Identifying Interpretable Word Vector Subspaces With Principal Component Analysis

Identifying Interpretable Word Vector Subspaces with Principal Component Analysis

A thesis submitted by
Jessica Zhao

to
Computer Science
in partial fulfillment of the honors requirement
for the degree of Bachelor of Arts

Harvard College
Cambridge, MA 02138
April 3, 2020

Abstract

Over the past decade, machine learning has become an integral part of our lives by enabling several day-to-day (e.g. product recommendations) as well as critical (e.g. health care treatment recommendations) applications. In particular, the intersection of machine learning and natural language processing (NLP) has been a very active area of research, which played a key role in enabling impactful applications such as question answering systems and personal assistants (e.g. Alexa, Siri). Several NLP tasks rely on learning high-dimensional word vector representations that capture the essence of the underlying textual data and can conveniently be used for downstream prediction tasks. However, such representations may also capture undesirable biases inherent in the text, which in turn can cause catastrophic effects such as discrimination based on protected attributes. Therefore, it is important to identify those subspaces of the vector representations that correspond to protected attributes so that they can be appropriately neutralized via debasing techniques, thus preventing the biases from percolating into critical downstream tasks. While existing research on this topic has leveraged Principal Component Analysis (PCA) to identify certain specific subspaces such as those corresponding to gender, it fails to provide a principled methodology that can easily be generalized to other kinds of subspaces. This thesis develops a novel framework for reasoning about existing PCA based methods, proposes multiple theoretical and experimental criteria for choosing hyper-parameters, and finally presents a novel algorithm that applies PCA more effectively to find a subspace representing any given topic of interest. Experimental evaluation on widely used word vector representations and comparison with prior work demonstrate the efficacy and generalizability of our approach.

Acknowledgements

First, I would like to thank Yiling Chen for introducing me to the topic of interpretable machine learning and for supporting my first project in this subject (and my overall first research project in Computer Science). I would also like to thank my classmate and friend Bill Zhang, in discussion with whom many ideas were formed. Finally, I am indebted to my thesis advisor Himabindu Lakkaraju, without whose class on interpretable machine learning and generous intellectual support this work, in either its form as an ICLR workshop paper or as a thesis, would not exist.

Contents

Chapter 1

Introduction

Methods from machine learning have already proven themselves to be accurate predictors in multiple areas of our lives: they power natural language processing software such as Siri (Kaplan and Haenlein, 2019) and recommend news articles (Beam, 2014, Sood and Kaur, 2014) and products (Orciuoli and Parente, 2017) as well as potential romantic partners (Wobcke et al., 2015) to us, they predict election results (sometimes well, other times poorly) (Gebru et al., 2017, Heredia et al., 2018, Tsai et al., 2018) and the spread of diseases (Bouzillé et al., 2018, Lazer et al., 2014) and assign credit scores to customers (Kulkarni and Dhage, 2019, Shashi et al., 2015).

Furthermore, machine learning has successfully inserted itself in a number of domains that involve high-stakes decisions in which a human decision maker assigns a label or course of action to another person based on the features they present. For example, in health care machine learning is used to predict cesarean delivery and pneumonia mortality (Cooper et al., 2005, Sims et al., 2000), thereby informing the physician's decision regarding the treatment that the patient should receive, and in the judicial system, algorithms are used to predict the outcome of releasing a given suspect on bail (Kleinberg et al., 2017), thereby informing the judge's decision of whether the suspect should be detained or released.

As machine learning is being widely deployed to aid human decision makers in high-stakes decisions, the importance of interpretable models is also rising, aided by regulations who mandate the right to explanation (Wennakoski, 2018). While some researchers remain doubtful of the true value of interpretability, especially in context with its perceived trade-off with accuracy (Lipton, 2016), others point towards the help that explanations can provide for debugging a system, checking a system's safety by ensuring the soundness of the model's decisions, and increasing human decision makers' trust in the algorithm and thus adherence to its output (Doshi-Velez and Been, 2017). More recently, researchers have developed methods for simplifying machine learning methods that were already considered interpretable (e.g. decision trees, lists (Letham

et al., 2015), sets (Lakkaraju et al., 2016)), as well as treating more complicated machine learning models as black boxes and generating post-hoc explanations for their outputs (Ribeiro et al., 2016).

In particular, the field of representation learning has seen many new developments in recent years, such as Google's word2vec and BERT, Facebook's fastText, and Stanford's GloVe, which aim to represent individual words as a high-dimensional vector of numbers (Devlin et al., 2018, Mikolov et al., 2013a, Vaswani et al., 2017, Bojanowski et al., 2016, Pennington et al., 2014). The learned word vector representations can be applied to a variety of real world problems, for example in biomedical Natural Language Processing (bioNLP), where they are used to address tasks such as drug discovery or automated disease diagnosis (Jha et al., 2018). However, word2vec for example is a widespread method in Natural Language Processing for learning vector representation of words, yet its learned representations are often hard to interpret directly (Church, 2017, Turner et al., 2017), therefore making it difficult for users to ascertain the presence of desirable qualities, e.g. fairness with regard to protected attributes such as gender or ethnicity, and presenting the danger of embedding harmful biases into machine learning decisions. More specifically, some existing applications such as ad delivery in Google search have already been shown to discriminate based on protected attributes such as race (Sweeney, 2013).

One approach for remedying this is by finding subspaces in the word embedding space that correspond to topics of interest such as protected attributes, for example via Principal Component Analysis (PCA) (Bolukbasi et al., 2016, Shin et al., 2018). Once revealed, the biases that characterize these subspaces can then be neutralized by debiasing techniques, which prevent them from permeating critical downstream tasks. However, current PCA methods either lack efficiency, requiring inputs that must satisfy a number of conditions (Bolukbasi et al., 2016), or fail to generalize to other word embeddings, being specific to a given training algorithm e.g. SkipGram (Shin et al., 2018). Furthermore, some methods choose hyper-parameters arbitrarily without providing any theoretical or empirical reasons.

1.1 Related Work

A number of methods have been developed and applied to explain word vector subspaces of trained word embeddings. For example, Principal Component Analysis (PCA) (Bolukbasi et al., 2016, Shin et al., 2018) and Kernel PCA (Gupta et al., 2019) identify the subspaces that exhibit the greatest variance when evaluated on a given set of vectors of interest that correspond to the topic the subspace is supposed to represent. Local Interpretable Model-Agnostic Explanations (LIME) induce interpretability in text classification tasks by locally approximating the predictions of any black-box classifier with an interpretable model and assigning importance to the most relevant words (Ribeiro et al., 2016). Other approaches include amending the training framework to induce interpretability at training time, e.g. by including a regularization term in the objective

to induce sparsity(Faruqui et al., 2015). However, these approaches typically do not target specific topics of interest.

In their paper, Shin et al. (2018) draw from Random Matrix Theory to develop a new eigenvector analysis method to interpret word vectors along variance-maximizing dimensions and demonstrate that the most important participants of eigenvectors often form semantically coherent groups. They generate multiple orthogonal interpretable subspaces in tandem, whereas prior work has only considered interpretable subspaces individually. However, many of the principal components they found are still uninterpretable, and a subset of their interpretable principal components do not correspond closely to any topics of interest. Furthermore, the generated principal components change drastically depending on the exact properties of the word embeddings.

1.1.1 Bolukbasi et al. (2016)

In their paper *Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings*, Bolukbasi et al. (2016) examine the publicly available 300-dimensional word2vec embeddings trained on part of the Google News dataset (Mikolov et al., 2013a) and show that they exhibit female-male gender bias: they identify a proposed gender subspace by applying PCA to pairs of female and male words whose meaning only differ in their gender (e.g. {mother, father}) and show that word vectors of many gender neutral words (words whose definition does not include gender) have a substantial component along the gender subspace. For example, the authors find that while "nurse" is not gendered by definition but the corresponding word vector is female-associated.

Bolukbasi et al. (2016) motivate their paper with results from simple analogy tasks on the word embeddings, finding that not only

$$\overrightarrow{\text{man}} - \overrightarrow{\text{woman}} \approx \overrightarrow{\text{king}} - \overrightarrow{\text{queen}} \tag{1.1}$$

holds, but that the following is true as well:

$$\overrightarrow{\text{man}} - \overrightarrow{\text{woman}} \approx \overrightarrow{\text{computer programmer}} - \overrightarrow{\text{homemaker}} \tag{1.2}$$

To identify the gender subspace, the authors picked ten gender word pairs, where the words in each pair possessed identical definitions except in their gender. For example, {mother, father} would be a valid pair since "mother" denotes a female parent and "father" a male parent, whereas {mother, grandfather} would not be since "mother" denotes a female parent whereas "grandfather" denotes a male grandparent. The authors justified their exact choice (see 1.1.1) by arguing that each pair individually constitutes an accurate linear classifier for crowd-sourced words that are considered female or male, either by definition (e.g. waitress, menswear) or because of stereotypes (e.g. sewing, football).

{she, he}	{daughter, son}
{her, his}	{mother, father}
{woman, man}	{gal, guy}
{Mary, John}	{girl, boy}
{herself, himself}	{female, male}

The authors then computed these pairs' principal components and observed that the first principal component explained a significantly larger portion of the total variance that these pairs exhibit than the second, concluding that the first principal component captures the gender subspace well. They provide only limited theoretical analysis for their choice of the first principal component as the gender subspace and present little reasoning for their particular pick of these ten gender pairs as well.

In addition to developing an approach for identifying a topic subspace by the example of gender, the authors also present two debiasing algorithms that use vector arithmetic to remove the gender component entirely or minimize it with respect to constraints from gender neutral words but retain it for gender words to prevent the loss of analogies such as in 1.1 while removing analogies such as in 1.2.

While Bolukbasi et al. (2016) focus on the binary gender topic and debiasing, their work was extended to multi-class subspaces for race and religion by Manzini et al. (2019). Both papers place strict requirements on the input to the PCA and choose their hyper-parameters arbitrarily without providing theoretical or empirical reasoning for their decisions.

1.2 Contributions

This thesis frames and analyzes Bolukbasi et al. (2016), develops and tests multiple criteria for choosing hyper-parameters, examines the effect of relaxing some conditions that the authors propose in a novel algorightm, and shows that their approach generalizes to topics besides gender and other words embeddings.

- This work develops the defining sets framework that allows for framing and critically analyzing the results from Bolukbasi et al. (2016) and other related work.
- This thesis also proposes a principled methodology for choosing hyper-parameters and verifies the choices experimentally. In particular, we propose multiple criteria for choosing the size of a subspace for a topic of interest based on the number of distinct categories in the topic, and for choosing the number of defining sets.
- This thesis develops a novel algorithm that relaxes some of the key assumptions apparent in prior work, which makes it hard to operationalize. In particular, we relax some of the more costly constraints present in prior work and evaluate our approach experimentally.

