Bias in Text

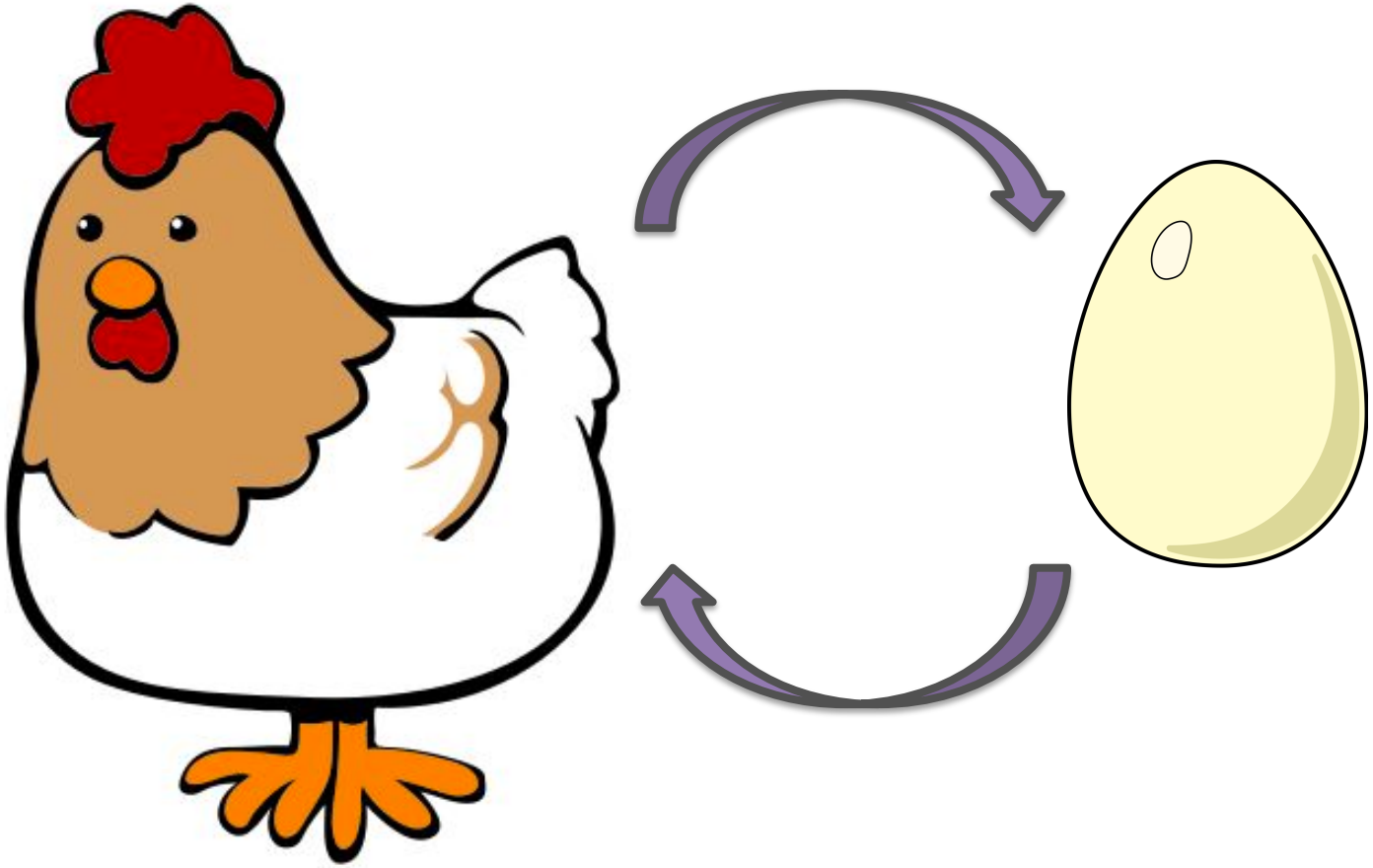
Rachel Rudinger

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 can learn about the world from 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

But reporting bias!

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

...and implicit human biases!

We can learn about the world from text!

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.

