0.2317	0.5904	0.1368	0.8684
Layer 12	$(0, 1, 0)$	0.3158	0.5420	0.8841	0.1106	0.0978	0.8195
Layer 12	$(0, 1, 1)$	0.3425	0.6364	0.8841	0.1336	0.1181	0.8397
Layer 12	$(1, 0, 0)$	0.2906	0.5424	0.1707	0.6006	0.1025	0.8190
Layer 12	$(1, 0, 1)$	0.3099	0.6246	0.1768	0.5573	0.0985	0.8744
Layer 12	$(1, 1, 0)$	0.2907	0.5318	0.1829	0.5995	0.1097	0.8225
Layer 12	$(1, 1, 1)$	0.3115	0.6180	0.1829	0.5491	0.1004	0.8762
Layer 11	$(0, 0, 0, 0, 0)$	0.2483	0.5228	0.5183	0.7131	0.3696	0.7059
Layer 11	$(0, 0, 0, 0, 1)$	0.2693	0.6207	0.7988	0.7740	0.6183	0.8273

Layer 11	$(0, 0, 0, 1, 0)$	0.2465	0.5184	0.5854	0.7736	0.4528	0.7786
Layer 11	$(0, 0, 0, 1, 1)$	0.2623	0.5856	0.6098	0.7630	0.4653	0.8302
Layer 11	$(0, 0, 1, 0, 0)$	0.2481	0.5255	0.6098	0.7625	0.4649	0.7116
Layer 11	$(0, 0, 1, 0, 1)$	0.2668	0.6219	0.8232	0.7978	0.6567	0.8356
Layer 11	$(0, 0, 1, 1, 0)$	0.2478	0.5244	0.6220	0.7751	0.4821	0.7105
Layer 11	$(0, 0, 1, 1, 1)$	0.2659	0.6155	0.8110	0.8006	0.6493	0.8370
Layer 11	$(0, 1, 0, 0, 0)$	0.2688	0.4535	0.8476	0.8786	0.7447	0.7311
Layer 11	$(0, 1, 0, 0, 1)$	0.2916	0.5243	0.7622	0.8398	0.6401	0.7339
Layer 11	$(0, 1, 0, 1, 0)$	0.3346	0.6331	0.9512	0.9642	0.9171	0.6966
Layer 11	$(0, 1, 0, 1, 1)$	0.3467	0.6715	0.9451	0.9559	0.9035	0.7063
Layer 11	$(0, 1, 1, 0, 0)$	0.2700	0.4578	0.8841	0.8999	0.7956	0.7456
Layer 11	$(0, 1, 1, 0, 1)$	0.2928	0.5333	0.8354	0.8639	0.7217	0.7484
Layer 11	$(0, 1, 1, 1, 0)$	0.2735	0.4486	0.9268	0.9204	0.8530	0.7414
Layer 11	$(0, 1, 1, 1, 1)$	0.3026	0.5324	0.8659	0.8907	0.7712	0.7474
Layer 11	$(1, 0, 0, 0, 0)$	0.2658	0.5392	0.4024	0.6984	0.2810	0.8026
Layer 11	$(1, 0, 0, 0, 1)$	0.2856	0.6261	0.4085	0.6727	0.2748	0.8530
Layer 11	$(1, 0, 0, 1, 0)$	0.2651	0.5344	0.9390	0.0781	0.0733	0.8195
Layer 11	$(1, 0, 0, 1, 1)$	0.2796	0.5975	0.9146	0.1033	0.0945	0.8400
Layer 11	$(1, 0, 1, 0, 0)$	0.2641	0.5372	0.3902	0.6719	0.2622	0.8260
Layer 11	$(1, 0, 1, 0, 1)$	0.2853	0.6198	0.3902	0.6359	0.2482	0.8661
Layer 11	$(1, 0, 1, 1, 0)$	0.2639	0.5362	0.4146	0.6393	0.2651	0.8407
Layer 11	$(1, 0, 1, 1, 1)$	0.2842	0.6152	0.3841	0.6147	0.2362	0.8705
Layer 11	$(1, 1, 0, 0, 0)$	0.2662	0.5198	0.5427	0.7660	0.4157	0.7736
Layer 11	$(1, 1, 0, 0, 1)$	0.2843	0.6069	0.5549	0.7425	0.4120	0.8387
Layer 11	$(1, 1, 0, 1, 0)$	0.3223	0.5623	0.8476	0.1617	0.1370	0.8291
Layer 11	$(1, 1, 0, 1, 1)$	0.3461	0.6620	0.7927	0.2118	0.1679	0.8490
Layer 11	$(1, 1, 1, 0, 0)$	0.2673	0.5223	0.4634	0.7212	0.3342	0.8114
Layer 11	$(1, 1, 1, 0, 1)$	0.2855	0.6080	0.4695	0.6826	0.3205	0.8618