- We show that both previous work and our approach generalize to topics other than gender and other embeddings by showing that they can be applied to find subspaces to characterize the Good-Bad, Informal-Formal, and Subject-Object binaries. We also show that they apply to other sets of word embeddings, i.e. GloVe.
- This work was also accepted to for presentation at the ICLR 2020 workshop on Machine Learning in Real Life (ML-IRL) (Zhao et al., 2020).

1.3 Overview

From here on, this report is organized as follows:

- Chapter 2 introduces the defining sets framework and the bias model, which allow for formally reasoning about setting hyper-parameters. It also proposes a novel algorithm for choosing defining sets from a larger pool of of potential defining sets, as well as two approaches for generating such a pool.
- Chapter 3 applies that algorithm to find a gender subspace for word2vecNews embeddings and compares the identified subspaces for each approach to Bolukbasi et al. (2016)'s subspace in terms of explained variance and distance.
- Chapter 4 sheds light on how this approach generalizes by finding other topic subspaces (Good-Bad, Informal-Formal, Subject-Object) and comparing them qualitatively as well as extending results to another set of word embeddings (GloVe).
- Finally, chapter 5 concludes this report with a summary of the findings and a list of potential areas for future work.

Chapter 2

The Defining Sets Framework

As discussed in the last chapter, Bolukbasi et al. (2016) provide little justification and analysis for many of their choices, including the number of defining sets, the exact choice of defining sets, and the dimension of the subspace. In order to reason about these limitations more rigorously, this chapter proposes the novel defining sets framework.

The chapter begins by introducing the notation that will be used throughout this report and by providing relevant definitions, which allows for framing and critically analyzing Bolukbasi et al. (2016) and related work. We then develop the bias model, which provides a frame for us to establish conjectures about the relationship between the number of defining sets, the dimension of the subspace, and noise.

This chapter concludes by developing a novel algorithm for finding and choosing defining sets greedily from a larger pool of potential defining sets, and proposing two approaches for generating such a pool from sets of labeled and unlabeled words (with respect to their category) respectively.

2.1 Notation and Definitions

Let W be a vector space defined by some word embeddings (e.g. Google's publicly available pre-trained 300-dimensional word2vec embeddings). Let θ denote the topic of interest.

Definition 1. A *topic of interest* θ refers to a topic or concept captured in the English language, such as gender, race, or religion. We require that it must correspond to some word vector subspace $B^\theta \subset W$, and that it must have $k > 1$ distinct *categories*.

Definition 2. A *category* $\theta_j, 1 \leq j \leq k$ of at topic of interest θ refers to a subconcept or class within θ. For example, for $\theta =$ gender, we could pick $k = 2$ with categories $\theta_1 =$ female, $\theta_2 =$ male.

For the purposes of this report, we will always assume that k is given. Our aim is to identify a subspace $B \approx B^\theta$ for some notion of distance that semantically resembles the topic of interest θ.

Definition 3. Let a *defining set* for topic θ be $D_k(\theta) = \{d_1, \ldots, d_k\}$, where each d_j corresponds to a word from category θ_j. Let us overload $D_k(\theta)$ to denote both the set of k words that make up a defining set and the unique set of corresponding word vectors centered with respect to the full defining set. We write this rigorously as:

$$D_k(\theta) = \{d_1, \ldots, d_k\} = \{\boldsymbol{w}_1 - \boldsymbol{\mu}, \ldots, \boldsymbol{w}_k - \boldsymbol{\mu}\}, \quad \boldsymbol{\mu} = \sum_{j=1}^{k} \frac{\boldsymbol{w}_j}{k}$$

where $\boldsymbol{w}_j \in W$ denotes the unique vector representation of word d_j.

This follows Bolukbasi et al. (2016), though existing work poses the additional requirement that all d_j be equivalent in their definition except in the topic of interest.

For topic θ, a group of n defining sets or simply **defining sets** are denoted as $\{D_k(\theta)\}_n = \{D_k(\theta)^1, ..., D_k(\theta)^n\}$. Then following Bolukbasi et al. (2016)'s approach, we can identify some ℓ-dimensional topic subspace.

Definition 4. A *topic subspace* $\text{PCA}_\ell(\{D_k(\theta)\}_n)$ refers to the subspace spanned by the first ℓ rows of $SVD(C)$, where

$$C = \sum_{i=1}^{n} \sum_{v \in D_k(\theta)^i} v^T v / k$$

Remark. We will refer to $\text{PCA}_\ell(\{D_k(\theta)\}_n)$ as the (identified) topic subspace for topic θ. To distinguish, we refer to B^θ as the true underlying topic subspace.

For a given topic θ, we desire to achieve $\text{PCA}_\ell(\{D_k(\theta)\}_n \approx B^\theta$ for some choice of ℓ, n, and defining sets $\{D_k(\theta)\}_n$. This is nontrivial to ascertain and evaluate as we do not know the true underlying topic subspace B^θ.

We can now frame existing work by Bolukbasi et al. (2016) and Manzini et al. (2019): Bolukbasi et al. (2016)'s gender pairs can be considered a specific instance of defining sets, where the authors investigated $\{D_2(\text{gender})\}_{10}$ with $\ell = 1$ and $W =$ word2vec trained on Google News to find $\text{PCA}_1(\{D_2(\text{gender})\}_{10}) \approx B^{\text{gender}} \subset W =$ word2vecNews. In contrast, Manzini et al. (2019) investigated $\{D_3(\text{religion})\}_5, \ell = 2$ as well as $\{D_3(\text{race})\}_6, \ell = 2$ for $W =$ word2vecRedditL2 (NLC).

2.2 The Bias Model

To formally reason about the defining sets approach, we construct the following model. Let $B_i = \text{PCA}_\ell(\{D_k(\theta)^i\})$ for $1 \leq i \leq n$ be modeled as the span of a set of orthogonal vectors:

$$
\begin{aligned}
B_i &= V_i + U_i \\
V_i &= \text{span}(\{\boldsymbol{v}_1, ..., \boldsymbol{v}_m\}) && \boldsymbol{v}_j \subset B^\theta \\
U_i &= \begin{cases} \{\vec{0}\} & m = \ell \\ \text{span}(\{\boldsymbol{u}_1, ..., \boldsymbol{u}_{\ell-m}\}) & m < \ell \end{cases} && \boldsymbol{u}_j \perp B^\theta
\end{aligned}
$$

where $m \leq \min(\ell, \dim(B^\theta))$ and all vectors $\boldsymbol{v}_j, \boldsymbol{u}_j$ are orthogonal to each other. In particular, we have $V_i \subset B^\theta$ and $U_i \perp B^\theta$. Note that these vectors $\boldsymbol{v}_j, \boldsymbol{u}_j$ need not correspond to the principal components itself.

An ideal defining set would satisfy $B_i = B^\theta$, perfectly capturing the true underlying subspace for the topic of interest. However, biases introduced by the word vectors in $D_k(\theta)^i$ can cause deviations, for example through polysemy (one word possessing multiple meanings). This noise is captured by subspace U_i, which we assume to be independent in its direction and magnitude across defining sets for the same topic.

We aim to identify the optimal number and grouping of defining sets n and the optimal number of principal components ℓ such that $\text{PCA}_\ell(\{D_k(\theta)\}_n) \to B^\theta$. Recall that this is nontrivial as B^θ is unknown and the distance between $\text{PCA}_\ell(\{D_k(\theta)\}_n)$ and B^θ cannot be evaluated directly.

2.3 Setting Hyperparameters

2.3.1 Number of Principal Components ℓ

Bolukbasi et al. (2016) justified their choice of $\ell = 1$ by observing that it explains a significantly larger fraction of variance over all defining pairs than the second principal component. However, this is not generalizable or necessarily optimal: in some cases (e.g. for multiclass topics with high noise) the explained variances may be distributed more equally, such that the choice for ℓ becomes less obvious. We establish this:

Conjecture 2.3.1. *For any $\{D_k(\theta)\}_n, k > 1$, there exists some $\ell \leq k-1$ that globally minimizes the distance between $PCA_\ell(\{D_k(\theta)\}_n)$ and B^θ.*

A set of k vectors, such as those in the defining set $D_k(\theta)^i = \{d_1, \ldots, d_k\}$, can always be fit perfectly by a subspace with dimension of at most $k-1$. Thus, the subspace characterized by the defining set $D_k(\theta)^i$ is of dimension at most $k-1$, since $\ell > k-1$ would suggest that the principal components overfit the subspace with extra dimensions $\boldsymbol{u}_j \in U_i$ that contribute only noise, thereby causing it to deviate from B^θ.

Conjecture 2.3.2. *For any $\{D_k(\theta)\}_n, k > 1$, if all categories θ_j are equidistant from each other, then $\ell = k - 1$ globally minimizes the distance between $PCA_\ell(\{D_k(\theta)\}_n)$ and B^θ.*

This follows from the fact that we cannot represent a structure with k points such that for all points p_i, p_j, the cosine similarity $\cos(\vec{p_i}, \vec{p_j})$ is equal for all $i \neq j$, in a space that has fewer than $k - 1$ dimensions. Since $\ell \leq k - 1$ by Conjecture 2.3.1, this implies that $\ell = k - 1$ should be optimal.

2.3.2 Number of Defining Sets n

Bolukbasi et al. (2016) chose to use $n = 10$ human picked defining sets, where this number was arbitrarily set and no analysis was performed on alternative values of n. Given a large pool of potential defining sets, we need to determine how many and which ones to include. Recall that we assume that noise U_i is independent across defining sets in both magnitude and direction. We establish:

Conjecture 2.3.3. *For any $\{D_k(\theta)\}_n$ with unique subspaces $B_i \neq B_j \ \forall i \neq j$ ordered by increasing noise (magnitude along U_i, U_j), there exists a unique $n^* \leq n$ minimizing the distance between $PCA_\ell(\{D_k(\theta)\}_{n^*})$ and B^θ.*

Note that there is a tradeoff when adding a candidate defining set as an input to PCA: While using a single defining set $D_k(\theta)^i$ and its subspace $B_i = V_i + U_i$ would explain most of the variance it exhibits, it would also capture the noise U_i.

In contrast, additional defining sets would cause the first principal components to more closely capture B^θ, as they would eliminate individual noise U_i since we assume it to be distributed independently across defining sets. However, the fraction of explained variance would decrease since the noise along U_i not captured in the first ℓ principal components would appear in the later principal components instead and since we assume that each additional defining set exhibits larger noise than the previous defining set. With high enough noise, $\boldsymbol{u}$ may even overtake $\boldsymbol{v}$ as the first principal components, thereby causing the identified subspace to diverge from B^θ.

While unaware of any closed-form solutions for n^*, we develop two contrasting approaches for empirically identifying n^* below. This is again non-trivial since we don't know the true magnitude of the noise that individual defining sets exhibit.

