Bias in Text

Rachel Rudinger
JHU CLSP Seminar
December 1, 2017
For a system to successfully read and understand natural language texts, it must have knowledge of the world.

Knowledge about the world is encoded in vast quantities of natural language text.
We use words to talk about the world. Therefore to understand what words mean, we must have a prior explication of how we view the world.

—J. R. Hobbs, 1987

Do not make your contribution more informative than is required.

—H. P. Grice, 1975

...text corpora contain recoverable and accurate imprints of our historic biases, whether morally neutral as towards insects or flowers, problematic as towards race or gender, or even simply veridical, reflecting the status quo distribution of gender with respect to careers or first names.

—Caliskan et al., 2016

But reporting bias!

We can learn about the world from text!

...and implicit human biases!
1) AI systems need to acquire knowledge about the world to operate.
2) It is prohibitively expensive to encode this information by hand.
3) Natural language texts encode information about the world.
4) We have LOTS of text data. (internet, books, newspapers, magazines, blogs, wikipedia...)
5) CONCLUSION: Let’s build large-scale systems that automatically extract knowledge from these sources.

“knowledge acquisition bottleneck”

Types of Knowledge
- Commonsense
- Generic
- Specific/Factoid
- Frames
- Scripts
- ...

Systems
- Knext
- TextRunner
- NELL
- Whirl
- Lore
- ...

We can learn about the world from text!
Observed:  Text about an event in the world.

Unobserved:  The world.

Inference Problem:  \( P(\text{WORLD} \mid \text{TEXT}) \)
A friend and I went to a steakhouse last Saturday night. As the server was taking our order...

**Rules of Cooperative Conversation** (Grice)

**BLOG ENTRY**

A friend and I went to a steakhouse last Saturday night. As the server was taking our order...

\[ P(\text{WORLD} | \text{TEXT}, ???) \]
Grice’s Maxims of Cooperative Conversation

Maxim of Quantity

1. Make your contribution as informative as is required (for the current purposes of the exchange).
2. Do not make your contribution more informative than is required.

A: What are you doing?
B: Reading a paper for class.
B1: Breathing, and occasionally blinking. [flouts 1]
B2: Reading a paper for class, printed on recycled paper but poorly stapled. [flouts 2]
Maxim of Quality

1. Do not say what you believe to be false.
2. Do not say that for which you lack adequate evidence.

A: How long will this flight to Chicago take?
B: About two hours.
B1: About four days. [flouts 1]
B2: Two hours, four minutes, and six-point-two seconds. [flouts 2]
Grice’s Maxims of Cooperative Conversation

Maxim of Relation
Be relevant.

Grice’s Maxims of Cooperative Conversation

Maxim of Manner

1. Avoid obscurity of expression.
2. Avoid ambiguity.
3. Be brief (avoid unnecessary prolixity).
4. Be orderly.
**Reporting Bias and Knowledge Acquisition**

Gordon and Van Durme, 2013

**KNEXT**, a knowledge-from-text acquisition system.

Example: *The man put his hands in his pockets.* → A person may have hands.

Problem: **KNEXT** finds textual evidence for both a heliocentric model (*The earth revolves around the sun*) and a geocentric model (*The sun revolves around the earth*). Which to trust?

Solution(?): An **inductive** approach to textual evidence.

I.e., More textual evidence for heliocentric (*n=107*) than geocentric (*n=50*),

**THUS:**

\[ p(\text{heliocentric}) > p(\text{geocentric}) \]
If fact $X$ appears more frequently in text than fact $Y$, does it follow that $p(X)$ is greater than $p(Y)$?

---

**Table 1: N-gram frequencies for (his|her|my|your) <body part> and the number of times Knext learns A <body part> may pertain to a person. Plurals are included when appropriate.**

<table>
<thead>
<tr>
<th>Body Part</th>
<th>Teraword</th>
<th>Knext</th>
<th>Body Part</th>
<th>Teraword</th>
<th>Knext</th>
</tr>
</thead>
<tbody>
<tr>
<td>Head</td>
<td>18,907,427</td>
<td>1,004,300</td>
<td>Liver</td>
<td>246,937</td>
<td>9,452</td>
</tr>
<tr>
<td>Eye(s)</td>
<td>18,455,030</td>
<td>934,721</td>
<td>Kidney(s)</td>
<td>183,973</td>
<td>3,289</td>
</tr>
<tr>
<td>Arm(s)</td>
<td>6,345,039</td>
<td>399,120</td>
<td>Spleen</td>
<td>47,216</td>
<td>1,568</td>
</tr>
<tr>
<td>Ear(s)</td>
<td>3,543,711</td>
<td>309,708</td>
<td>Pancreas</td>
<td>24,230</td>
<td>1,186</td>
</tr>
<tr>
<td>Brain</td>
<td>3,277,326</td>
<td>144,511</td>
<td>Gallbladder</td>
<td>17,419</td>
<td>991</td>
</tr>
</tbody>
</table>
Reporting Bias and Knowledge Acquisition
Gordon and Van Durme, 2013

If fact $X$ appears more frequently in text than fact $Y$, does it follow that $p(X)$ is greater than $p(Y)$?

