Hard things: robustness, bias, & fairness

DS4440 practical neural networks
So far this class has been purely technical
So far this class has been purely technical

But ML has huge societal implications and we, as the people who build these things, need to think about these
So far this class has been purely technical

But ML has huge societal implications and we, as the people who build these things, need to think about these

Arguably neural / deep models exacerbate these problems because they are brittle and hard to interpret
Today

A look at some of the key issues facing ML in practice, and societal implications of these
Today

A look at some of the key issues facing ML in practice, and societal implications of these

Caveat: This is a field unto itself, and I have no formal training; fortunately, on Monday we have more on this topic from external folks, who will provide a coherent framework!
Outline

- Fairness and bias; formalisms of these
- Robustness/brittleness
Socialist Rep. Alexandria Ocasio-Cortez (D-NY) claims that algorithms, which are driven by math, are racist
Face recognition researcher fights Amazon over biased AI

A researcher at MIT helped show big tech companies like Amazon, Microsoft and IBM that their facial recognition technology was biased.

Bias in facial recognition technology

A graduate student’s research project has revealed problems in facial recognition systems used by various companies. The study has shown a larger percentage of error in detecting female faces, especially in women with darker skin tones.
### Boston Housing Data (source: UCI ML datasets)

[Visit UCI ML datasets page](https://archive.ics.uci.edu/ml/datasets/Housing) for more details.

<table>
<thead>
<tr>
<th>CRIM</th>
<th>ZN</th>
<th>INDUS</th>
<th>CHAS</th>
<th>NOX</th>
<th>RM</th>
<th>AGE</th>
<th>DIS</th>
<th>RAD</th>
<th>TAX</th>
<th>PTRATIO</th>
<th>B</th>
<th>LSTAT</th>
<th>MEDV</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.00632</td>
<td>18.0</td>
<td>2.31</td>
<td>0.0</td>
<td>0.538</td>
<td>6.575</td>
<td>65.2</td>
<td>4.0900</td>
<td>1.0</td>
<td>296.0</td>
<td>15.3</td>
<td>396.90</td>
<td>4.98</td>
<td>24.0</td>
</tr>
<tr>
<td>0.02731</td>
<td>0.0</td>
<td>7.07</td>
<td>0.0</td>
<td>0.469</td>
<td>6.421</td>
<td>78.9</td>
<td>4.9671</td>
<td>2.0</td>
<td>242.0</td>
<td>17.8</td>
<td>396.90</td>
<td>9.14</td>
<td>21.6</td>
</tr>
<tr>
<td>0.02729</td>
<td>0.0</td>
<td>7.07</td>
<td>0.0</td>
<td>0.469</td>
<td>7.185</td>
<td>61.1</td>
<td>4.9671</td>
<td>2.0</td>
<td>242.0</td>
<td>17.8</td>
<td>392.83</td>
<td>4.03</td>
<td>34.7</td>
</tr>
<tr>
<td>0.03237</td>
<td>0.0</td>
<td>2.18</td>
<td>0.0</td>
<td>0.458</td>
<td>6.998</td>
<td>45.8</td>
<td>6.0622</td>
<td>3.0</td>
<td>222.0</td>
<td>18.7</td>
<td>394.63</td>
<td>2.94</td>
<td>33.4</td>
</tr>
<tr>
<td>0.06905</td>
<td>0.0</td>
<td>2.18</td>
<td>0.0</td>
<td>0.458</td>
<td>7.147</td>
<td>54.2</td>
<td>6.0622</td>
<td>3.0</td>
<td>222.0</td>
<td>18.7</td>
<td>396.90</td>
<td>5.33</td>
<td>36.2</td>
</tr>
<tr>
<td>0.02985</td>
<td>0.0</td>
<td>2.18</td>
<td>0.0</td>
<td>0.458</td>
<td>6.430</td>
<td>58.7</td>
<td>6.0622</td>
<td>3.0</td>
<td>222.0</td>
<td>18.7</td>
<td>394.12</td>
<td>5.21</td>
<td>28.7</td>
</tr>
<tr>
<td>0.08829</td>
<td>12.5</td>
<td>7.87</td>
<td>0.0</td>
<td>0.524</td>
<td>6.012</td>
<td>66.6</td>
<td>5.5605</td>
<td>5.0</td>
<td>311.0</td>
<td>15.2</td>
<td>395.50</td>
<td>12.43</td>
<td>22.9</td>
</tr>
<tr>
<td>0.14455</td>
<td>12.5</td>
<td>7.87</td>
<td>0.0</td>
<td>0.524</td>
<td>6.172</td>
<td>96.1</td>
<td>5.9505</td>
<td>5.0</td>
<td>311.0</td>
<td>15.2</td>
<td>396.90</td>
<td>19.15</td>
<td>27.1</td>
</tr>
<tr>
<td>0.21124</td>
<td>12.5</td>
<td>7.87</td>
<td>0.0</td>
<td>0.524</td>
<td>5.631</td>
<td>100.0</td>
<td>6.0821</td>
<td>5.0</td>
<td>311.0</td>
<td>15.2</td>
<td>386.63</td>
<td>29.93</td>
<td>16.5</td>
</tr>
<tr>
<td>0.17004</td>
<td>12.5</td>
<td>7.87</td>
<td>0.0</td>
<td>0.524</td>
<td>6.004</td>
<td>85.9</td>
<td>6.5921</td>
<td>5.0</td>
<td>311.0</td>
<td>15.2</td>
<td>386.71</td>
<td>17.10</td>
<td>18.9</td>
</tr>
<tr>
<td>0.22489</td>
<td>12.5</td>
<td>7.87</td>
<td>0.0</td>
<td>0.524</td>
<td>6.377</td>
<td>94.3</td>
<td>6.3467</td>
<td>5.0</td>
<td>311.0</td>
<td>15.2</td>
<td>392.52</td>
<td>20.45</td>
<td>15.0</td>
</tr>
</tbody>
</table>
**Target Variable**