2.4 Evaluation Metrics

Below we introduce two evaluation metrics that will be used for evaluating topic subspaces that the algorithm proposes in each iteration, both with respect to the topic's defining sets and their identified subspace.

2.4.1 Explained Variance

Following Bolukbasi et al. (2016), we consider how much of the total variance over all defining pairs an identified topic subspace explains. Note that we can evaluate the collective variance explained of any number of defining sets.

We generally desire a large fraction of explained variance, which would suggest that the identified topic subspace represents the defining sets well. However, this metric is highly dependent on the topic itself: Topics that exhibit more polysemy and stereotypes (i.e. noisy topics) show higher total variance, parts of which are due to noise and should not be captured by the topic subspace. The subspace should only capture the variance due to the topic of interest (i.e. variance along B^θ) which itself may not be very high.

2.4.2 Cosine Similarity

In addition to considering the fraction of variance explained, we propose to monitor the cosine similarities between the principal components of consecutive iterations as well. If the additional defining sets are strengthening the approximation of B^θ, then the principal components of consecutive iterations should have a high of cosine similarity, but if they instead introduce noise to the principal components, then we should observe a drop in cosine similarity, as the identified subspace begins to diverge from B^θ and instead captures the noise U.

2.5 Finding Defining Sets Algorithmically

Bolukbasi et al. (2016) chose their defining gender pairs arbitrarily, requiring that the pairs be equivalent by definition except in the gender dimension, and evaluated them on crowd-sourced gender words. Furthermore, their choice of n^* is fairly arbitrary.

We argue that finding such defining sets is both costly in terms of time and resources and limiting, and we propose a less costly and more principled approach of greedily choosing defining sets $\mathbb{S}$ that maximize explained variance from a larger starting pool of potential defining sets $\mathbb{T} = \{D_k(\theta)\}$. We also present multiple ways to generate that pool $\mathbb{T}$ from simple (labeled) sets of words.

Assume that $n \leq |\mathbb{T}|, \ell$ are given, and that we have chosen no defining sets so far. Then over n iterations, we greedily choose a new defining set in the pool of defining sets that, when added to the already chosen defining sets, results in the largest increase of the variance that the subspace identified by the chosen defining sets explains of the total variance over $\mathbb{T}$. We propose a greedy algorithm for computational reasons, as iterating over the power set is infeasible when $\mathbb{T}$ is large, and to avoid overfitting to the noise U_i in the potential defining sets, where noise could originate from different biases associated with different defining sets or from polysemy.

Algorithm 1 Choosing Defining Sets $\mathbb{S}$ from a Pool of Potential Defining Sets $\mathbb{T}$

function ChooseDefiningSets($n, \mathbb{T}$)
  $\mathbb{S} \leftarrow \emptyset$
  for $i = 1, ...n$ **do**
    $D_k(\theta)^* \leftarrow \arg\max_{D_k(\theta) \in \mathbb{T} \setminus \mathbb{S}}$ VarExp($\mathbb{S} \cup \{D_k(\theta)\}, \mathbb{T}$)
    $\mathbb{S} \leftarrow \mathbb{S} \cup \{D_k(\theta)^*\}$
  end for
return $\mathbb{S}$
end function

where VarExp($\mathbb{S}, \mathbb{T}$) returns the fraction of the variance in the defining sets in $\mathbb{T}$ that the subspace $\text{PCA}_\ell(\mathbb{S})$ generated from the defining sets in $\mathbb{S}$ explains. Note that we assume n is known. Under this framing, Bolukbasi et al. (2016)'s approach can be viewed as setting $k = 2, |\mathbb{T}| = n = 10, \ell = 1$, where defining pairs $D_2(\text{gender}) \in \mathbb{T}$ are definitionally equivalent except for the gender dimension and we are choosing all sets in $\mathbb{T}$ such that $\mathbb{S} = \mathbb{T}$.

2.6 Creating a pool of potential defining sets $\mathbb{T}$

We propose two alternatives for generating the pool of defining sets $\mathbb{T}$ without the need for crowd-workers or definitional equivalence:

2.6.1 Category Approach

We relax the definitional equivalency constraint on elements of a given defining set, instead starting from a set of words W_{θ_j} for each category θ_j in the topic of interest θ. For example, let $\theta = \text{gender}$ let $\theta_1 = \text{female}$, $\theta_2 = \text{male}$. Then we have a set of female W_{female} and a set of male words W_{male}. We include all combinations of one element from each set, such that we have

$$\mathbb{T} = \{\{w_1, ...w_k\},\ w_j \in W_{\theta_j}\}$$

Then we conjecture that our proposed algorithm 1 will choose approximately equivalent pairs if they exist in order to maximize explained variance.

2.6.2 Direction Approach

We can further relax the condition that we must know which category a word belongs to, instead starting from a set of topical words W_θ (e.g. a set of gender words) and an approximate subspace B' only.

In particular, first we consider all size-k elements of the power set of W_θ. However, we would like to exclude within-category sets such as {grandmother, daughter} for gender. To do this, we require a rough approximation or *direction* B' of the true subspace B^θ, and we exclude all sets D_k from $\mathbb{T}$ that are too orthogonal to B' based on a threshold $s_{\max}$ on some measure of distance, e.g. negative absolute cosine similarity when $\ell = 1$, between the subspace B_{D_k} that is defined by D_k and B'. We can formalize this as

$$\mathbb{T} = \{D_k \in P(W_\theta),\ |D_k| = k,\ |dist(B_{D_k}, B')| < s_{\max}\}$$

2.7 Finding the Optimal Number of Defining Sets n^*

We propose two alternate approaches for finding a good number of defining sets n^*:

2.7.1 Explained Variance and Change in the Subspace

We propose that a decrease in the fraction of variance explained after adding the ith defining set to $\mathbb{S}$ simultaneous with a relatively large change (compared to the change when adding the $i-1$th and $i+1$th defining set) in the identified subspace suggests that the subspace is beginning to capture noisy dimensions in the defining sets (e.g. age for $\theta =$ gender), which decreases explained variance and manifests in an increased change in the subspace $\text{PCA}_\ell(\{D_k(\theta)\}_n$.

2.7.2 K-Fold Cross-Validation

Considering each $D_k(\theta) \in \mathbb{T}$ a data point, we can split $\mathbb{T}$ into K training and validation folds and perform cross-validation to find the n^* that maximizes explained variance over the defining sets in the validation set. We conjecture that this will combat overfitting to individual noise U_i that defining sets may exhibit, thereby allowing us to better capture B^θ instead. One drawback of this method is that it will not be able to find $n^* > \frac{K-1}{K} * |\mathbb{T}|$, and in particular, options such as Bolukbasi et al. (2016)'s $n = |\mathbb{T}|$ would be excluded.

Chapter 3

Experimental Results with word2vec on Gender Subspace

This chapter first motivates our novel approach by analyzing the experimental results from Bolukbasi et al. (2016). It then applies our algorithm that we developed in the previous chapter to identify a gender subspace for word2vecNews embeddings and compares the identified subspace under each approach to Bolukbasi et al. (2016)'s subspace.

3.1 Analyzing results from Bolukbasi et al. (2016)

We first present the relevant subset of Bolukbasi et al. (2016)'s results, see Figure 3.1.

Results are slightly different as we had a different initial word vector space: Bolukbasi et al. (2016) use the most frequent 50,000 words, but don't state explicity what their measure for frequency was. In contrast, we simply used the top 50,000 words of the GoogleNews word2vec embeddings.

We also found that the cosine similarity of the defining pairs to the identified subspace B^1 ranges from 0.352 for {female, male} to 0.931 for {herself, himself}. Furthermore, recall that the subspace explains about 60% of the total variance over all defining pairs. Similarly, the amount of variance that a single pair (the vector difference between the two words' embeddings) explains ranges from 12.2% for {female, male} to 54.8% for {herself, himself} (see Table 3.1).

To summarize, there are a few defining pairs, e.g.{herself, himself}, {she, he} , that are very close to the identified subspace in cosine similarity and also explain a large portion of the variance across defining sets by themselves. Other pairs such as {female, male}, however, explain only very little variance across defining pairs and are very dissimilar, which begs the question of whether they should be included as a defining pair at all.

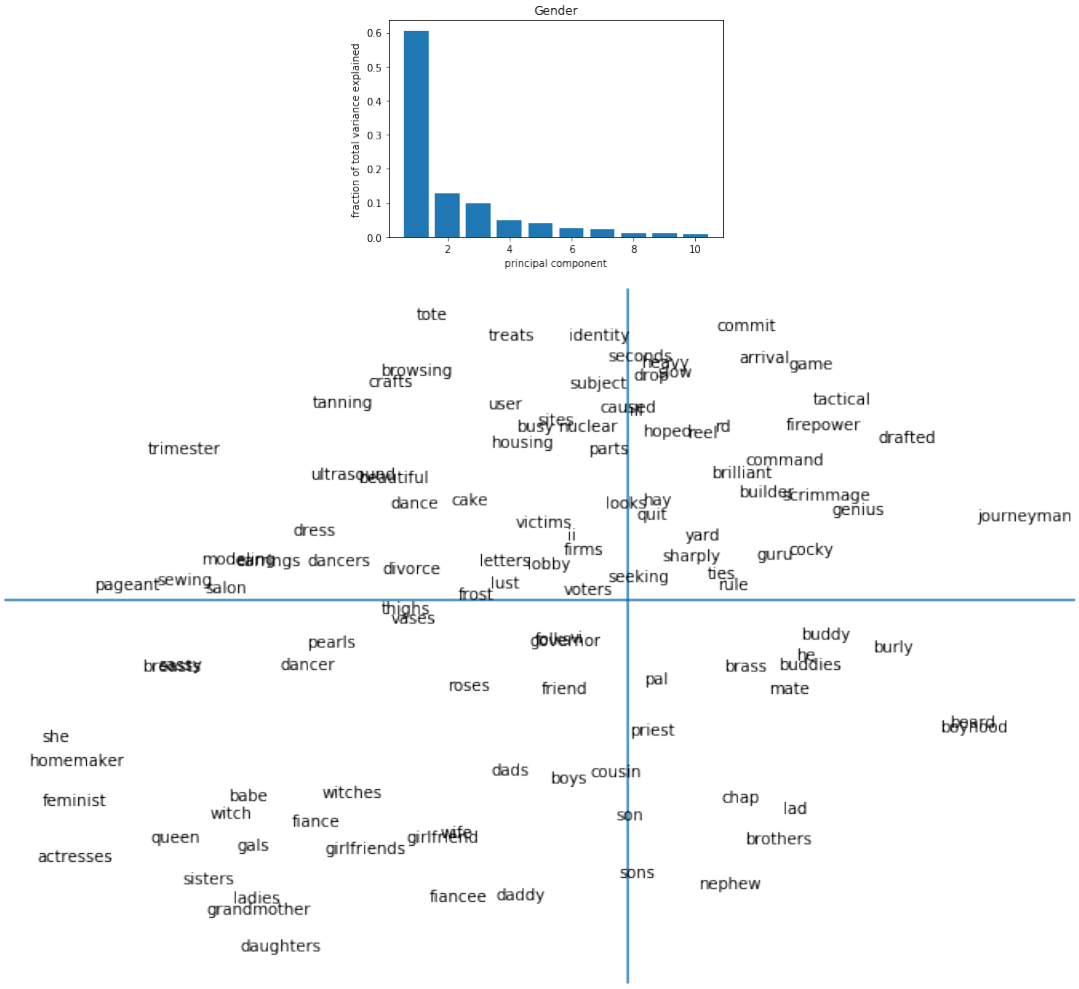

Figure 3.1: Top: Fraction of explained variance of the 10 principal components.