<table>
<thead>
<tr>
<th>Word</th>
<th>Teraword</th>
<th>Knext</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spoke</td>
<td>11,577,917</td>
<td>372,042</td>
</tr>
<tr>
<td>Laughed</td>
<td>3,904,519</td>
<td>179,395</td>
</tr>
<tr>
<td>Murdered</td>
<td>2,843,529</td>
<td>16,890</td>
</tr>
<tr>
<td>Inhaled</td>
<td>984,613</td>
<td>5,617</td>
</tr>
<tr>
<td>Breathed</td>
<td>725,034</td>
<td>41,215</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Word</th>
<th>Teraword</th>
<th>Knext</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hugged</td>
<td>610,040</td>
<td>11,453</td>
</tr>
<tr>
<td>Blinked</td>
<td>390,692</td>
<td>21,973</td>
</tr>
<tr>
<td>Was late</td>
<td>368,922</td>
<td>31,168</td>
</tr>
<tr>
<td>Exhaled</td>
<td>168,985</td>
<td>4,052</td>
</tr>
<tr>
<td>Was on time</td>
<td>23,997</td>
<td>14</td>
</tr>
</tbody>
</table>
Reporting Bias and Knowledge Acquisition
Gordon and Van Durme, 2013

If fact $X$ appears more frequently in text than fact $Y$, does it follow that $p(X)$ is greater than $p(Y)$?

<table>
<thead>
<tr>
<th>Type</th>
<th>Miles Travelled</th>
<th>Crashes</th>
<th>Miles/ Crash</th>
<th>Teraword</th>
</tr>
</thead>
<tbody>
<tr>
<td>Car</td>
<td>1,682,671 million</td>
<td>4,341,688</td>
<td>387,562</td>
<td>1,748,832</td>
</tr>
<tr>
<td>Motorcycle</td>
<td>12,401 million</td>
<td>101,474</td>
<td>122,209</td>
<td>269,158</td>
</tr>
<tr>
<td>Airplane</td>
<td>6,619 million</td>
<td>83</td>
<td>79,746,988</td>
<td>603,933</td>
</tr>
</tbody>
</table>
Reporting Bias and Knowledge Acquisition
Gordon and Van Durme, 2013

Five hypotheses about reporting bias:

1. **The more expected something, the less likely people are to convey it as the primary intent of an utterance.**
   - “A blue pencil” ✔
   - “A yellow pencil” ✗

2. **The more value people attach to something, the more likely they are to give information about it, even if the information is unsurprising.**
   - # people killed in a forest fire ✔
   - # deer or chipmunks killed in a forest fire ✗

3. **Conversely, even unusual facts are unlikely to be mentioned if they are trivial.**
   - That someone has a scratch on their left bicep. ✗
   - That someone is pregnant. ✔

4. **Reporting bias varies by literary genre.**
   - E.g., sports magazine (game results) vs Wall Street Journal (stock market events)
   - *Different audiences presumed to know different things.

5. **There are fundamental kinds of lexical and world knowledge that are needed for understanding and inference that don’t get stated in text.**
   - E.g., things children learn before language: an object can’t be in two locations at once, solid objects tend to persist, people have motives, ...
Reporting Bias and Knowledge Acquisition
Gordon and Van Durme, 2013

Tricks for getting around reporting bias: *implicit vs explicit* content.

1. **Presuppositions**
   - ‘Both my legs hurt.’ → *A person normally has two legs.*
   - ‘I forgot the money to buy groceries.’ → *A person may use money to buy things.*

2. **Disconfirmed expectations**
   - ‘Sally crashed her car into a tree but wasn’t hurt.’ → *If a person crashes her car, she may be hurt.*
   - ‘I dropped the glass but it didn’t break.’ → *If a person drops a glass, it will often break.*

3. **Implicit denials**
   - *Explicit statements, pragmatically required to be informative, contain implicit denials that what they’re saying is usually the case.*
   - ‘The tree had no branches.’ → *Trees usually have branches.*
   - ‘Molly handed me a blue pencil.’ → *Probably pencils are not always blue.*
Case Study: VERBPHYSICS
Forbes and Choi, 2017

“While natural language text is a rich source to obtain broad knowledge about the world, compiling trivial commonsense knowledge from unstructured text is a nontrivial feat...