<table>
<thead>
<tr>
<th>CRIM</th>
<th>ZN</th>
<th>INDUS</th>
<th>CHAS</th>
<th>NOX</th>
<th>RM</th>
<th>AGE</th>
<th>DIS</th>
<th>RAD</th>
<th>TAX</th>
<th>PRTATIO</th>
<th>B</th>
<th>LSTAT</th>
<th>MEDV</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.00632</td>
<td>18.0</td>
<td>2.31</td>
<td>0.0</td>
<td>0.538</td>
<td>6.575</td>
<td>65.2</td>
<td>4.0900</td>
<td>1.0</td>
<td>296.0</td>
<td>15.3</td>
<td>396.90</td>
<td>4.98</td>
<td>24.0</td>
</tr>
<tr>
<td>0.02731</td>
<td>0.0</td>
<td>7.07</td>
<td>0.0</td>
<td>0.469</td>
<td>6.421</td>
<td>78.9</td>
<td>4.9671</td>
<td>2.0</td>
<td>242.0</td>
<td>17.8</td>
<td>396.90</td>
<td>9.14</td>
<td>21.6</td>
</tr>
<tr>
<td>0.02729</td>
<td>0.0</td>
<td>7.07</td>
<td>0.0</td>
<td>0.469</td>
<td>7.185</td>
<td>61.1</td>
<td>4.9671</td>
<td>2.0</td>
<td>242.0</td>
<td>17.8</td>
<td>392.83</td>
<td>4.03</td>
<td>34.7</td>
</tr>
<tr>
<td>0.03237</td>
<td>0.0</td>
<td>2.18</td>
<td>0.0</td>
<td>0.458</td>
<td>6.998</td>
<td>45.6</td>
<td>6.0622</td>
<td>3.0</td>
<td>222.0</td>
<td>16.7</td>
<td>394.63</td>
<td>2.94</td>
<td>33.4</td>
</tr>
<tr>
<td>0.06905</td>
<td>0.0</td>
<td>2.18</td>
<td>0.0</td>
<td>0.458</td>
<td>7.147</td>
<td>54.2</td>
<td>6.0622</td>
<td>3.0</td>
<td>222.0</td>
<td>18.7</td>
<td>396.90</td>
<td>5.33</td>
<td>36.2</td>
</tr>
<tr>
<td>0.02985</td>
<td>0.0</td>
<td>2.18</td>
<td>0.0</td>
<td>0.458</td>
<td>6.430</td>
<td>58.7</td>
<td>6.0622</td>
<td>3.0</td>
<td>222.0</td>
<td>18.7</td>
<td>394.12</td>
<td>5.21</td>
<td>28.7</td>
</tr>
<tr>
<td>0.08829</td>
<td>12.5</td>
<td>7.87</td>
<td>0.0</td>
<td>0.524</td>
<td>6.012</td>
<td>66.6</td>
<td>5.5605</td>
<td>5.0</td>
<td>311.0</td>
<td>15.2</td>
<td>395.60</td>
<td>12.43</td>
<td>22.9</td>
</tr>
<tr>
<td>0.14455</td>
<td>12.5</td>
<td>7.87</td>
<td>0.0</td>
<td>0.524</td>
<td>6.172</td>
<td>96.1</td>
<td>5.9505</td>
<td>5.0</td>
<td>311.0</td>
<td>15.2</td>
<td>396.90</td>
<td>19.15</td>
<td>27.1</td>
</tr>
<tr>
<td>0.21124</td>
<td>12.5</td>
<td>7.87</td>
<td>0.0</td>
<td>0.524</td>
<td>5.631</td>
<td>100.0</td>
<td>6.0821</td>
<td>5.0</td>
<td>311.0</td>
<td>15.2</td>
<td>386.63</td>
<td>29.93</td>
<td>16.5</td>
</tr>
<tr>
<td>0.17004</td>
<td>12.5</td>
<td>7.87</td>
<td>0.0</td>
<td>0.524</td>
<td>6.004</td>
<td>85.9</td>
<td>6.5921</td>
<td>5.0</td>
<td>311.0</td>
<td>15.2</td>
<td>386.71</td>
<td>17.10</td>
<td>18.9</td>
</tr>
<tr>
<td>0.22489</td>
<td>12.5</td>
<td>7.87</td>
<td>0.0</td>
<td>0.524</td>
<td>6.377</td>
<td>94.3</td>
<td>6.3467</td>
<td>5.0</td>
<td>311.0</td>
<td>15.2</td>
<td>392.52</td>
<td>20.45</td>
<td>15.0</td>
</tr>
</tbody>
</table>

**MEDV**: Median value of owner-occupied homes in $1000's
Hmmm...

CRIM: Per capita crime rate by town
ZN: Proportion of residential land zoned for lots over 25,000 sq. ft
INDUS: Proportion of non-retail business acres per town
CHAS: Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
NOX: Nitric oxide concentration (parts per 10 million)
RM: Average number of rooms per dwelling
AGE: Proportion of owner-occupied units built prior to 1940
DIS: Weighted distances to five Boston employment centers
RAD: Index of accessibility to radial highways
TAX: Full-value property tax rate per $10,000
PTRATIO: Pupil-teacher ratio by town
B: 1000(Bk - 0.63)^2, where Bk is the proportion of [people of African American descent] by town
LSTAT: Percentage of lower status of the population
MEDV: Median value of owner-occupied homes in $1000s
Hmmm…

Q: Is it ok to use to B here?
In general how do we define bias?

- Discrimination on the basis of things (*features*, if you will) that we feel morally should have no bearing.
In general how do we define bias?

- Discrimination on the basis of things (*features*, if you will) that we feel morally should have no bearing

- *Especially* for domains in which predictions may have a large impact on individuals (criminal justice, education, housing …)
Legally “protected classes”

**Race** (Civil Rights Act of 1964); **Color** (Civil Rights Act of 1964); **Sex** (Equal Pay Act of 1963; Civil Rights Act of 1964); **Religion** (Civil Rights Act of 1964); **National origin** (Civil Rights Act of 1964); **Citizenship** (Immigration Reform and Control Act); **Age** (Age Discrimination in Employment Act of 1967); **Pregnancy** (Pregnancy Discrimination Act); **Familial status** (Civil Rights Act of 1968); **Disability status** (Rehabilitation Act of 1973; Americans with Disabilities Act of 1990); **Veteran status** (Vietnam Era Veterans' Readjustment Assistance Act of 1974; Uniformed Services Employment and Reemployment Rights Act); **Genetic information** (Genetic Information Nondiscrimination Act)

*Legally recognized as unsound bases to treat people differently!*
Can’t we just withhold features that contain this info?

- No: There are often *proxy features* that implicitly capture this
  - e.g., zip-code may strongly correlate with race
When things go wrong
'Coded Bias' Review: When the Bots Are Racist

This clear-eyed documentary explores how machine-learning algorithms can perpetuate society's existing class-, race- and gender-based inequities.