Layer 11	$(1, 1, 1, 1, 0)$	0.2706	0.5059	0.4451	0.7120	0.3169	0.8310
Layer 11	$(1, 1, 1, 1, 1)$	0.2924	0.6013	0.4451	0.6665	0.2967	0.8648
+ attn	$(0, 0, 0, 0, 0)$	0.2200	0.4224	0.3537	0.6317	0.2234	0.7156
+ attn	$(0, 0, 0, 0, 1)$	0.2508	0.5300	0.7256	0.7883	0.5720	0.7859
+ attn	$(0, 0, 0, 1, 0)$	0.2196	0.4208	0.7195	0.7911	0.5692	0.7425
+ attn	$(0, 0, 0, 1, 1)$	0.2359	0.4716	0.8476	0.8271	0.7011	0.7951
+ attn	$(0, 0, 1, 0, 0)$	0.1938	0.3701	0.4878	0.7197	0.3511	0.6706
+ attn	$(0, 0, 1, 0, 1)$	0.2163	0.4483	0.8354	0.8528	0.7124	0.7753
+ attn	$(0, 0, 1, 1, 0)$	0.1934	0.3681	0.5122	0.7397	0.3788	0.6623
+ attn	$(0, 0, 1, 1, 1)$	0.2144	0.4387	0.8476	0.8583	0.7274	0.7751
+ attn	$(0, 1, 0, 0, 0)$	0.3172	0.4949	0.9329	0.9457	0.8822	0.6918
+ attn	$(0, 1, 0, 0, 1)$	0.3315	0.5656	0.8415	0.8805	0.7409	0.6996
+ attn	$(0, 1, 0, 1, 0)$	0.3588	0.5368	0.8720	0.9075	0.7913	0.6836
+ attn	$(0, 1, 0, 1, 1)$	0.3672	0.5709	0.8415	0.8865	0.7459	0.6886
+ attn	$(0, 1, 1, 0, 0)$	0.2891	0.5151	0.9329	0.9438	0.8805	0.6909
+ attn	$(0, 1, 1, 0, 1)$	0.3045	0.6246	0.7866	0.8613	0.6775	0.7037
+ attn	$(0, 1, 1, 1, 0)$	0.2967	0.4985	0.9146	0.9327	0.8531	0.6743
+ attn	$(0, 1, 1, 1, 1)$	0.3161	0.6155	0.7805	0.8594	0.6707	0.6951
+ attn	$(1, 0, 0, 0, 0)$	0.2452	0.4927	0.6280	0.2803	0.1760	0.7737
+ attn	$(1, 0, 0, 0, 1)$	0.2570	0.5615	0.6707	0.5836	0.3914	0.8517
+ attn	$(1, 0, 0, 1, 0)$	0.2465	0.4604	0.8110	0.2713	0.2201	0.7496
+ attn	$(1, 0, 0, 1, 1)$	0.2609	0.5299	0.8110	0.2882	0.2337	0.8074
+ attn	$(1, 0, 1, 0, 0)$	0.2177	0.3707	0.4390	0.5434	0.2386	0.7391
+ attn	$(1, 0, 1, 0, 1)$	0.2518	0.4623	0.6220	0.6606	0.4108	0.8530
+ attn	$(1, 0, 1, 1, 0)$	0.2170	0.3703	0.5122	0.5783	0.2962	0.7569
+ attn	$(1, 0, 1, 1, 1)$	0.2493	0.4568	0.6159	0.6561	0.4041	0.8526
+ attn	$(1, 1, 0, 0, 0)$	0.2526	0.4905	0.1280	0.5243	0.0671	0.6561
+ attn	$(1, 1, 0, 0, 1)$	0.2688	0.5599	0.6402	0.7712	0.4938	0.8481

+ attn	$(1, 1, 0, 1, 0)$	0.2799	0.5102	0.9085	0.1910	0.1735	0.7614
+ attn	$(1, 1, 0, 1, 1)$	0.2868	0.5470	0.8963	0.1984	0.1779	0.8171
+ attn	$(1, 1, 1, 0, 0)$	0.2114	0.3973	0.5122	0.5170	0.2648	0.7568
+ attn	$(1, 1, 1, 0, 1)$	0.2436	0.4953	0.6646	0.7182	0.4774	0.8528
+ attn	$(1, 1, 1, 1, 0)$	0.2137	0.4016	0.7866	0.7467	0.5873	0.7716
+ attn	$(1, 1, 1, 1, 1)$	0.2478	0.5026	0.6951	0.7571	0.5263	0.8497