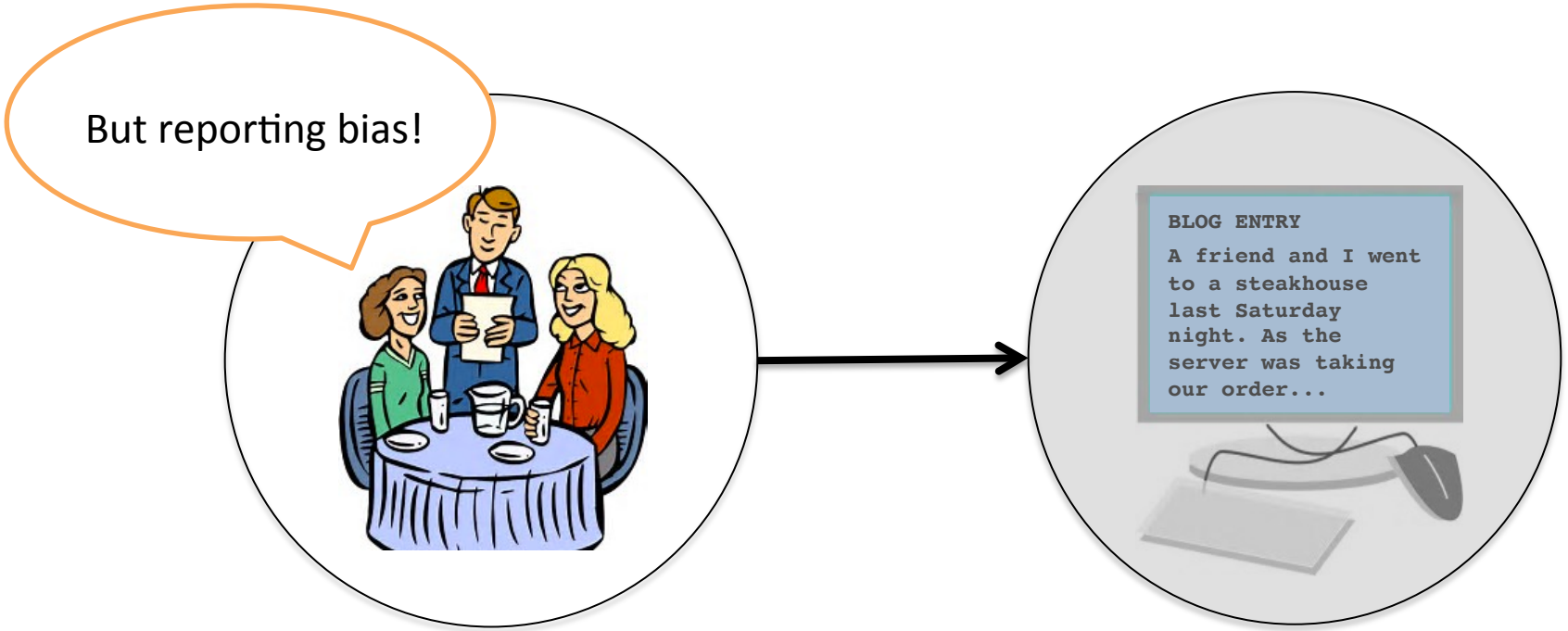
"knowledge acquisition bottleneck"

Types of Knowledge

- Commonsense
- Generic
- Specific/Factoid
- Frames
- Scripts
- ...

Systems

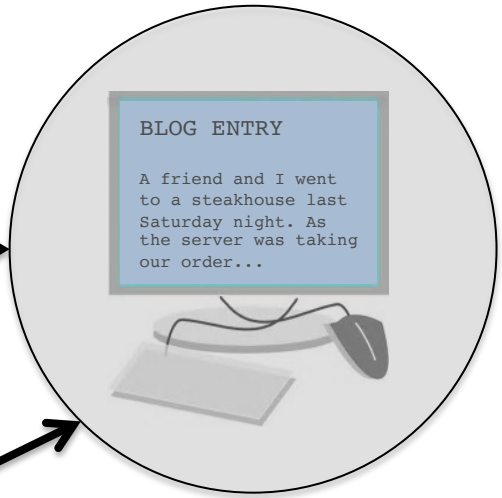
- KNEXT
- TEXTRUNNER
- NELL
- WHIRL
- LORE
- ...



Observed: Text about an event in the world.

Unobserved: The world.

Inference Problem: P(WORLD | TEXT)



Rules of Cooperative Conversation (Grice)

P(WORLD | 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]

Grice, H. Paul, Peter Cole, and Jerry Morgan. "Logic and conversation." 1975

Grice's Maxims of Cooperative Conversation

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, H. Paul, Peter Cole, and Jerry Morgan. "Logic and conversation." 1975

Grice's Maxims of Cooperative Conversation

Maxim of Relation

Be relevant.

Grice, H. Paul, Peter Cole, and Jerry Morgan. "Logic and conversation." 1975

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.

Grice, H. Paul, Peter Cole, and Jerry Morgan. "Logic and conversation." 1975

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(heliocentric) > p(geocentric)

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 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.