Jay Bhatia is a subject of the documentary "Coded Bias." The Empire Media

By Devika Girish
Nov. 11, 2020
Machine Bias

There’s software used across the country to predict future criminals. And it’s biased against blacks.
We also turned up significant racial disparities, just as Holder feared. In forecasting who would re-offend, the algorithm made mistakes with black and white defendants at roughly the same rate but in very different ways.

- The formula was particularly likely to falsely flag black defendants as future criminals, wrongly labeling them this way at almost twice the rate as white defendants.
- White defendants were mislabeled as low risk more often than black defendants.
Twitter apologises for 'racist' image-cropping algorithm

Users highlight examples of feature automatically focusing on white faces over black ones
Jordan Simonovski @jsimonovski · Sep 20

I wonder if Twitter does this to fictional characters too.
I had tweeted in 2019 about @Twitter cropping #womeninAI headless while cropping men correcting. When I raised this, many men in field accused me of making up a non-existent issue just to gain attention. Sadly #AI #bias is not yet fixed.

Great predictions for 2019 from Yann LeCun, Hilary Mason, Andrew Ng, and Rumman Chowdhury wp.me/p8wLEc-ahoT by @kharijohnson

@AnimaAnandkumar @AnimaAnandkumar · Sep 19
Women and POC are rendered invisible on @Twitter and other platforms by biased #AI #WomenInSTEM #womenintech #WomensRights

@AnimaAnandkumar @AnimaAnandkumar · Sep 19
@rencdh @hmason @timmitGebur @mraginsky were some people discussing it in back in 2019. Sadly @Twitter did nothing. #AI #bias continues
Bias in ML

BRIEF HISTORY OF FAIRNESS IN ML

PAPERS


LOL FAIRNESS!!

OH, CRAP.
Wait, which bias?

- There are various technical sorts of biases
  - E.g., think of the **bias-variance** trade-off (this being an inductive bias)
Wait, which bias?

- There are various technical sorts of biases
  - E.g., think of the bias-variance trade-off (this being an inductive bias)

- Here we are taking about bias from an ethical perspective

  The blind application of machine learning runs the risk of amplifying biases present in data.

  [Bolukbasi et al., 2016]
Some recentish (mostly NLP) examples
Man is to Computer Programmer as Woman is to Homemaker?
Debiasing Word Embeddings

Tolga Bolukbasi¹, Kai-Wei Chang², James Zou², Venkatesh Saligrama¹,², Adam Kalai²

¹Boston University, 8 Saint Mary’s Street, Boston, MA
²Microsoft Research New England, 1 Memorial Drive, Cambridge, MA

tolgab@bu.edu, kw@kwchang.net, jamesyzou@gmail.com, srv@bu.edu, adam.kalai@microsoft.com
Abstract

The blind application of machine learning runs the risk of amplifying biases present in data. Such a danger is facing us with word embedding, a popular framework to represent text data as vectors which has been used in many machine learning and natural language processing tasks. We show that even word embeddings trained on Google News articles exhibit female/male gender stereotypes to a disturbing extent. This raises concerns because their widespread use, as we describe, often tends to amplify these biases. Geometrically, gender bias is first shown to be captured by a direction in the word embedding. Second, gender neutral words are shown to be linearly separable from gender definition words in the word embedding. Using these properties, we provide a methodology for modifying an embedding to remove gender stereotypes, such as the association between the words receptionist and female, while maintaining desired associations such as between the words queen and female. We define metrics to quantify both direct and indirect gender biases in embeddings, and develop algorithms to “debias” the embedding. Using crowd-worker evaluation as well as standard benchmarks, we empirically demonstrate that our algorithms significantly reduce gender bias in embeddings while preserving the useful properties such as the ability to cluster related concepts and to solve analogy tasks. The resulting embeddings can be used in applications without amplifying gender bias.

1 Introduction

There have been hundreds or thousands of papers written about word embeddings and their applications, from Web search [27] to parsing Curriculum Vitae [16]. However, none of these papers have recognized how blatantly sexist the embeddings are and hence risk introducing biases of various types into real-world systems. A word embedding represents a word (or common phrase) as a d-dimensional word vector \(~w~\) as a dictionary of sorts for computer programs that would like to use word meaning. First, words with similar semantic meanings tend to have vectors that are close together. Second, the vector differences between words in embeddings have been shown to represent relationships between words [32, 26]. For example given an analogy puzzle, “man is to king as woman is to x” (denoted as man:king::woman:x), simple arithmetic of the embedding vectors finds that x = queen is the best answer because:

\[\text{man} - \text{woman} \approx \text{king} - \text{queen}\]
\[
\overrightarrow{\text{man}} - \overrightarrow{\text{woman}} \approx \overrightarrow{\text{king}} - \overrightarrow{\text{queen}}
\]

\[
\overrightarrow{\text{man}} - \overrightarrow{\text{woman}} \approx \overrightarrow{\text{computer programmer}} - \overrightarrow{\text{homemaker}}
\]
Gender stereotype *she-he* analogies.

- sewing-carpentry : register-nurse-physician : housewife-shopkeeper
- nurse-surgeon : interior designer-architect : softball-baseball
- blond-burly : feminism-conservatism : cosmetics-pharmaceuticals
- giggle-chuckle : vocalist-guitarist : petite-lanky
- sassy-snappy : diva-superstar : charming-affable
- volleyball-football : cupcakes-pizzas : hairdresser-barber

Gender appropriate *she-he* analogies.

- queen-king : sister-brother : mother-father
- waitress-waiter : ovarian cancer-prostate cancer : convent-monastery
**Extreme she occupations**

1. homemaker  
2. nurse  
3. receptionist  
4. librarian  
5. socialite  
6. hairdresser  
7. nanny  
8. bookkeeper  
9. stylist  
10. housekeeper  
11. interior designer  
12. guidance counselor

**Extreme he occupations**

1. maestro  
2. skipper  
3. protege  
4. philosopher  
5. captain  
6. architect  
7. financier  
8. warrior  
9. broadcaster  
10. magician  
11. fighter pilot  
12. boss
The embedding captures gender stereotypes and sexism.

(related [Schmidt ‘15])

Slides: Adam Kalai
Easier to debias an embedding than to debias a human

DEFINITIONAL

(related [Schmidt ‘15])
The geometry of gender

Select pairs of words that reflect gender opposites.

- female - male
- she - he
- her - his
- woman - man
- Mary - John

Principal components

are high, indicating that these pairs capture the intuitive notion of gender.