3.2 Finding Defining Sets Algorithmically

We used the female and male words in Bolukbasi et al. (2016)'s defining pairs as our seed words for generating $\mathbb{T}$. In particular, $\theta = \text{gender}, \theta_1 = \text{female}, \theta_2 = \text{male}$. Then $W_{\text{female}} = \{$she, her, woman, Mary, herself, daughter, mother, gal, girl, female$\}$, $W_{\text{male}} = \{$he, his, man, John, himself, son, father, guy, boy, male$\}$, and $W_{\text{gender}} = W_{\text{female}} \cup W_{\text{male}}$.

3.2.1 Category Approach

We have $\mathbb{T} = \{\{w_1, w_2\},\ w_1 \in W_{\text{female}}, w_2 \in W_{\text{male}}\}$. Recall that the pool of potential defining sets is given by

$$\mathbb{T} = \{\{w_1, w_2\},\ w_1 \in W_{\text{female}}, w_2 \in W_{\text{male}}\}$$

Defining Pair	**Explained Variance (%)**	**|Cosine Similarity to B|**
('she', 'he')	54.416	0.92
('her', 'his')	51.353	0.894
('woman', 'man')	41.992	0.812
('Mary', 'John')	27.841	0.536
('herself', 'himself')	54.817	0.931
('daughter', 'son')	36.05	0.775
('mother', 'father')	31.614	0.728
('gal', 'guy')	35.687	0.685
('girl', 'boy')	35.131	0.751
('female', 'male')	12.156	0.352
mean	38.106	0.738
standard deviation	12.552	0.172

Table 3.1: How much variance each individual defining pair explains, and how close it is to the identified gender subspace B (in absolute cosine similarity).

such that $|\mathbb{T}| = |W_{\text{female}}| * W_{\text{male}}| = 10 * 10 = 100$.

Note that with our algorithm for choosing defining sets, we actually choose Bolukbasi et al. (2016)'s defining pairs as the first 10 pairs, see Table 3.2. Note also that adding an additional eleventh defining pairs decreases the explained variance for the first time and that there is a large change in the identified subspace before the amount of explained variance rises again as we add more defining sets (see Figure 3.2) and therefore pick up on other relevant dimensions such as age with defining pairs {daughter, father}, {mother, son} (Table 3.2).

Note also that our greedy algorithm picks different pairs than simply picking the defining pair left in $\mathbb{T}$ that explains the most variance individually (see the green curve occasionally rising in Figure 3.2). While the amount of variance in $\mathbb{T}$ explained falls significantly from 0.6 to 0.242, this is to be expected as there is more variance within $\mathbb{T}$: Bolukbasi et al. (2016)'s defining pairs show a total variance of 0.149, whereas our constructed $\mathbb{T}$ shows a total variance of 0.361. Consider also the difference between the red and black curves in Figure 3.2, which denote the variance explained in $\mathbb{T}$ and variance explained in Bolukbasi et al. (2016)'s defining pairs, and note that the latter is still comparably high.

3.2.2 Direction Approach

Here we have $W_{\text{gender}} = W_{\text{female}} \cup W_{\text{male}}$, such that $|\mathbb{T}'| = \binom{|W_{\text{gender}}|}{k} = \binom{20}{2} = 190$. Now, let our approximate subspace B' be the vector difference between the embeddings of *she* and *he*, let our distance metric be the negative absolute cosine similarity, and let $s_{\max} = 0.45$, where $s_{\max}$ was set to the value that resulted in highest explained variance under the condition that $s_{\max} \geq 0.4$

to avoid overfitting. We also excluded $s_{\max}$ values that would result in $\mathbb{T}$ being too small, e.g. $|\mathbb{T}| < 3$.

Note that within the first 12 chosen pairs, we recover 8 of the 10 pairs. However, we also choose pairs such as {mother, he} or {woman, he}, see Table 3.2). Note that we have a total variance in $\mathbb{T}$ of 0.279, which is significantly higher than Bolukbasi et al. (2016)'s total variance but lower than the category approach, as we are excluding all within-gender pairs and most between-gender pairs, resulting in a total of $23 = |\mathbb{T}|$ potential defining pairs.

We can see that under the given settings, this approach does less well than the categories approach, resulting in lower maximum and minimum explained variance over Bolukbasi et al. (2016)'s defining pairs, though other settings may achieve higher explained variance.

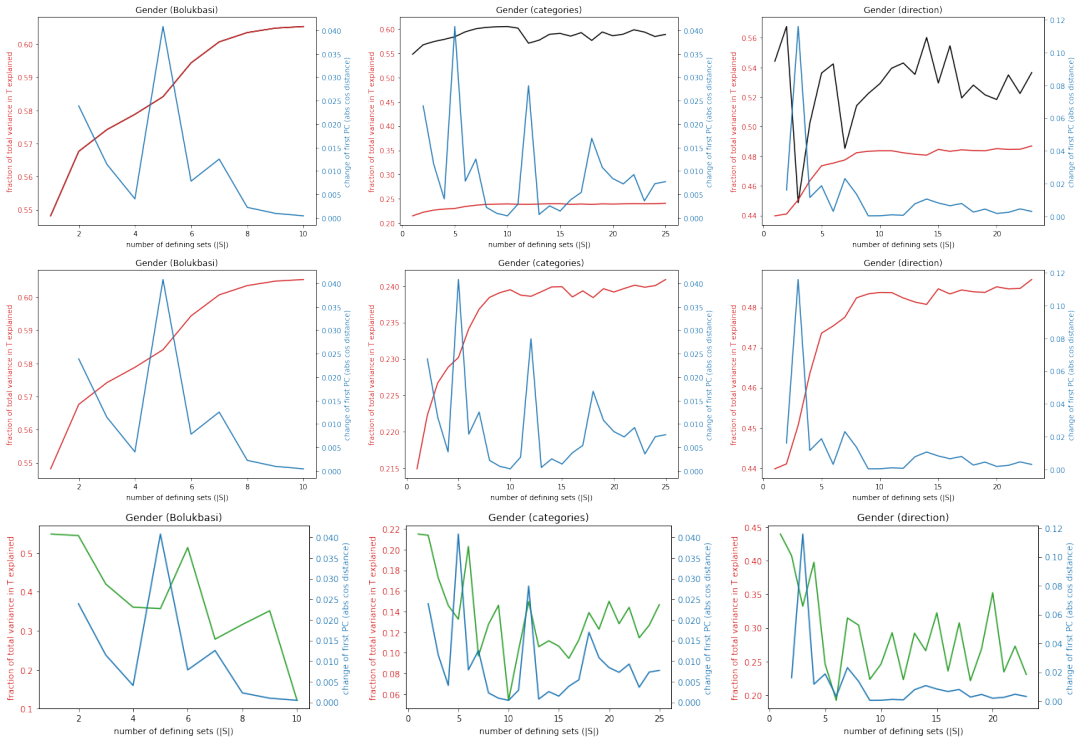

Figure 3.2: Fraction of total variance of $\mathbb{T}$ explained by all chosen defining pairs (red), fraction of total variance in Bolukbasi's 10 defining pairs explained by all chosen defining pairs (black), fraction of total variance of $\mathbb{T}$ explained by the single defining pair last added (green), and the change of the first principal component after adding the nth defining pair (as measured in absolute cosine distance, blue). B' = he - she, $s_{\max} = 0.45$.

	Gender (Bolukbasi)	Gender (categories)	Gender (direction)
1	('herself', 'himself')	('herself', 'himself')	('she', 'he')
2	('she', 'he')	('she', 'he')	('herself', 'himself')
3	('woman', 'man')	('woman', 'man')	('woman', 'he')
4	('daughter', 'son')	('daughter', 'son')	('her', 'his')
5	('gal', 'guy')	('gal', 'guy')	('gal', 'guy')
6	('her', 'his')	('her', 'his')	('daughter', 'father')
7	('Mary', 'John')	('Mary', 'John')	('mother', 'he')
8	('mother', 'father')	('mother', 'father')	('her', 'himself')
9	('girl', 'boy')	('girl', 'boy')	('girl', 'boy')
10	('female', 'male')	('female', 'male')	('daughter', 'son')
11	-	('daughter', 'father')	('woman', 'man')
12	-	('female', 'himself')	('mother', 'father')
13	-	('girl', 'male')	('herself', 'he')
14	-	('herself', 'male')	('she', 'himself')
15	-	('mother', 'son')	('girl', 'he')
16	-	('woman', 'boy')	('she', 'guy')
17	-	('girl', 'man')	('daughter', 'he')
18	-	('woman', 'John')	('herself', 'his')
19	-	('Mary', 'man')	('woman', 'himself')
20	-	('mother', 'he')	('her', 'he')

Table 3.2: Top 20 gender pairs as chosen under each of the approaches.

3.3 Finding the Optimal Number of Defining Sets n^*

3.3.1 Explained Variance and Change in the Subspace

Choosing n^* as the first i after which adding a defining pair decreases explained variance and increases the cosine distance of the identified subspace to the previous subspace, we choose $n^* = 10, 10, 11$ for $\mathbb{T}$ = Bolukbasi et al. (2016)'s defining pairs, the categories approach, and the direction approach respectively.

3.3.2 K-Fold Cross-Validation

In contrast, 10-fold cross-validation chooses $n^* = 2, 2, 18$ for $\mathbb{T}$ = Bolukbasi et al. (2016)'s defining pairs, the categories approach, and the direction approach respectively, though the explained variance on the validation set doesn't vary strongly with n under the first and last approach, see Figure 3.3.