“The key insight is this: there is consistency in the way people describe how they interact with the world, which provides vital clues to reverse engineer the common knowledge shared among people.”

– Forbes and Choi, 2017
Case Study: VERB PHYSICS
Forbes and Choi, 2017

Idea: Some verbs carry implications about the relative physical size, weight, speed, strength, and rigidity of their arguments.

Example:
“*She barged into the stable.*”
→ human SMALLER-TAN stable

Goals:
1. Learn relative (ordered) physical properties of objects.
2. Learn which verbs carry which implications.

Figure 1: An overview of our approach. A verb’s usage in language (top) implies physical relations between objects it takes as arguments. This allows us to reason about properties of specific objects (middle), as well as the knowledge implied by the verb itself (bottom).
Case Study: Scripts

**Scripts** are a type of world knowledge in the form of a *common sequence of events*. (Schank and Abelson, 1977)

Famous example, the RESTAURANT SCRIPT:
E.g., “enter restaurant,” “get seated,” “look at menu,” “place order,” ...
Script Induction

**Script Induction**: Scripts are expensive to hand-code. Learn them automatically from text.

**Evaluation**: One proposed evaluation is the *narrative cloze task* (Chambers and Jurafsky, 2008)

- Identify a “main protagonist” in a document, and the sequence of every event (verb) they participate in.
- Occlude one event in the sequence and predict it.
Prime Minister Tony Blair announced today that he would leave office June 27 after a decade in power in which he sacrificed his popularity to the war in Iraq and struggled at home to improve schools, policing, and hospitals. Labor Party Chancellor Gordon Brown will likely succeed Blair...
Key insight: If evaluating on narrative cloze, don’t use PMI-based script induction models; instead train a language model directly on the event sequences. Big performance boost!

Why this works: Models that generate outputs that look “scripty” do so by penalizing high-frequency events in text (e.g., “say.”)

Cautionary tale: If goal is to learn world statistics, be careful evaluating on text statistics.
A brief detour into computer vision...
Reporting Bias and Vision

Misra et al., 2016

(a) A woman standing next to a bicycle with basket.

(b) A city street filled with lots of people walking in the rain.

(c) A yellow Vespa parked in a lot with other cars.

(d) A store display that has a lot of bananas on sale.


Slides by R. Rudinger, 2017
Reporting Bias and Vision
Misra et al., 2016

Data: Images with human caption labels
(MS COCO, Yahoo Flickr100M)

Direct Approach:
Predict human captions conditioned on image.

Accounting for Reporting Bias:
Let model decide whether an object is *visually* present in the image, and then whether it is *relevant* (i.e., reported in the human caption).

...and implicit human biases!
Implicit Human Bias in Text
Review: Distributional Semantics and Word Embeddings

“You shall know a word by the company it keeps.”
-- J. R. Firth

• “Knowledge from text”
  ≈> Acquiring lexical semantics (word meanings) from text

• **Word Embeddings**
  – Represent each word in a vocabulary as a vector $w \in \mathbb{R}^d$
  – Mathematical properties indicative of semantic properties
    • king – man + woman = queen
    • Semantically similar words are geometrically close in the vector space
  – Trained from word co-occurrences in large amounts of text
  – Ubiquitous in NLP methods and research
Review: Distributional Semantics and Word Embeddings

“You shall know a word by the company it keeps.”

-- J. R. Firth

Summary: Word embeddings exhibit the same implicit biases as humans, as measured by Implicit Association Tests (IATs).
Implicit Association Tests (IATs)

Measuring implicit association between target and attribute concepts with reaction times.

- **Male Names**: Ben, Paul, Daniel, John, Jeffrey
- **Female Names**: Rebecca, Michelle, Emily, Julia, Anna
- **Career Terms**: Career, Corporation, Salary, Office, Professional, Management, Business
- **Family Terms**: Wedding, Marriage, Parents, Relatives, Family, Home, Children