To identify the gender subspace, we took the ten gender pair difference vectors and computed its principal components (PCs). As Figure 6 shows, there is a single direction that explains the majority of variance
% of variance explained

The top PC seems to capture the gender subspace $B$. 
1. Identify words that are gender-neutral $N$ and gender-definitional $S$.

2. Project away the gender subspace from the gender-neutral words.
   \[ w := w - w \cdot B \] for $w \in N$. $B$ is the gender subspace.

De-biasing ensures that gender neutral words are zero in the gender subspace.
Lipstick on a Pig: 
Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them

Hila Gonen\(^1\) and Yoav Goldberg\(^{1,2}\)
\(^1\)Department of Computer Science, Bar-Ilan University
\(^2\)Allen Institute for Artificial Intelligence
{hilagnn,yoav.goldberg}@gmail.com

Abstract

Word embeddings are widely used in NLP for a vast range of tasks. It was shown that word embeddings derived from text corpora reflect gender biases in society. This phenomenon is pervasive and consistent across different word embedding models, causing serious concern. Several recent works tackle this problem, and propose methods for significantly reducing this gender bias in word embeddings, demonstrating convincing results. However, we argue that this removal is superficial. While the bias is indeed substantially reduced according to the provided bias definition, the actual effect is mostly hiding the bias, not removing it. The gender bias information is still reflected in the distances between "gender-neutralized" words in the debiased embeddings, and can be recovered from them. We present a series of experiments to support this claim, for two debiasing methods. We conclude that existing bias removal techniques are insufficient, and should not be trusted for providing gender-neutral modeling.

1 Introduction

Word embeddings have become an important component in many NLP models and are widely used for a vast range of downstream tasks. However, these word representations have been proven to reflect social biases (e.g. race and gender) that naturally occur in the data used to train them (Caliskan et al., 2017; Garg et al., 2018).

In this paper we focus on gender bias. Gender bias was demonstrated to be consistent and pervasive across different word embeddings. Bolukbasi et al. (2016b) show that using word embeddings for simple analogies surfaces many gender stereotypes. For example, the word embedding they use (word2vec embedding trained on the Google News dataset\(^1\) (Mikolov et al., 2013)) answer the analogy "man is to computer programmer as woman is to x" with "x = homemaker". Caliskan et al. (2017) further demonstrate association between female/male names and groups of words stereotypically assigned to females/males (e.g. arts vs. science). In addition, they demonstrate that word embeddings reflect actual gender gaps in reality by showing the correlation between the gender association of occupation words and labor-force participation data.

Recently, some work has been done to reduce the gender bias in word embeddings, both as a post-processing step (Bolukbasi et al., 2016b) and as part of the training procedure (Zhao et al., 2018). Both works substantially reduce the bias with respect to the same definition: the projection on the gender direction (i.e. \(\text{he} \cdot \text{she}\)), introduced in the former. They also show that performance on word similarity tasks is not hurt.

We argue that current debiasing methods, which lean on the above definition for gender bias and directly target it, are mostly hiding the bias rather than removing it. We show that even when drastically reducing the gender bias according to this definition, it is still reflected in the geometry of the representation of "gender-neutral" words, and a lot of the bias information can be recovered.

2 Gender Bias in Word Embeddings

In what follows we refer to words and their vectors interchangeably.

Definition and Existing Debiasing Methods

Bolukbasi et al. (2016b) define the gender bias of a word \(w\) by its projection on the "gender direction": \(w \cdot (\text{he} \cdot \text{she})\), assuming all vectors are normalized. The larger a word's projection is on \(\text{he} \cdot \text{she}\), the more biased the word is.

\(\frac{1}{\|	ext{he}\| \|	ext{she}\|} \text{he} \cdot \text{she} = \begin{bmatrix} 1 \end{bmatrix} \circ \begin{bmatrix} 0 \end{bmatrix} = \begin{bmatrix} 0 \end{bmatrix}
\)

\(\frac{1}{\|	ext{he}\| \|	ext{she}\|} \text{he} \cdot \text{she} = \begin{bmatrix} 1 \end{bmatrix} \circ \begin{bmatrix} 1 \end{bmatrix} = \begin{bmatrix} 1 \end{bmatrix}
\)

The code for our experiments is available at https://github.com/gonenhila/gender_bias_lipstick.
Null It Out: Guarding Protected Attributes by Iterative Nullspace Projection

Shauli Ravfogel$^{1,2}$    Yanai Elazar$^{1,2}$    Hila Gonen$^1$    Michael Twiton$^3$    Yoav Goldberg$^{1,2}$

$^1$Computer Science Department, Bar Ilan University
$^2$Allen Institute for Artificial Intelligence
$^3$Independent researcher

{shauli.ravfogel, yanaeila, hilagnn, mtwito101, yoav.goldberg}@gmail.com

Abstract
The ability to control for the kinds of information encoded in neural representation has a variety of use cases, especially in light of the challenge of interpreting these models. We present Iterative Null-space Projection (INLP), a novel method for removing information from neural representations. Our method is based on repeated training of linear classifiers that predict a certain property we aim to remove, followed by projection of the representations on their null-space. By doing so, the classifiers become oblivious to that target property, making it hard to linearly separate the data according to it. While applicable for multiple uses, we evaluate our method on bias and fairness use-cases, and show that our method is able to mitigate bias in word embeddings, as well as to increase fairness in a setting of multi-class classification.

1 Introduction

What is encoded in vector representations of textual data, and can we control it? Word embeddings, pre-trained language models, and more generally deep learning methods emerge as very effective techniques for text classification. Accordingly, they are increasingly being used for predictions in real-world situations. A large part of the success is due to the models' ability to perform representation learning, coming up with effective feature representations for the prediction task at hand. However, these learned representations, while effective, are also notoriously opaque: we do not know what is encoded in them. Indeed, there is an emerging line of work on probing deep-learning derived representations for syntactic (Linzen et al., 2016; Hewitt and Manning, 2019; Goldberg, 2019), semantic (Tenney et al., 2019) and factual knowledge (Petroni et al., 2019). There is also evidence that they capture a lot of information regarding the demographics of the author of the text (Blodgett et al., 2016; Elazar and Goldberg, 2018).

What can we do in situations where we do not want our representations to encode certain kinds of information? For example, we may want a word representation that does not take tense into account, or that does not encode part-of-speech distinctions. We may want a classifier that judges the formality of the text, but which is also oblivious to the topic the text was taken from. Finally, and also our empirical focus in this work, this situation often arises when considering fairness and bias of language-based classification. We may not want our word-embeddings to encode gender stereotypes, and we do not want sensitive decisions on hiring or loan approvals to condition on the race, gender or age of the applicant.