Body Part	*Teraword*	*Knext*	*Body Part*	*Teraword*	*Knext*
Head	18,907,427	1,004,300	Liver	246,937	9,452
Eye(s)	18,455,030	934,721	Kidney(s)	183,973	3,289
Arm(s)	6,345,039	399,120	Spleen	47,216	1,568
Ear(s)	3,543,711	309,708	Pancreas	24,230	1,186
Brain	3,277,326	144,511	Gallbladder	17,419	991

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 2: N-gram frequencies for various verbal events and the number of times Knext learns that *A person may* ⟨*x*⟩, including appropriate arguments, e.g., *A person may hug a person*.

Word	*Teraword*	*Knext*	*Word*	*Teraword*	*Knext*
Spoke	11,577,917	372,042	Hugged	610,040	11,453
Laughed	3,904,519	179,395	Blinked	390,692	21,973
Murdered	2,843,529	16,890	Was late	368,922	31,168
Inhaled	984,613	5,617	Exhaled	168,985	4,052
Breathed	725,034	41,215	Was on time	23,997	14

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 3: *Miles Travelled*, *Crashes*, and *Miles/Crash* are for travel in the United States in 2006 [31]. A plane crash is considered any event in which the plane was damaged. *Teraword* results are for the patterns *car* (*crash*|*accident*), *motorcycle* (*crash*|*accident*), and (*airplane*|*plane*) (*crash*|*accident*).

Type	*Miles Travelled*	*Crashes*	*Miles/Crash*	*Teraword*
Car	1,682,671 million	4,341,688	387,562	1,748,832
Motorcycle	12,401 million	101,474	122,209	269,158
Airplane	6,619 million	83	79,746,988	603,933

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: VERBPHYSICS

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-THAN stable

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

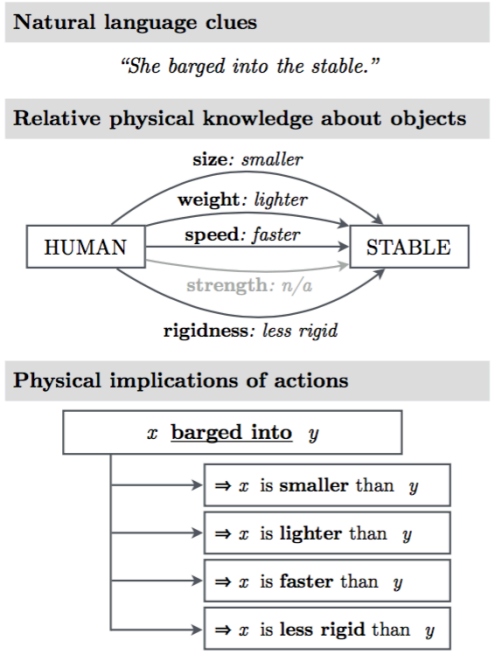


Figure 1: An overview of our approach. A verb's usage in language (top) implies physical relations between objects it takes as arguments. This allows us to reason about properties of specific objects (middle), as well as the knowledge implied by the verb itself (bottom).

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,” ...