IAT Results Reproduced in Word Embeddings

<table>
<thead>
<tr>
<th>Target words</th>
<th>Attrib. words</th>
<th>Original Finding</th>
<th>Our Finding</th>
</tr>
</thead>
<tbody>
<tr>
<td>Flowers vs insects</td>
<td>Pleasant vs unpleasant</td>
<td>(5) 32 1.35 10^{-8}</td>
<td>25×2 25×2 1.50 10^{-7}</td>
</tr>
<tr>
<td>Instruments vs weapons</td>
<td>Pleasant vs unpleasant</td>
<td>(5) 32 1.66 10^{-10}</td>
<td>25×2 25×2 1.53 10^{-7}</td>
</tr>
<tr>
<td>Eur.-American vs Afr.-American names</td>
<td>Pleasant vs unpleasant</td>
<td>(5) 26 1.17 10^{-5}</td>
<td>32×2 25×2 1.41 10^{-8}</td>
</tr>
<tr>
<td>Eur.-American vs Afr.-American names</td>
<td>Pleasant vs unpleasant from (5)</td>
<td>(7) Not applicable</td>
<td>16×2 25×2 1.50 10^{-4}</td>
</tr>
<tr>
<td>Eur.-American vs Afr.-American names</td>
<td>Pleasant vs unpleasant from (5)</td>
<td>(7) Not applicable</td>
<td>16×2 8×2 1.28 10^{-3}</td>
</tr>
<tr>
<td>Male vs female names</td>
<td>Career vs family</td>
<td>(9) 39k 0.72 &lt; 10^{-2}</td>
<td>8×2 8×2 1.81 10^{-3}</td>
</tr>
<tr>
<td>Math vs arts</td>
<td>Male vs female terms</td>
<td>(9) 28k 0.82 &lt; 10^{-2}</td>
<td>8×2 8×2 1.06 .018</td>
</tr>
<tr>
<td>Science vs arts</td>
<td>Male vs female terms</td>
<td>(10) 91 1.47 10^{-24}</td>
<td>8×2 8×2 1.24 10^{-2}</td>
</tr>
<tr>
<td>Mental vs physical disease</td>
<td>Temporary vs permanent</td>
<td>(23) 135 1.01 10^{-3}</td>
<td>6×2 7×2 1.38 10^{-2}</td>
</tr>
<tr>
<td>Young vs old people’s names</td>
<td>Pleasant vs unpleasant</td>
<td>(9) 43k 1.42 &lt; 10^{-2}</td>
<td>8×2 8×2 1.21 10^{-2}</td>
</tr>
</tbody>
</table>

\[ N = \# \text{ participants} \]
\[ N_T = \# \text{ target words} \]
\[ N_A = \# \text{ attribute words} \]
\[ d = \text{ effect size} \]
\[ p = \text{ p-value} \]
Word Embedding Association Test (WEAT)

X, Y := two sets of target words of same size, e.g., X={Ben, Paul...}, Y={Rebecca, Anna...}
A,B := two sets of association words, e.g., A={career, salary...}, B={home, family, children...}

Test statistic:

\[ s(w, A, B) = \text{mean}_{a \in A} \cos(w, a) - \text{mean}_{b \in B} \cos(w, b) \]
\[ \rightarrow \text{“measures the association of } w \text{ with the attribute.”} \]

\[ s(X, Y, A, B) = \sum_{x \in X} s(x, A, B) - \sum_{y \in Y} s(y, A, B) \]
\[ \rightarrow \text{“measures the differential association of the two sets of target words with the attribute.”} \]

\[ p\text{-value (Permutation test):} \]
\[ \Pr_i[s(X_i, Y_i, A, B) > s(X, Y, A, B)] \]

\[ \text{Effect Size:} \]
\[ \frac{\text{mean}_{x \in X} s(x, A, B) - \text{mean}_{y \in Y} s(y, A, B)}{\text{std-dev}_{w \in X \cup Y} s(w, A, B)} \]
Gender Bias in Word Embedding Correlates with Real-World Gender Bias

Figure 1: Occupation-gender association. Pearson’s correlation coefficient $\rho = 0.90$ with $p$-value $< 10^{-18}$.

Figure 2: Name-gender association. Pearson’s correlation coefficient $\rho = 0.84$ with $p$-value $< 10^{-13}$.

Word Embedding

Factual Association Test (WEFAT)

$W := \text{set of target words, e.g., } W = \{\text{technician, accountant, therapist, mechanic, hairdresser...}\}$

$A, B := \text{two sets of association words, e.g., } A = \{\text{female, woman, girl,...}\}, B = \{\text{male, man, boy...}\}$

Compute statistic:

$s(w, A, B) = \frac{\text{mean}_{a \in A} \cos(\overrightarrow{w}, \overrightarrow{a}) - \text{mean}_{b \in B} \cos(\overrightarrow{w}, \overrightarrow{b})}{\text{std-dev}_{x \in A \cup B} \cos(\overrightarrow{w}, \overrightarrow{x})}$

Each target word has corresponding real-world association $p_w$. E.g., if $w=\text{“mechanic,”}$ then $p_w = \%$ mechanics who are women, according to Bureau of Labor Statistics.