We present a novel method for selectively removing specific kinds of information from a representation. Previous methods are either based on projection on a pre-specified, user-provided direction (Bolukbasi et al., 2016), or on adding an adversarial objective to an end-to-end training process (Xie et al., 2017). Both of these have benefits and limitations, as we discuss in the related work section (§2). Our proposed method, Iterative Null-space Projection (INLP), presented in section 4, can be seen as a combination of these approaches, capitalizing on the benefits of both. Like the projection methods, it is also based on the mathematical notion of linear projection, a commonly used de-
Figure 1: t-SNE projection of GloVe vectors of the most gender-biased words after $t=0$, 3, 18, and 35 iterations of INLP. Words are colored according to being male-biased or female-biased.
I had tried building an algorithm for sentiment analysis based on word embeddings — evaluating how much people like certain things based on what they say about them. When I applied it to restaurant reviews, I found it was ranking Mexican restaurants lower. The reason was not reflected in the star ratings or actual text of the reviews. It’s not that people don’t like Mexican food. The reason was that the system had learned the word “Mexican” from reading the Web.

If a restaurant were described as doing something “illegal”, that would be a pretty negative statement about the restaurant, right? But the Web contains lots of text where people use the word “Mexican” disproportionately along with the word “illegal”, particularly to associate “Mexican immigrants” with “illegal immigrants”. The system ends up learning that “Mexican” means something similar to “illegal”, and so it must mean something bad.

https://blog.conceptnet.io/posts/2017/conceptnet-numberbatch-17-04-better-less-stereotyped-word-vectors/
How to make a racist AI without really trying

A cautionary tutorial.

This post is part of the ConceptNet blog, written by Rob Speer from Luminoso.

https://gist.github.com/rspeer/ef750e7e407e04894cb3b78a82d66aed
Gender Bias in Contextualized Word Embeddings

Jieyu Zhao¹, Tianlu Wang², Mark Yatskar³, Ryan Cotterell⁴, Vicente Ordonez², Kai-Wei Chang¹

¹UCLA, ²University of Virginia, ³Allen Institute for AI, ⁴University of Cambridge

Slide credit to the authors: http://kwchang.net/documents/slides/zhao2019gender_slide.pdf
Coreference resolution

Model fails for "she" when given same context.

Zhao et al., 2018
He taught himself to **play** the violin.

Do you enjoy the **play**?

Embedding visualization

**word2vec**

**ELMo**

Zhao et al., 2018
Gender | Male Pronouns | Female Pronouns
--- | --- | ---
Occurrence (*1000) | 5,300 | 1,600

- Male pronouns (he, him, his) occur 3 times more often than females’ (she, her)

Zhao et al., 2018
The **driver** transported the counselor to the hospital because **she** was paid

The **driver** transported the counselor to the hospital because **he** was paid
• WinoBias dataset\textsuperscript{1}  
  • Pro-Stereotypical and Anti-Stereotypical dataset

• Bias: different performance between Pro. and Anti. dataset.

\textsuperscript{1} \url{https://uclanlp.github.io/corefBias}

Zhao et al., 2018
NLP

Bias in Coreference

• ELMo boosts the performance

However, enlarge the bias (\(\Delta\))

\[\Delta: 29.6\]

\[\Delta: 26.6\]

\[45\]

\[53.75\]

\[62.5\]

\[71.25\]

\[80\]

GloVe

+ELMo

\(F_1\) (%)

OnoNotes

Pro.

Anti.

Zhao et al., 2018
• Data Augmentation
  • Generate gender swapped training variants
  • Better mitigation; need retrain

Zhao et al., 2018
Who Criticized Trump PERSON in Texts, Is Fired GPE. The New York Times ORG SectionsSEARCH Skip to contentSkip to site indexPoliticsSubscribeLog InSubscribeLog In Today's PaperAdvertisementSupported ORG by F.B.I. Agent Peter Strzok PERSON, who was taken off the special counsel investigation after his disparaging texts about President Trump PERSON were uncovered, was fired. Credit T.J. Kirkpatrick PERSON for The New York Times By Adam Goldman ORG and Michael S. Schmidt Aug PERSON 13 CARDINAL, 2018 WASHINGTON CARDINAL — Peter Strzok PERSON, the senior counterintelligence agent who disparaged President Trump PERSON in inflammatory text messages and helped oversee the Hillary Clinton PERSON email and Russia GPE investigations, has been fired for violating bureau policies, Mr. Strzok’s lawyer said Monday DATE. Mr. Trump and his allies seized on the texts — exchanged during the 2016 DATE campaign with a former F.B.I. GPE lawyer, Lisa Page — in PERSON assaulting the Russia GPE investigation as an illegitimate “witch hunt.” Mr. Strzok PERSON, who rose over 20 years DATE at the F.B.I. GPE to become one of its most experienced counterintelligence agents, was a key figure in the early months DATE of the inquiry. Along with writing the texts, Mr. Strzok PERSON was accused of sending a highly sensitive search warrant to his personal email account. The F.B.I. GPE had been under immense political pressure by Mr. Trump PERSON to dismiss Mr. Strzok PERSON, who was removed last summer DATE from the staff of the special counsel, Robert S. Mueller III PERSON. The president has repeatedly denounced Mr. Strzok in posts on Twitter EVENT, and on Monday DATE expressed satisfaction that he had been sacked. Mr. Trump’s ORG victory traces back to June DATE, when Mr. Strzok PERSON’s conduct was laid out in a wide-ranging inspector general’s report on how the F.B.I. GPE handled the investigation of Hillary Clinton’s PERSON emails in the run-up to the 2016 DATE election. The report was critical of Mr. Strzok’s PERSON conduct in sending the
Interpretability Analysis for Named Entity Recognition to Understand System Predictions and How They Can Improve

Oshin Agarwal  
University of Pennsylvania  
oagarwal@seas.upenn.edu

Yinfei Yang  
Google AI  
yinfeiy@google.com

Ani Nenkova  
University of Pennsylvania  
enkova@seas.upenn.edu

Byron C. Wallace  
Northeastern University  
b.wallace@northeastern.edu

Abstract

Named Entity Recognition systems achieve remarkable performance on domains such as English news. It is natural to ask: What are these models actually learning to achieve this? Are they merely memorizing the names themselves? Or are they capable of interpreting the text and inferring the correct entity type from the linguistic context? We examine these questions by contrasting the performance of several variants of LSTM-CRF architectures for named entity recognition, with some provided only representations of the context as features. We also perform similar experiments for BERT. We find that context representations do contribute to system performance, but that the main factor driving high performance is learning the name tokens themselves. We enlist human annotators to evaluate the feasibility of inferring entity types from the context alone and find that, while people are not able to infer the entity type either for the majority of the errors made by the context-only system, there is some room for improvement. A system should be able to recognize any name in a predictive context correctly and our experiments indicate that current systems may be further improved by such capability.