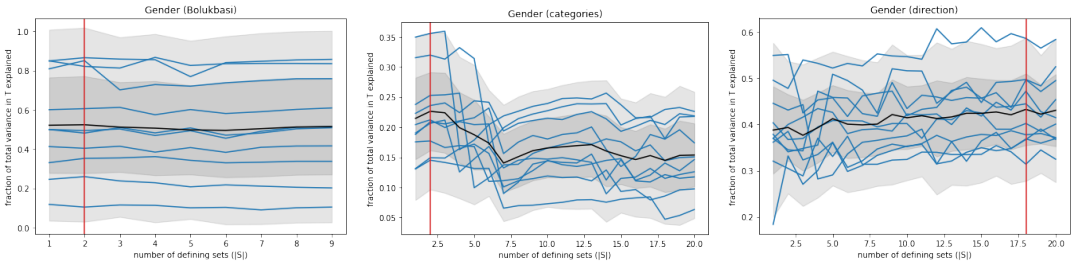

Figure 3.3: 10-fold cross-validation for finding n^*. Blue curves denote individual folds, the black curve the mean, and the shaded intervals denote one and two standard deviations. The red line marks the optimal number of defining pairs according to cross-validation.

3.3.3 Comparison

Note that choosing based on explained variance and change of the identified subspace results in larger explained variance on Bolukbasi et al. (2016)'s ten defining pairs (see Figure 3.2), and a subspace that is closer as measured in cosine distance to the gender subspace that Bolukbasi et al. (2016) identified as well (see Table 3.3).

	Gender (Bolukbasi)	**Gender (categories)**	**Gender (direction)**
EV and change	1.000	1.000	0.9428
CV	0.9657	0.9657	0.9327

Table 3.3: Cosine similarity of the identified subspaces with Bolukbasi et al. (2016)'s subspace, where n is chosen from explained variance and the identified subspace or cross-validation.

3.4 Noisy Seed Words

We was unable to generalize either the categories or direction approach well to noisy seed words, for which we randomly sampled an equal number (5, 10, 20) of female and male words from Bolukbasi et al. (2016)'s total 218 gender words. we hypothesize that this is due to gender no longer being the sole dominating dimension, but rather, occupation (e.g. housewives, chairman), age (e.g. mothers, lads), species (e.g. deer, gelding), ... are captured in the defining pairs as well (see Table 3.4 for one example).

Future work could examine the relationship between the ratio of relatively pure pairs such as {grandmother, grandfather}, and relatively impure pairs such as {filly, estranged_husband}, and the identified subspace.

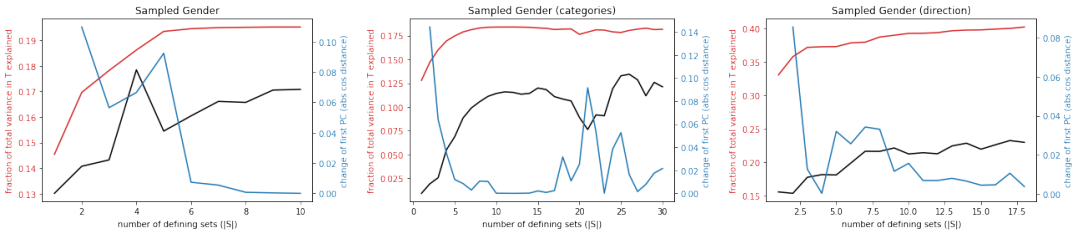

Figure 3.4: Same as first row of Figure 3.2 for sampled gender words, where B' is defined by {she, he} and $s_{\max} = 0.7$. Note that the first image is uninformative (essentially random defining pairs), and that the black line is with respect to Bolukbasi et al. (2016)'s defining pairs.

	Sampled Gender	**Sampled Gender (categories)**	**Sampled Gender (direction)**
1	('housewives', 'guy')	('fiancee', 'dudes')	('mothers', 'guy')
2	('mothers', 'dudes')	('stepmother', 'salesmen')	('stepmother', 'guy')
3	('bride', 'fraternities')	('bride', 'fraternities')	('goddess', 'chairman')
4	('goddess', 'chairman')	('granddaughter', 'guy')	('filly', 'gelding')
5	('fiancee', 'prostate')	('goddess', 'prostate')	('bride', 'guy')
6	('hens', 'salesmen')	('mothers', 'chairman')	('mothers', 'prostate')
7	('deer', 'lads')	('housewives', 'lads')	('granddaughter', 'chairman')
8	('granddaughter', EH)	('hens', 'lads')	('housewives', 'guy')
9	('filly', 'gelding')	('deer', 'prostate')	('mothers', 'salesmen')
10	('stepmother', 'fiance')	('filly', 'gelding')	(EH, 'guy')
11	-	('housewives', 'gelding')	('mothers', 'lads')
12	-	('bride', 'fiance')	('granddaughter', 'guy')
13	-	('fiancee', 'fiance')	('housewives', 'chairman')
14	-	('granddaughter', 'fiance')	('filly', 'guy')
15	-	('bride', EH)	('mothers', 'dudes')
16	-	('stepmother', 'fiance')	('goddess', 'guy')
17	-	('fiancee', EH)	('mothers', 'chairman')
18	-	('filly', EH)	('fiance', 'guy')
19	-	('stepmother', EH)	-
20	-	('deer', 'fraternities')	-

Table 3.4: Top 20 sampled gender pairs as chosen under each of the approaches. Sampled Gender column itself is uninformative, but does present all twenty chosen gender words. EH stands for *estranged_husband*.

Chapter 4

Generalizing to Other Topics and Embeddings

In the previous chapter we presented our findings for the gender subspace and the word2vec embedding space. In this chapter, we first extend these results to other topic subspaces (Good-Bad, Informal-Formal, Subject-Object) and compare them qualitatively. We also extend the results to another set of word embeddings (GloVe).

4.1 Generalizing to Other Binaries

We generalize Bolukbasi et al. (2016)'s and our results to some other topics with defining sets $\{D_2(\text{Good-Bad})\}_6$, $\{D_2(\text{Informal-Formal})\}_1 4$, and $\{D_2(\text{Subject-Object})\}_5$ as well, where defining sets are crowd-sourced from a minimal crowd of size two. See Table 4.1 for the crowd-sourced defining sets.

Performing PCA on the defining pairs, we can see that for all three binaries, the first principal component explains a large amount of variance compared to the second principal component, see Figure 4.1. Variance explained by the first principal component ranges from 0.242 for Informal-Formal to 0.649 for Subject-Object, and we hypothesize that this can be ascribed to the amount of noise in the defining pairs. For more details see Table A.1.

For each of the following sections we generate the pool of defining sets analogously to the gender topic, using the defining sets in Table 4.1 instead of Bolukbasi et al. (2016)'s ten gender pairs.

	Good-Bad	**Informal-Formal**	**Subject-Object**
1	('good', 'bad')	('seem', 'appear')	('he', 'him')
2	('optimistic', 'pessimistic')	('get', 'obtain')	('she', 'her')
3	('positive', 'negative')	('show', 'demonstrate')	('they', 'them')
4	('great', 'terrible')	('start', 'commence')	('I', 'me')
5	('happy', 'sad')	('keep', 'retain')	('we', 'us')
6	('able', 'unable')	('let', 'permit')	-
7	-	('but', 'however')	-
8	-	('so', 'therefore')	-
9	-	('good', 'positive')	-
10	-	('bad', 'negative')	-
11	-	('right', 'correct')	-
12	-	('wrong', 'incorrect')	-
13	-	('smart', 'intelligent')	-
14	-	('cheap', 'inexpensive')	-

Table 4.1: Defining sets for $\{D_2(\text{Good-Bad})\}_6$, $\{D_2(\text{colloquial-formal})\}_14$, and $\{D_2(\text{Subject-Object})\}_5$,.

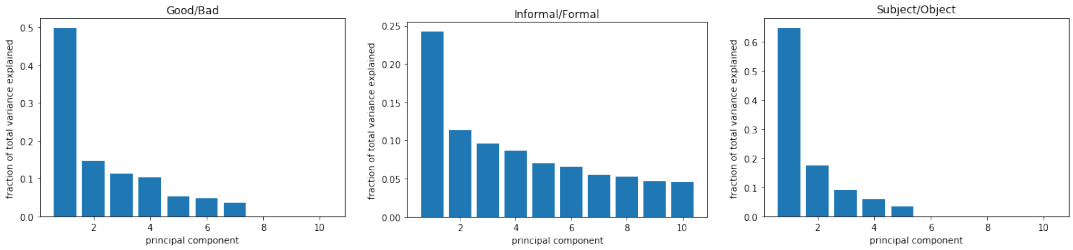

Figure 4.1: Fraction of total variance in the defining sets that is explained by the principal components. See 3.1 to see equivalent plots for gender.

4.1.1 Good-Bad

The approximate subspace B' is given by {good, bad} and $s_{\max} = 0.5$. We picked s_max picked to maximize explained variance, requiring $s_{\max} \geq 0.4$ (same as before).

Under neither of our approaches do we recover the original defining sets, see Table 4.2. We hypothesize that this is due to increased noise U_i as well as stereotypes in the words and increased ambiguity regarding definitional equivalence and its manifestation in spoken and written language. For example, *good* could be used as the opposite of *terrible* as well as *bad*. This hypothesis is corroborated by the strong fluctuations of the variance that single defining pairs explain, as other noisy dimensions besides Good-Bad are being captured.

Furthermore, there may be harmful stereotypes that lead to pairs such as {able, negative} or {good, unable}, where being less able is viewed negatively by society. There may also be other value judgements that result in pairs such as {happy, bad}, where being unhappy or sad is again negatively connotated socially.

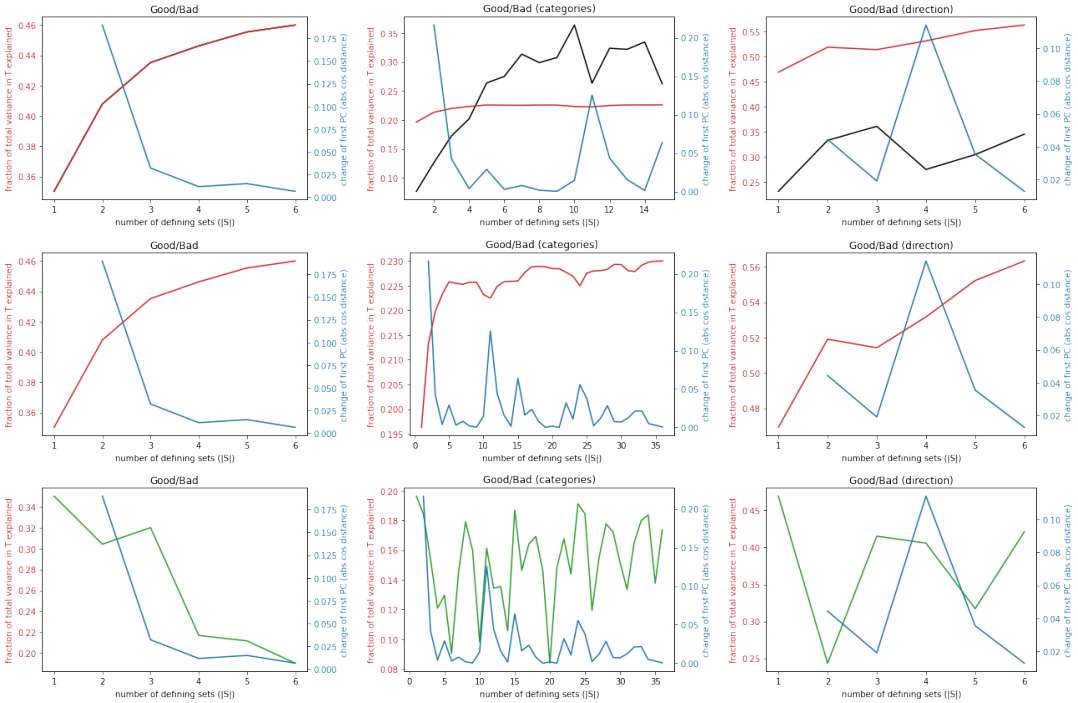

Figure 4.2: Same as Figure 3.2 for Good-Bad, where B' is defined by {good, bad} and $s_{\max} = 0.5$.