WEFAT: Compute Pearson correlation between $s(w, A, B)$ and $p_w$ over all tested values of $w$ (e.g., occupations).

Conclusions from Word Embedding Study

• Validation of original IAT studies. (Results replicated in different setting.)
• WEAT a possible method for discovering or comparing existing biases.
• IAT for historical populations via historical texts?
• Building AI systems that understand language may inherently carry the biases in that language?
• Automated decision-making technology may behave with prejudice.

Stereotyping in AI and NLP Datasets

Unsupervised word embeddings train over large quantities of raw text.

What about supervised methods in AI/NLP that use artificially constructed datasets to train models for a particular task?
Case Study: Natural Language Inference

Premise: A man is walking his two dogs in the park.
Hypothesis: They are outside.
ENTAILMENT

Premise: A man is walking his two dogs in the park.
Hypothesis: The man is wearing a red shirt.
NEUTRAL

Premise: A man is walking his two dogs in the park.
Hypothesis: The man is driving.
CONTRADICTION
Stanford Natural Language Inference Dataset (SNLI)
Bowman et al., 2015

Flickr30k Dataset
- Five crowd annotators independently caption each image from a 30k photo sample of Flickr.

Natural language inference
- Given a premise, determine whether a hypothesis is true (entailment), false (contradiction), or neither (neutral).

SNLI Dataset
- A crowd annotator writes three new sentences per caption: entailment, neutral, and contradiction.
  - The image caption is called the premise.
  - The elicited inference is called the hypothesis.

Image: Rudinger and May, 2017
Social Bias in Elicited Natural Language Inferences
Rudinger*, May*, and Van Durme, 2017

Measuring strength of association between two terms using estimated pointwise mutual information (PMI):

$$PMI(x, y) = \log \left( \frac{p(x, y)}{p(x)p(y)} \right)$$

Uncover stereotyped associations by querying top-k PMI results:

$$\operatorname{argmax}_y PMI(\text{“woman”}, y)$$
Gender Stereotypes in SNLI

Top hypothesis words (y) by PMI with premise word (x).

Female stereotypes: emotional warmth (affection, smile), “pink collar jobs” (hairdresser), talkativeness (gossiping, chat), sexualized (strippers, bikinis), girl’s activities (selfies, slumber [parties])

Male stereotypes: physical labor (roofers, cowboys, lumberjacks), technical professionals (computers, engineers, supervisors), crime and violence (thieves, guns, fighting)

Top hypothesis words (y) by PMI with premise word (x).
### Explicit Bias from Elicitation in SNLI

...in the form of *harmful stereotypes* and *pejorative language*

<table>
<thead>
<tr>
<th>Premise</th>
<th>Hypothesis (contr.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>An African American man looking at some butchered meat that is hanging from a rack outside a building.</td>
<td>A black man is in jail</td>
</tr>
<tr>
<td>New sport is being played to show appreciation to the kids who can not walk.</td>
<td>People are playing a sport in honor of crippled people.</td>
</tr>
<tr>
<td>Several people, including a shirtless man and a woman in purple shorts which say “P.I.N.K.” on the back, are walking through a crowded outdoor area.</td>
<td>The woman is wearing slutty shorts.</td>
</tr>
<tr>
<td>Adult with red boots and purse walking down the street next to a brink wall.</td>
<td>A whore looking for clients.</td>
</tr>
<tr>
<td>Several Muslim worshipers march towards Mecca.</td>
<td>The Muslims are terrorists.</td>
</tr>
<tr>
<td>A man dressed as a woman and other people stand around tables with checkered tablecloths and a ladder.</td>
<td>The man is a transvestite.</td>
</tr>
</tbody>
</table>
Sources of Bias in SNLI

• Image bias: Distribution of images in Flickr30k
• Caption bias: How original images were captioned by crowdsourcing workers.
• Elicitation bias: How inferences were elicited from crowdsourcing workers.
  – Human biases
  – Crowd worker quality control
  – *Neutral* label invites stereotypic reasoning
  – Explicit bias
Recap

• Text carries information about world that we want to reverse engineer.

• Text frequencies don’t match real-world frequencies because of reporting bias. (Think Grice.)
  – Some work exists on clever ways to get around reporting bias or incorporate it in modeling. (KNEXT, VerbPhysics, Image Captioning...)
  – Don’t conflate text statistics and “world” statistics in evaluation of knowledge acquisition systems. (Script Induction)

• Word embeddings trained on raw text capture implicit human biases as measured by IAT, including gender and race.

• Labeled NLP datasets may contain both implicit and explicit forms of social bias.