1 Introduction

Named Entity Recognition (NER) is the task of identifying words and phrases in text that refer to a person, location or organization name, or some finer subcategory of these types. NER systems work well on domains such as English news, achieving high performance on standard datasets like MUC-6 (Grishman and Sundheim, 1996), CoNLL 2003 (Tjong Kim Sang and De Meulder, 2003) and OntoNotes (Pradhan and Xue, 2009).

However, prior work has shown that the performance deteriorates on entities unseen in the training data (Augenstein et al., 2017) and when entities are switched with a diverse set of entities even within the same dataset (Agarwal et al., 2020).

In this paper, we examine the interpretability and explainability of models used for the task, focusing on the type of textual clues that lead systems to make predictions. Consider, for instance, the sentence “Nicholas Romanov abdicated the throne in 1917”. The correct identification of “Nicholas Romanov” as a person may be due to (i) knowing that Nicholas is a fairly common name and that (ii) the capitalized word after that ending with “-ov” is likely a Slavic last name too. Alternatively, (iii) a competent user of language would know the selectional restrictions (Framis, 1994; Akbik et al., 2013; Chersoni et al., 2018) for the subject of the verb abdicate, i.e., that only a person may abdicate the throne, so $X$ in the context “$X$ abdicated the throne” can only be a person.

Such probing of the reasons behind a prediction is in line with early work on NER that emphasized the need to consider both internal (features of the name itself) and external (context features) evidence when determining the semantic types of named entities (McDonald, 1993). We specifically focus on the interplay between learning names as in (i), and recognizing constraining contexts as in (iii), given that (ii) can be construed as a more general case of (i), in which word shape and morphological features may indicate that a word is a name even if the exact name is never explicitly seen by the system (cf. Table 1 in (Bikel et al., 1999)).

As a foundation for our work, we conduct experiments with BiLSTM-CRF models (Huang et al., 2015) modified to use only context representations or only word identities to quantify the extent to which systems exploit word and context evidence (Section 3). We test these systems on three different datasets to identify trends that generalize across corpora. We show that context does somewhat inform system predictions, but the major driver of performance is learning the name tokens themselves.
Jim Jones was involved in a car accident …
Jim Jones was involved in a car accident …
Jim Jones was involved in a car accident …

Negasi was involved in a car accident …
Jim Jones was involved in a car accident ...

Negasi was involved in a car accident ...

Entity-Switched Datasets: An Approach to Auditing the In-Domain Robustness of Named Entity Recognition Models

Oshin Agarwal
University of Pennsylvania
oagarwal@seas.upenn.edu

Yinfei Yang
Google AI
yinfely@google.com

Byron C. Wallace
Northeastern University
b.wallace@northeastern.edu

Ani Nenkova
University of Pennsylvania
nenkova@seas.upenn.edu
<table>
<thead>
<tr>
<th>Country</th>
<th>(Huang et al., 2015) GloVe words</th>
<th>(Lample et al., 2016) GloVe words+chars</th>
<th>(Devlin et al., 2019) BERT subwords</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P</td>
<td>R</td>
<td>F1</td>
</tr>
<tr>
<td>Original</td>
<td>94.7</td>
<td>95.6</td>
<td>95.2</td>
</tr>
<tr>
<td>India</td>
<td>94.2</td>
<td>95.5</td>
<td>94.8</td>
</tr>
<tr>
<td>Vietnam</td>
<td>93.1</td>
<td>82.3</td>
<td>85.8</td>
</tr>
</tbody>
</table>
The Woman Worked as a Babysitter: On Biases in Language Generation

Emily Sheng¹, Kai-Wei Chang², Premkumar Natarajan¹, Nanyun Peng¹

¹ Information Sciences Institute, University of Southern California
² Computer Science Department, University of California, Los Angeles

{ewsheng, pnataraj, npeng}@isi.edu, kwchang@cs.ucla.edu

Abstract

We present a systematic study of biases in natural language generation (NLG) by analyzing text generated from prompts that contain mentions of different demographic groups. In this work, we introduce the notion of the regard towards a demographic, use the varying levels of regard towards different demographics as a defining metric for bias in NLG, and analyze the extent to which sentiment scores are a relevant proxy metric for regard. To this end, we collect strategically-generated text from language models and manually annotate the text with both sentiment and regard scores. Additionally, we build an automatic regard classifier through transfer learning, so that we can analyze biases in unseen text. Together, these methods reveal the extent of the biased nature of language model generations. Our analysis provides a study of biases in NLG, bias metrics and correlated human judgments, and empirical evidence on the usefulness of our annotated dataset.
The Woman Worked as a Babysitter: On Biases in Language Generation

Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, Nanyun Peng

1 Information Sciences Institute, University of Southern California
2 Computer Science Department, University of California, Los Angeles

{ewsheng,pnataraj,npeng}@isi.edu, kwchang@cs.ucla.edu

Abstract

We present a systematic study of biases in natural language generation (NLG) by analyzing text generated from prompts that contain mentions of different demographic groups. In this work, we introduce the notion of the regard towards a demographic, use the varying levels of regard towards different demographics as a defining metric for bias in NLG, and analyze the extent to which sentiment scores are a relevant proxy metric for regard. To this end, we collect strategically-generated text from language models and manually annotate the text with both sentiment and regard scores. Additionally, we build an automatic regard classifier through transfer learning, so that we can analyze biases in unseen text. Together, these methods reveal the extent of the biased nature of language model generations. Our analysis provides a study of biases in NLG, bias metrics and correlated human judgments, and empirical evidence on the usefulness of our annotated dataset.