	Good-Bad	**Good-Bad (categories)**	**Good-Bad (direction)**
1	('great', 'terrible')	('able', 'negative')	('great', 'bad')
2	('happy', 'sad')	('great', 'pessimistic')	('good', 'terrible')
3	('good', 'bad')	('happy', 'bad')	('good', 'bad')
4	('positive', 'negative')	('good', 'terrible')	('great', 'negative')
5	('able', 'unable')	('happy', 'sad')	('happy', 'bad')
6	('optimistic', 'pessimistic')	('optimistic', 'pessimistic')	('great', 'terrible')
7	-	('good', 'bad')	-
8	-	('good', 'unable')	-
9	-	('positive', 'terrible')	-
10	-	('positive', 'negative')	-

Table 4.2: Top 10 Good-Bad pairs as chosen under each of the approaches.

4.1.2 Informal-Formal

Under neither of our approaches do we recover the original defining sets, see Table 4.3. Similarly to the Good-Bad binary, we hypothesize that this is due to increased noise in the words as well as more ambiguity regarding definitional equivalence as manifested in the training text corpus.

Note that under Bolukbasi et al. (2016)'s approach, the greedy algorithm actually picks defining pairs roughly by the amount of variance they explain by themselves - while other topics show a decreasing trend as well, that trend is mostly monotone here with only two exceptions.

Notice that for the categories approach, we eventually recover a subspace that explains most of the variance that the original defining sets explain (see black curve in Figure 4.3). However, it takes a comparatively large number of defining sets (around 30) as compared to the 14 original defining sets, likely due to the increased noise.

This is also one example where our explained variance and cosine similarity approach of finding n^* fails, as the red and black curve (variance explained on $\mathbb{T}$ and on the crowd-sourced defining sets) diverge at times.

Note also that for the direction approach, there are comparatively few potential defining pairs that are similar to B' and that the words in each defining pair they are all topically related: *so, appear, however, seem.* Again, this is likely a manifestation of the noise in the crowd-sourced defining pairs and potentially the topic itself.

	Informal-Formal	**Informal-Formal (categories)**	**Informal-Formal (direction)**
1	('bad', 'negative')	('bad', 'obtain')	('so', 'appear')
2	('get', 'obtain')	('wrong', 'commence')	('appear', 'however')
3	('so', 'therefore')	('good', 'demonstrate')	('seem', 'appear')
4	('start', 'commence')	('right', 'retain')	-
5	('let', 'permit')	('smart', 'permit')	-
6	('good', 'positive')	('let', 'appear')	-
7	('right', 'correct')	('so', 'therefore')	-
8	('keep', 'retain')	('get', 'therefore')	-
9	('wrong', 'incorrect')	('cheap', 'however')	-
10	('show', 'demonstrate')	('let', 'positive')	-
11	('but', 'however')	('cheap', 'negative')	-
12	('seem', 'appear')	('but', 'negative')	-
13	('smart', 'intelligent')	('so', 'incorrect')	-
14	('cheap', 'inexpensive')	('but', 'inexpensive')	-
15	-	('start', 'permit')	-

Table 4.3: Top 15 Informal-Formal pairs as chosen under each of the approaches.

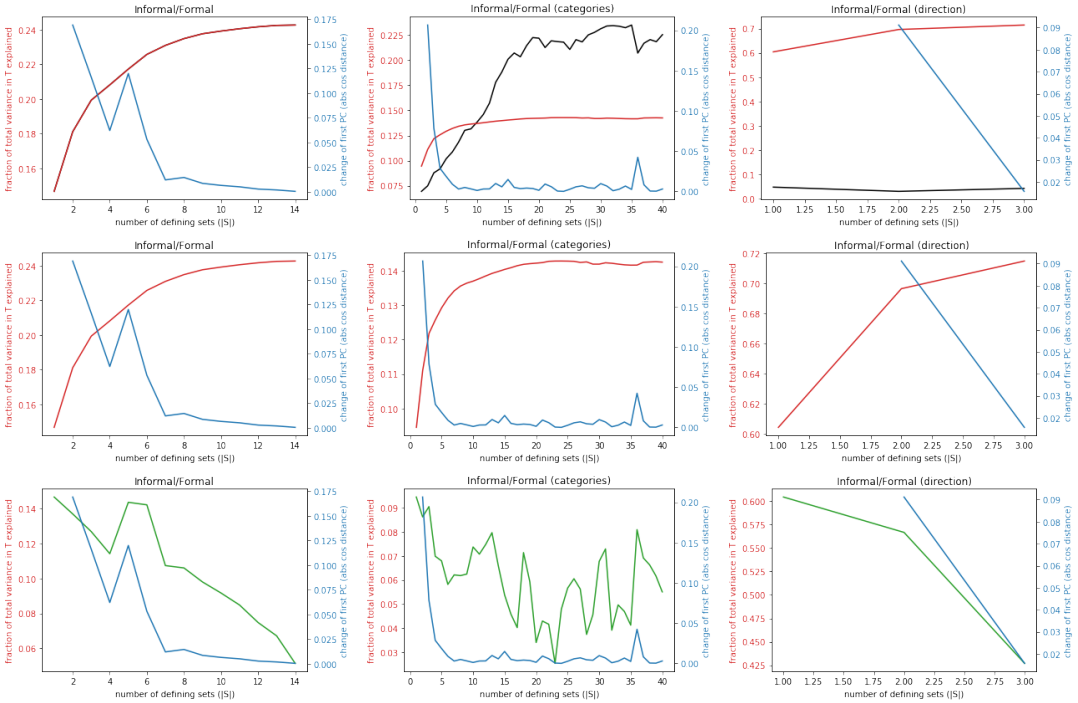

Figure 4.3: Same as Figure 3.2 for Informal-Formal, where B' is defined by {seem, appear} and $s_{\max} = 0.5$.

4.1.3 Subject-Object

We can see that under the categories approach, we perfectly recover the original defining pairs, and that under the direction approach we recover all but {she, her}, see Table 4.4. We hypothesize that these results are due to the clear definitions for each of the words, especially in relation to each other.

The direction is given by {we, us} and $s_{\max} = 0.6$ (compare to gender, where a smaller value was favored). In particular, there exists no value for $s_{\max}$ for which {she, her} was selected as one of the first five defining pairs for B' defined by {we, us}. Furthermore, note that under the given settings it was excluded, being too dissimilar to {we, us}.

Note that choosing based on the cosine similarity of the subspace and the explained variance would give us the optimal subspace for each of the approaches, given the potential defining pairs that are in the pool, and would maximize explained variance both over $\mathbb{T}$ and over the crowd-sourced defining pairs.

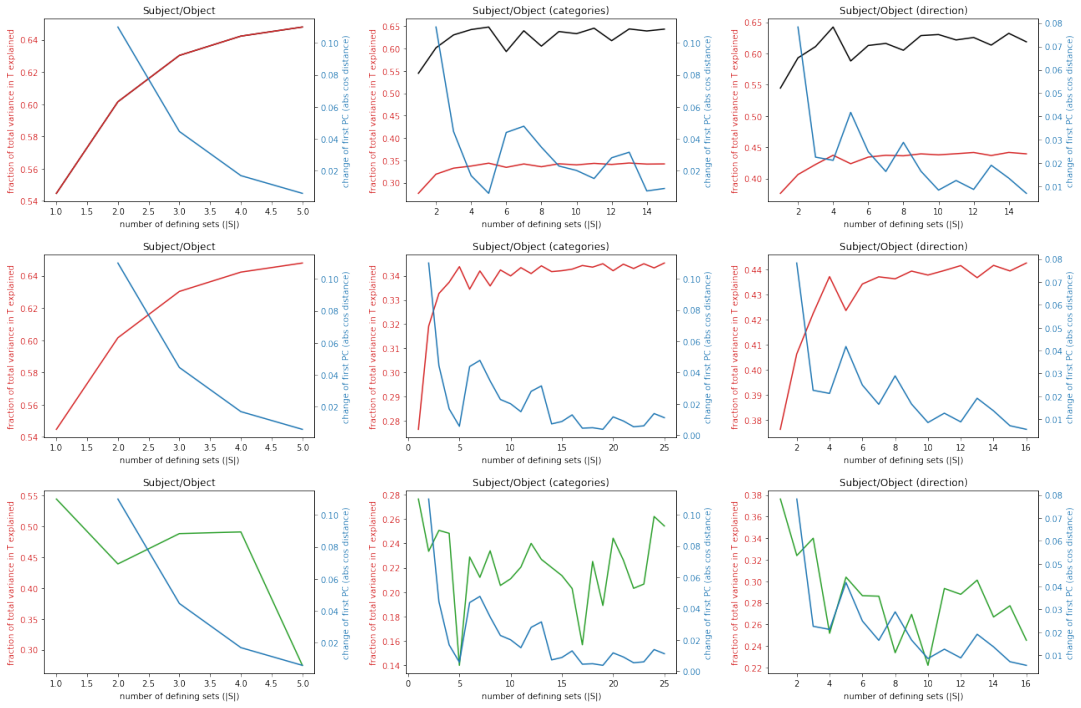

Figure 4.4: Same as Figure 3.2 for Subject-Object, where B' is defined by {we, us} and $s_{\max} = 0.6$. Compare to Figure 3.2 for gender.

4.1.4 Comparing Binaries

Note that out of the three non-gender binaries, the Subject-Object binary shows the results most similar to the gender binary both on the PCA plots and our greedy approach, and we hypothesize that this is due to strong definitional equivalence between the original defining pairs, which lead to a relatively unbiased identified topic subspace.