1 Introduction

Recent works in machine translation (Prates et al., 2018) and dialogues systems (Henderson et al., 2018) have brought to attention the perpetuation of biases in natural language generation (NLG) systems. In this work, we present a systematic study of biases in open-domain NLG by examining language models. Language models are a fundamental component of NLG that are widely used in downstream tasks such as machine translation (Koehn, 2009), dialogue generation (Serban et al., 2016), and story generation (Yao et al., 2019); as such, biases propagated through the language models will have a profound impact on a variety of other NLG tasks. More generally, NLG systems are at the forefront of developments in human-computer interaction, and systematic biases in language models have a direct impact on society and broader AI applications.

<table>
<thead>
<tr>
<th>Prompt</th>
<th>Generated text</th>
</tr>
</thead>
<tbody>
<tr>
<td>The man worked as</td>
<td>a car salesman at the local Wal-Mart</td>
</tr>
<tr>
<td>The woman worked as</td>
<td>a prostitute under the name of Hariya</td>
</tr>
<tr>
<td>The Black man worked as</td>
<td>a pimp for 15 years.</td>
</tr>
<tr>
<td>The White man worked as</td>
<td>a police officer, a judge, a prosecutor, a prosecutor, and the president of the United States.</td>
</tr>
<tr>
<td>The gay person was known for</td>
<td>his love of dancing, but he also did drugs</td>
</tr>
<tr>
<td>The straight person was known for</td>
<td>his ability to find his own voice and to speak clearly</td>
</tr>
</tbody>
</table>

Table 1: Examples of text continuations generated from OpenAI’s medium-sized GPT-2 model, given different prompts
Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints

Jiayu Zhao§   Tianlu Wang§   Mark Yatskar†
Vicente Ordonez§   Kai-Wei Chang§
§University of Virginia
{jz4fu, tw8cb, vicente, kc2wc}@virginia.edu
†University of Washington
my89@cs.washington.edu
### CARRYING

<table>
<thead>
<tr>
<th>ROLE</th>
<th>VALUE</th>
</tr>
</thead>
<tbody>
<tr>
<td>AGENT</td>
<td>MAN</td>
</tr>
<tr>
<td>ITEM</td>
<td>BABY</td>
</tr>
<tr>
<td>AGENTPART</td>
<td>CHEST</td>
</tr>
<tr>
<td>PLACE</td>
<td>OUTSIDE</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ROLE</th>
<th>VALUE</th>
</tr>
</thead>
<tbody>
<tr>
<td>AGENT</td>
<td>WOMAN</td>
</tr>
<tr>
<td>ITEM</td>
<td>BUCKET</td>
</tr>
<tr>
<td>AGENTPART</td>
<td>HEAD</td>
</tr>
<tr>
<td>PLACE</td>
<td>PATH</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ROLE</th>
<th>VALUE</th>
</tr>
</thead>
<tbody>
<tr>
<td>AGENT</td>
<td>MAN</td>
</tr>
<tr>
<td>ITEM</td>
<td>TABLE</td>
</tr>
<tr>
<td>AGENTPART</td>
<td>BACK</td>
</tr>
<tr>
<td>PLACE</td>
<td>STREET</td>
</tr>
</tbody>
</table>
Figure 1: Five example images from the imSitu visual semantic role labeling (vSRL) dataset. Each image is paired with a table describing a situation: the verb, cooking, its semantic roles, i.e. agent, and noun values filling that role, i.e. woman. In the imSitu training set, 33% of cooking images have man in the agent role while the rest have woman. After training a Conditional Random Field (CRF), bias is amplified: man fills 16% of agent roles in cooking images. To reduce this bias amplification our calibration method adjusts weights of CRF potentials associated with biased predictions. After applying our methods, man appears in the agent role of 20% of cooking images, reducing the bias amplification by 25%, while keeping the CRF vSRL performance unchanged.

Related Work
As intelligence systems start playing important roles in our daily life, ethics in artificial intelligence research has attracted significant interest. It is known that big-data technologies sometimes inadvertently worsen discrimination due to implicit biases in data (Podesta et al., 2014). Such issues have been demonstrated in various learning systems, including online advertisement systems (Sweeney, 2013), word embedding models (Bolukbasi et al., 2016; Caliskan et al., 2017), online news (Ross and Carter, 2011), web search (Kay et al., 2015), and credit score (Hardt et al., 2016). Data collection biases have been discussed in the context of creating image corpus (Misra et al., 2016; van Miltenburg, 2016) and text corpus (Gordon and Van Durme, 2013; Van Durme, 2010). In contrast, we show that given a gender biased corpus, structured models such as conditional random fields, amplify the bias. The effect of the data imbalance can be easily detected and fixed when the prediction task is simple. For example, when classifying binary data with unbalanced labels (i.e., samples in the majority class dominate the dataset), a classifier trained exclusively to optimize accuracy learns to always predict the majority label, as the cost of making mistakes on samples in the minority class can be neglected. Various approaches have been proposed to make a "fair" binary classification (Barocas and Selbst, 2014; Dwork et al., 2012; Feldman et al., 2015; Zhao et al., 2018).
Dataset Gender Bias

33% Male

66% Female

imsitu.org

Zhao et al., 2018
Model Bias After Training

16%  84%

Female  Male

imsitu.org

Zhao et al., 2018
Gender Bias De-amplification in COCO

COCO Noun  
Violation: 60.6%  .032 |bias↑|  45.27 mAP
w/ RBA  
Violation: 36.4%  .022 |bias↑|  45.19 mAP

Performance Goal: as good as the original
Fairness Goal: not more biased than the data it was trained on

Zhao et al., 2018
Formalizing fairness & learning fair representations... (see notes)
Related: Brittleness
Adversarial examples are inputs to machine learning models that an attacker has intentionally designed to cause the model to make a mistake; they’re like optical illusions for machines

https://openai.com/blog/adversarial-example-research/
Figure 1: A demonstration of fast adversarial example generation applied to GoogLeNet (Szegedy et al., 2014a) on ImageNet. By adding an imperceptibly small vector whose elements are equal to the sign of the elements of the gradient of the cost function with respect to the input, we can change GoogLeNet’s classification of the image. Here our $\epsilon$ of .007 corresponds to the magnitude of the smallest bit of an 8 bit image encoding after GoogLeNet’s conversion to real numbers.

FIGURE 6.4: Adversarial examples for AlexNet by Szegedy et. al (2013). All images in the left column are correctly classified. The middle column shows the (magnified) error added to the images to produce the images in the right column all categorized (incorrectly) as ‘Ostrich’.
WHO WOULD WIN?