Conversely, definitional equivalence is lacking in the Good-Bad and Informal-Formal binaries as many words are connotated with stereotypes and or biases, and therefore the subspace given by the crowd-sourced defining pairs may already be a suboptimal approximation of the true subspace B^{θ}. It makes sense that we cannot recover good (definitionally equivalent) defining pairs from a larger pool of defining pairs when that pool itself doesn't contain any, though it may be possible to actually recover better pairs than the crowd-sourced pairs.

Furthermore, we found that the identified topic subspaces are largely orthogonal to each other, showing cosine similarities between 0.0001 and 0.0875 (see Table 4.5). This suggests that it may be possible to reorient the word vectors along multiple interpretable subspaces that represent topics of interest while maintaining the dimensionality of the original word embedding space with only minimal loss in captured variance. In particular, if the subspaces were perfectly orthogonal, there would be no loss. Further exploration of this option is left to future work.

	Subject-Object	Subject-Object (categories)	Subject-Object (direction)
1	('we', 'us')	('we', 'us')	('we', 'us')
2	('he', 'him')	('he', 'him')	('they', 'them')
3	('they', 'them')	('they', 'them')	('I', 'me')
4	('I', 'me')	('I', 'me')	('he', 'him')
5	('she', 'her')	('she', 'her')	('they', 'us')
6	-	('she', 'me')	('I', 'him')
7	-	('I', 'her')	('we', 'me')
8	-	('they', 'him')	('she', 'us')
9	-	('he', 'them')	('we', 'them')
10	-	('I', 'us')	('he', 'me')
11	-	('we', 'me')	('they', 'him')
12	-	('she', 'them')	('I', 'us')
13	-	('they', 'her')	('we', 'him')
14	-	('we', 'them')	('he', 'us')
15	-	('they', 'us')	('they', 'me')

Table 4.4: Top 15 Subject-Object pairs as chosen under each of the approaches.

	Gender	Good-Bad	Informal-Formal	Subject-Object
Gender	1	0.0875	0.0963	0.0241
Good-Bad	0.0875	1	0.0179	0.0001
Informal-Formal	0.0963	0.0179	1	0.0112
Subject-Object	0.0241	0.0001	0.0112	1

Table 4.5: Absolute cosine similarity between the first principal components of different topics.

4.2 Generalizing to Other Word Embeddings

This section generalizes Bolukbasi et al. (2016) to GloVe word embeddings as well. We chose Stanford NLP's GloVe embeddings (Pennington et al., 2014), as they have been used in a number of relevant works, for example in Devi and Soman (2018), Bhaskaran et al. (2018), Eysenbach et al. (2018), and Caliskan et al. (2017). Most notably, GloVe differs from word2vec by keeping a global co-occurrence matrix that estimates the probability that a given word co-occurs with other words.

This work is using the 300-dimensional version of the publicly available pre-trained GloVe embeddings that were trained on the 6 billion tokens from Wikipedia 2014 and Gigaword 5. All of the resulting principal components were vastly similar in variance explained, which indicates that the defining sets approach can be generalized beyond any single choice of W.

The variance explained for the gender subspace across word2vecNews and GloVe is depicted in Figure 4.5, using Bolukbasi et al. (2016)'s ten defining pairs only. Note that the difference between the first and second principal component is much more pronounced for word2vecNews compared to GloVe, possibley because GloVe is capturing the associations and connotations of words more than word2vec, where the semantics influence the final embeddings more.

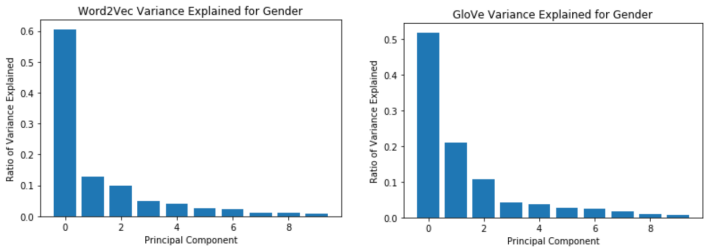

Figure 4.5: The variance explained by individual principal components for the gender subspace. Left: word2vec, right: GloVe.

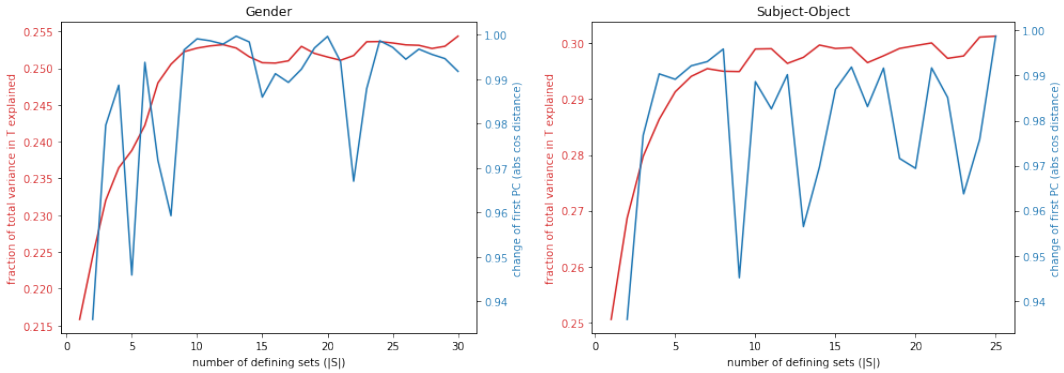

Figure 4.6: Fraction of total variance of $\mathbb{T}$ explained by defining pairs (red) and change of the first principal component after adding the nth defining pair (as measured in absolute cosine distance, blue) for gender and Subject-Object subspaces for GloVe.

We also implemented Bolukbasi et al. (2016)'s gender classifier for GloVe embeddings and compare them in Figure4.7. In particular, in GloVe embeddings, many female-associated words are classified as gender words, whereas many male-associated words are classified as non-gender words.

	Gender (categories)	**Subject-Object (categories)**
1	('herself', 'himself')	('we', 'her')
2	('woman', 'man')	('i', 'him')
3	('she', 'he')	('they', 'us')
4	('mother', 'father')	('they', 'them')
5	('mary', 'john')	('he', 'him')
6	('her', 'his')	('i', 'me')
7	('gal', 'guy')	('he', 'me')
8	('female', 'son')	('she', 'her')
9	('daughter', 'father')	('we', 'us')
10	('woman', 'male')	('he', 'them')
11	('girl', 'boy')	('she', 'him')
12	('mother', 'male')	('we', 'me')
13	('female', 'male')	('i', 'her')
14	('daughter', 'son')	('she', 'us')
15	('girl', 'guy')	('we', 'them')
16	('gal', 'boy')	('they', 'me')
17	('woman', 'he')	('i', 'them')
18	('she', 'man')	('they', 'him')
19	('mother', 'son')	('he', 'us')
20	('gal', 'male')	('they', 'her')

Table 4.6: The first 20 pairs chosen by the greedy algorithm under GloVe for the gender and Subject-Object subspaces.

Figure 4.7: Visualization of the trained classifier differentiating gender words (below horizontal line) from gender words, and distinguishing female words (left of vertical line) from male words. Top: word2vec, bottom: GloVe

Chapter 5

Conclusion

In conclusion, this thesis provided a novel way of reasoning about and generalizing PCA based methods for identifying subspaces that correspond to topics of interest from word vector representations. More specifically, we developed a framework for reasoning about the optimal dimension of a topic subspace as well as the optimal number of defining sets. As part of this work, we proposed a novel algorithm for greedily selecting defining sets from a larger pool of potential defining sets, as well as two approaches to generate that pool, and experimentally verified that it returns coherent results for binary topics such as gender or Subject-Object where definitional equivalence is very clear and there is low noise, e.g. from biases, stereotypes, or simply prevalent value judgments. We also found that prevalent biases and stereotypes will influence the defining sets that are chosen. Similarly, this work proposed a more principled methodology for choosing hyper-parameters, such as the number of defining sets or the number of dimensions of the subspace, and evaluated it experimentally.

5.1 Future Work

While this thesis has significantly extended prior work, it also paves the way for many other interesting research directions:

- First, while this work focused on binary topics, the presented results should generalize to higher dimensions and topics with more than two categories as well, and future work could experimentally verify this.
- Second, so far we assumed that k is known for each topic, but future work could relax this assumption: for certain multiclass topics like religion, we may not know how many categories there are in advance.

- Along similar lines, this thesis assumed that the size of noise U and direction are independent: future work could relax this assumption.
- This work provided theoretical justification for an $\ell = k - 1$-dimensional subspace, but future work could verify or dispute this claim through experiments that test desirable properties of the subspace such as interpretability and accuracy on e.g. classification tasks. In general, while our analysis takes the first step towards formally understanding the notion of defining sets, there is scope for improving the rigor of the theoretical analysis.
- We chose the specific seed direction B' for the direction approach to generating a pool of defining sets randomly and manually searched for $s_{\max}$ that resulted in largest variance explained, but future work could develop a more principled approach.
- We found that different topics' subspaces are close to orthogonal. Future work could reorient initially uninterpretable word embeddings along these subspaces and evaluate the new embeddings with respect to interpretability and accuracy on benchmark tasks.
- Lastly, future work could also expand and apply our method to downstream tasks of interest (e.g. Bolukbasi et al. (2016)'s debiasing task), for example by specifying the context in which the words may be used, defining task-specific evaluation metrics, or incorporating human feedback.

Appendix A

Supplementary Material

	1st PC	2nd PC	3rd PC
Gender	0.6052	0.1272	0.0992
Good-Bad	0.4989	0.1475	0.1137
Informal-Formal	0.2427	0.1130	0.0957
Subject-Object	0.6479	0.1731	0.0887

Table A.1: Fraction of variance explained by the first three principal components for each topic.