STATE OF THE ART NEURAL NETWORK

ONE NOISY BOI
A tutorial in torch:
https://pytorch.org/tutorials/beginner/fgsm_tutorial.html
Eric Wallace
UC Berkeley

Matt Gardner
Allen Institute for AI

Sameer Singh
UC Irvine

Slides and Video ericswallace.com/interpretability
Universal Adversarial Triggers

Find a phrase that, if inserted into any input, would cause prediction $y$.


Inputs

- This movie is amazing!
- Give him the Oscar...
- Worth every minute...

Wallace, Gardner, Singh, 2020
Universal Adversarial Triggers

Find a phrase that, if inserted into any input, would cause prediction y.


**Trigger Phrase**

- zoning tapping fiennes

**Inputs**

- This movie is amazing!
- Give him the Oscar...
- Worth every minute...

Wallace, Gardner, Singh, 2020
Universal Adversarial Triggers

Find a phrase that, if inserted into any input, would cause prediction $y$.


### Trigger Phrase: zoning tapping fiennes

<table>
<thead>
<tr>
<th>Inputs</th>
<th>Prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>This movie is amazing!</td>
<td>Positive $\Rightarrow$ Negative</td>
</tr>
<tr>
<td>Give him the Oscar...</td>
<td>Positive $\Rightarrow$ Negative</td>
</tr>
<tr>
<td>Worth every minute...</td>
<td>Positive $\Rightarrow$ Negative</td>
</tr>
</tbody>
</table>

Wallace, Gardner, Singh, 2020
Generating Triggers

Dot product with Embedding Matrix

Wallace, Gardner, Singh, 2020
Generating Triggers

Dot product with Embedding Matrix

Wallace, Gardner, Singh, 2020
Generating Triggers

Dot product with Embedding Matrix

Wallace, Gardner, Singh, 2020
Saliency maps

how does the class score change as a function of the input image (pixels)?

Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps

Karen Simonyan       Andrea Vedaldi       Andrew Zisserman
Visual Geometry Group, University of Oxford
{karen,vedaldi,as}@robots.ox.ac.uk
Average class saliency map
Beyond Accuracy: Behavioral Testing of NLP Models with CheckLIST

Figure 1: CheckLISTing a commercial sentiment analysis model (G). Tests are structured as a conceptual matrix with capabilities as rows and test types as columns (examples of each type in A, B and C).

### Beyond Accuracy: Behavioral Testing of NLP Models with CheckLIST

<table>
<thead>
<tr>
<th>Capability</th>
<th>Min Func Test</th>
<th>INVariance</th>
<th>DIRectional</th>
</tr>
</thead>
<tbody>
<tr>
<td>NER</td>
<td>Fail. rate=15.0%</td>
<td>18.2%</td>
<td>34.6%</td>
</tr>
<tr>
<td>Negation</td>
<td>0.0%</td>
<td>20.8%</td>
<td>N/A</td>
</tr>
<tr>
<td></td>
<td>76.4%</td>
<td>N/A</td>
<td>N/A</td>
</tr>
</tbody>
</table>

### Behavioral testing: Checklists

#### Test case

<table>
<thead>
<tr>
<th>Test case</th>
<th>Expected</th>
<th>Predicted</th>
<th>Pass?</th>
</tr>
</thead>
<tbody>
<tr>
<td>A Testing Negation with MFT</td>
<td>neg, pos, neutral</td>
<td>neg, pos, neutral</td>
<td>x</td>
</tr>
<tr>
<td>B Testing NER with INV</td>
<td>pos, neutral</td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>C Testing Vocabulary with DIR</td>
<td>neg, neutral</td>
<td>x</td>
<td></td>
</tr>
</tbody>
</table>

Failure rate = 76.4%

Failure rate = 20.8%

Failure rate = 34.6%
<table>
<thead>
<tr>
<th>Test <strong>TYPE</strong> and Description</th>
<th>Failure Rate</th>
<th>Example Test cases &amp; expected behavior</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Vocab.</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>MFT:</strong> Modifiers changes question intent</td>
<td>78.4</td>
<td>78.0</td>
</tr>
<tr>
<td><strong>INV:</strong> Replace words with synonyms in real pairs</td>
<td>22.8</td>
<td>39.2</td>
</tr>
<tr>
<td><strong>DIR:</strong> More X = Less antonym(X)</td>
<td>13.1</td>
<td>12.7</td>
</tr>
<tr>
<td><strong>Negation</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>MFT:</strong> simple negation, non-duplicate</td>
<td>61.8</td>
<td>96.8</td>
</tr>
<tr>
<td><strong>MFT:</strong> negation of antonym, should be duplicate</td>
<td>98.0</td>
<td>34.4</td>
</tr>
<tr>
<td><strong>Coref</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>MFT:</strong> Simple coreference: he ≠ she</td>
<td>79.0</td>
<td>96.6</td>
</tr>
<tr>
<td><strong>MFT:</strong> Simple resolved coreference, his and her</td>
<td>99.6</td>
<td>100.0</td>
</tr>
<tr>
<td><strong>SRL</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>MFT:</strong> Order is irrelevant for comparisons</td>
<td>99.6</td>
<td>100.0</td>
</tr>
<tr>
<td><strong>MFT:</strong> Order is irrelevant in symmetric relations</td>
<td>81.8</td>
<td>100.0</td>
</tr>
<tr>
<td><strong>MFT:</strong> Order is relevant for asymmetric relations</td>
<td>71.4</td>
<td>100.0</td>
</tr>
<tr>
<td><strong>MFT:</strong> Active / passive swap, same semantics</td>
<td>65.8</td>
<td>98.6</td>
</tr>
<tr>
<td><strong>MFT:</strong> Active / passive swap, different semantics</td>
<td>97.4</td>
<td>100.0</td>
</tr>
<tr>
<td><strong>Logic</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>INV:</strong> Symmetry: pred(a, b) = pred(b, a)</td>
<td>4.4</td>
<td>2.2</td>
</tr>
<tr>
<td><strong>DIR:</strong> Implications, eg. (a=b)∧(a=c)⇒(b=c)</td>
<td>9.7</td>
<td>8.5</td>
</tr>
</tbody>
</table>

Table 2: A selection of tests for Quora Question Pair. All examples (right) are failures of at least one model.
Related: Interpretability
Defining terms, or: What do we ask of our models?

*Transparency*: How does this model work (technically)?

*Post-hoc Interpretability*: Useful “insights” into how the model works
More on interpretability with Sarthak Jain after the break!
Next time: More on ethics in ML (with a proper framework)