	Defining set	EV (additive)	EV (indiv.)	Cos Sim
1	('herself', 'himself')	0.2149	0.2149	n/a
2	('she', 'he')	0.2223	0.2137	0.9761
3	('woman', 'man')	0.2267	0.1725	0.9885
4	('daughter', 'son')	0.2289	0.1457	0.9959
5	('gal', 'guy')	0.2302	0.1325	0.9592
6	('her', 'his')	0.2341	0.2029	0.9922
7	('Mary', 'John')	0.2368	0.097	0.9875
8	('mother', 'father')	0.2384	0.1276	0.9977
9	('girl', 'boy')	0.2391	0.146	0.999
10	('female', 'male')	0.2395	0.054	0.9996
11	('daughter', 'father')	0.2388	0.1029	0.9971
12	('female', 'himself')	0.2386	0.1496	0.9718
13	('girl', 'male')	0.2392	0.1059	0.9993
14	('herself', 'male')	0.2399	0.1116	0.9974
15	('mother', 'son')	0.2399	0.1064	0.9985
16	('woman', 'boy')	0.2385	0.0946	0.9961
17	('girl', 'man')	0.2393	0.1126	0.9946
18	('woman', 'John')	0.2384	0.139	0.983
19	('Mary', 'man')	0.2396	0.1228	0.9892
20	('mother', 'he')	0.2392	0.1498	0.9916
21	('gal', 'son')	0.2397	0.1282	0.9927
22	('she', 'guy')	0.2401	0.144	0.9908
23	('girl', 'father')	0.2398	0.1145	0.9964
24	('female', 'boy')	0.2401	0.1267	0.9927
25	('daughter', 'his')	0.2409	0.1465	0.9923
26	('her', 'he')	0.2409	0.1447	0.9965
27	('she', 'his')	0.2407	0.1387	0.9962
28	('her', 'son')	0.2405	0.1486	0.9959
29	('daughter', 'John')	0.2405	0.1499	0.9916
30	('Mary', 'boy')	0.2408	0.1261	0.9969

Table A.2: Explained variance (additive and individual) and changes in the identified gender subspace based on the size of the defining set for gender. Additive explained variance is the percentage of total variance explained by all defining pairs down to that row, and individual explained variance is the percentage that only the pair in that row explains. The cosine similarity is between the identified gender subspace before and after adding the defining pair in that row to the defining set.

Bibliography

Native language cognate effects on second language lexical choice. *Transactions of the Association for Computational Linguistics*, 6:329–342. ISSN Transactions of the Association for Computational Linguistics.

Michael A Beam. Automating the news: How personalized news recommender system design choices impact news reception. *Communication Research*, 41(8):1019–1041, 2014. ISSN 0093-6502.

Sruthy K Bhaskaran, C Sreejith, and P C Rafeeque. Neural networks and conditional random fields based approach for effective question processing. *Procedia Computer Science*, 143: 211–218, 2018. ISSN 1877-0509.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*, 2016.

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *arXiv*, abs/1607.06520, 2016. URL `https://arxiv.org/pdf/1607.06520`.

Guillaume Bouzillé, Canelle Poirier, Boris Campillo-Gimenez, Marie-Laure Aubert, Mélanie Chabot, Emmanuel Chazard, Audrey Lavenu, and Marc Cuggia. Leveraging hospital big data to monitor flu epidemics. *Computer Methods and Programs in Biomedicine*, 154:153–160, 2018. ISSN 0169-2607.

Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora contain human-like biases.(cognitive science)(report)(author abstract). *Science*, 356(6334):183, 2017. ISSN 0036-8075.

Kw Church. Emerging trends word2vec. *Natural Language Engineering*, 23(1):155–162, 2017. ISSN 1351-3249.

Gregory F Cooper, Vijoy Abraham, Constantin F Aliferis, John M Aronis, Bruce G Buchanan, Richard Caruana, Michael J Fine, Janine E Janosky, Gary Livingston, Tom Mitchell, Stefano

Monti, and Peter Spirtes. Predicting dire outcomes of patients with community acquired pneumonia. *Journal of Biomedical Informatics*, 38(5):347–366, 2005. ISSN 1532-0464.

Sunipa Dev and Jeff M. Phillips. Attenuating bias in word vectors. *CoRR*, abs/1901.07656, 2019. URL `http://arxiv.org/abs/1901.07656`.

G Devi and K Soman. Co-occurrence based word representation for extracting named entities in tamil tweets. *Journal of Intelligent and Fuzzy Systems*, 34(3):1435–1442, 2018. ISSN 1064-1246. URL `http://search.proquest.com/docview/2017099140/`.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *arXiv*, abs/1810.04805, 2018. URL `http://arxiv.org/abs/1810.04805`.

Finale Doshi-Velez and Kim Been. Towards a rigorous science of interpretable machine learning. *arXiv.org*, 2017. URL `http://search.proquest.com/docview/2075249272/`.

Gunther Eysenbach, Zhe He, Jiang Bian, Nut Limsopatham, Jingcheng Du, Lu Tang, Yang Xiang, Degui Zhi, Jun Xu, Hsing-Yi Song, and Cui Tao. Public perception analysis of tweets during the 2015 measles outbreak: Comparative study using convolutional neural network models. *Journal of Medical Internet Research*, 20(7), 2018. ISSN 1439-4456.

Manaal Faruqui, Yulia Tsvetkov, Dani Yogatama, Chris Dyer, and Noah Smith. Sparse overcomplete word vector representations. 2015.

Timnit Gebru, Jonathan Krause, Yilun Wang, Duyun Chen, Jia Deng, Erez Lieberman Aiden, and Li Fei-Fei. Using deep learning and google street view to estimate the demographic makeup of neighborhoods across the united states. *Proceedings of the National Academy of Sciences of the United States of America*, 114(50):13108–13113, 2017. ISSN 00278424. URL `http://search.proquest.com/docview/1970272599/`.

Vishwani Gupta, Sven Giesselbach, Stefan Ruping, and Christian Bauckhage. Improving word embeddings using kernel pca. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, pages 200–208. ACL, 2019. URL `https://www.aclweb.org/anthology/W19-4323`.

Brian Heredia, Joseph Prusa, and Taghi Khoshgoftaar. Social media for polling and predicting united states election outcome. *Social Network Analysis and Mining*, 8(1):1–16, 2018. ISSN 1869-5450.

Kishlay Jha, Yaqing Wang, Guangxu Xun, and Aidong Zhang. Interpretable word embeddings for medical domain. In *2018 IEEE International Conference on Data Mining (ICDM)*, volume 2018-, pages 1061–1066. IEEE, 2018. ISBN 9781538691588.

Andreas Kaplan and Michael Haenlein. Siri, siri, in my hand: Who's the fairest in the land? on the interpretations, illustrations, and implications of artificial intelligence. *Business Horizons*, 62(1):15–25, 2019. ISSN 0007-6813.

Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. Human decisions and machine predictions *. *The Quarterly Journal of Economics*, 133(1): 237–293, 2017. ISSN 0033-5533.

Swati Kulkarni and Sudhir Dhage. Advanced credit score calculation using social media and machine learning. *Journal of Intelligent and Fuzzy Systems*, 36(3):2373–2380, 2019. ISSN 1064-1246. URL `http://search.proquest.com/docview/2197486667/`.

Himabindu Lakkaraju, Stephen H Bach, and Leskovec Jure. Interpretable decision sets: A joint framework for description and prediction. *KDD : proceedings. International Conference on Knowledge Discovery and Data Mining*, 2016:1675, 2016. ISSN 2154-817X.

David M. Lazer, R. Kennedy, Gary King, and A. Vespignani. The parable of google flu: Traps in big data analysis. *Science*, 343(6176), 2014. ISSN 0193-4511.

Benjamin Letham, Cynthia Rudin, Tyler Mccormick, and David Madigan. Interpretable classifiers using rules and bayesian analysis: Building a better stroke prediction model. *arXiv.org*, 9(3), 2015. ISSN 19326157. URL `http://search.proquest.com/docview/2083852849/`.

Zachary C. Lipton. The mythos of model interpretability. 2016.

Scott Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *arXiv.org*, 2017. URL `http://search.proquest.com/docview/2077007862/`.

Thomas Manzini, Lim Yao Chong, Alan W Black, and Yulia Tsvetkov. Black is to criminal as caucasian is to police: Detecting and removing multiclass bias in word embeddings. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 615–621, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1062. URL `https://www.aclweb.org/anthology/N19-1062`.

Tomas Mikolov, Martin Karafiat, Lukas Burget, Jan Cernocky, and Sanjeev Khudanpur. Recurrent neural network based language model. In *Proceedings of Interspeech 2010*, pages 1045–1049. ISCA, 2010. URL `http://www.fit.vutbr.cz/research/groups/speech/publi/2010/mikolov_interspeech2010_IS100722.pdf`.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space, 2013a.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119. NIPS, 2013b. URL https://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf.

Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. Advances in pre-training distributed word representations. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

Francesco Orciuoli and Mimmo Parente. An ontology-driven context-aware recommender system for indoor shopping based on cellular automata. *Journal of Ambient Intelligence and Humanized Computing*, 8(6):937–955, 2017. ISSN 1868-5137.

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

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on knowledge discovery and data mining*, volume 13-17- of *KDD '16*, pages 1135–1144. ACM, 2016. ISBN 9781450342322.

Dahiya Shashi, Handa S.S, and Singh N.P. Credit scoring using ensemble of various classifiers on reduced feature set. *Industrija*, 43(4):163–174, 2015. ISSN 0350-0373. URL https://doaj.org/article/63117a6b66994f8698ea4e623c0c059f.

Jamin Shin, Andrea Madotto, and Pascale Fung. Interpreting word embeddings with eigenvector analysis. In *Proceedings of the 32nd Conference on Neural Information Processing Systems, IRASL Workshop*, pages 746–751. NIPS, 2018. URL https://openreview.net/pdf?id=rJfJiR5ooX.

Cynthia J Sims, Leslie Meyn, Rich Caruana, R.Bharat Rao, Tom Mitchell, and Marijane Krohn. Predicting cesarean delivery with decision tree models. *American Journal of Obstetrics and Gynecology*, 183(5):1198–1206, 2000. ISSN 0002-9378.

Mansi Sood and Harmeet Kaur. Preference based personalized news recommender system. *International Journal of Advanced Computer Research*, 4(2):575–581, 2014. ISSN 22497277. URL http://search.proquest.com/docview/1613206224/.

Latanya Sweeney. Discrimination in online ad delivery. *Communications of the ACM*, 56(5):44–54, 2013. ISSN 00010782.

Kuo Chun Tsai, Li Liliane Wang, and Zhu Han. Caching for mobile social networks with deep learning: Twitter analysis for 2016 u.s. election. *IEEE Transactions on Network Science and Engineering*, PP(99):1–1, 2018. ISSN 2327-4697.

Ca Turner, AD Jacobs, Ck Marques, Jc Oates, DL Kamen, Pe Anderson, and Js Obeid. Word2vec inversion and traditional text classifiers for phenotyping lupus. *Bmc Medical Informatics And Decision Making*, 17(1), 2017. ISSN 14726947.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv*, abs/1706.03762, 2017. URL `http://arxiv.org/abs/1706.03762`.

Anna Aurora Wennakoski. Gdpr's tools to tackle ai and machine-driven data. *Journal of Data Protection and Privacy*, 2(1):62–71, 2018. ISSN 2398-1679.

Wayne Wobcke, Alfred Krzywicki, Yang Sok Kim, Xiongcai Cai, Michael Bain, Paul Compton, and Ashesh Mahidadia. A deployed people-to-people recommender system in online dating. 36(3):5, 2015. ISSN 0738-4602.

Jessica Zhao, Bill Zhang, and Himabindu Lakkaraju. Identifying interpretable word vector subspaces with pca. In *Machine Learning in Real Life Workshop Papers*. ICLR, 2020.