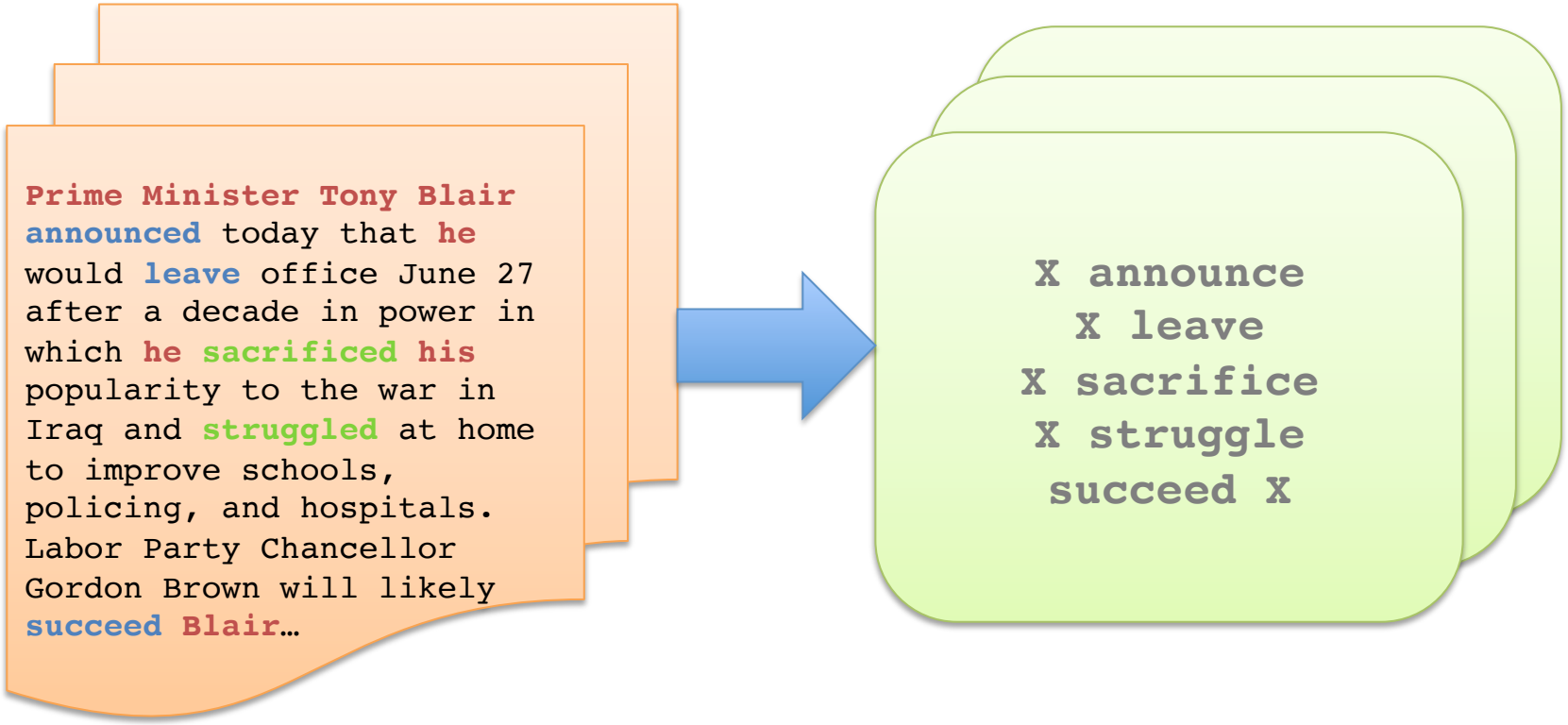
Script Induction

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.

Script Induction



Script Induction as Language Modeling

Rudinger et al., 2015

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.

	Human Label	Visual Label
Bicycle	✓	✓

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

	Human Label	Visual Label
Bicycle	✗	✓

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

	Human Label	Visual Label
Yellow	✓	✓

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

	Human Label	Visual Label
Yellow	✗	✓

Misra I, Lawrence Zitnick C, Mitchell M, Girshick R. Seeing through the human reporting bias: Visual classifiers from noisy human-centric labels. CVPR, 2016.

Reporting Bias and Vision

Misra et al., 2016

Data: Images with human caption labels

(MS COCO, Yahoo Flickr100M)

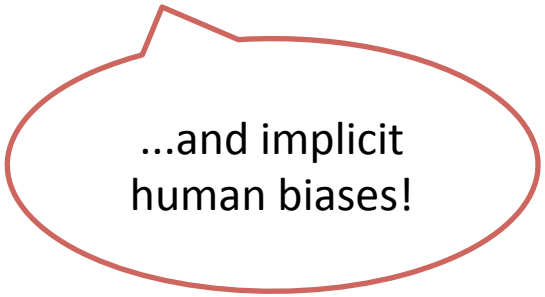
Direct Approach:

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).

Misra I, Lawrence Zitnick C, Mitchell M, Girshick R. Seeing through the human reporting bias: Visual classifiers from noisy human-centric labels. CVPR, 2016.


...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"

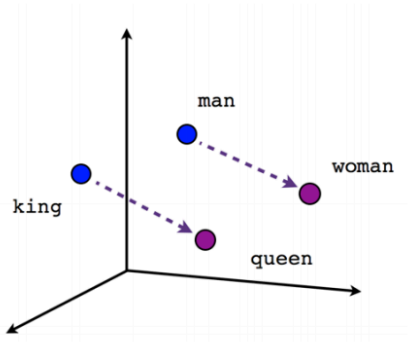
  ≈> Acquiring lexical semantics (word meanings) from text

- **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

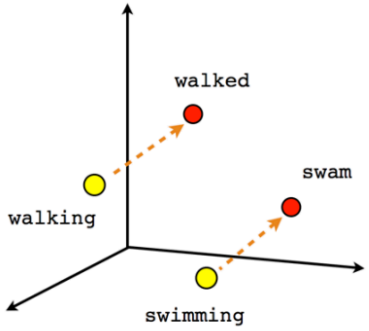
Review: Distributional Semantics and Word Embeddings

"You shall know a word by the company it keeps."

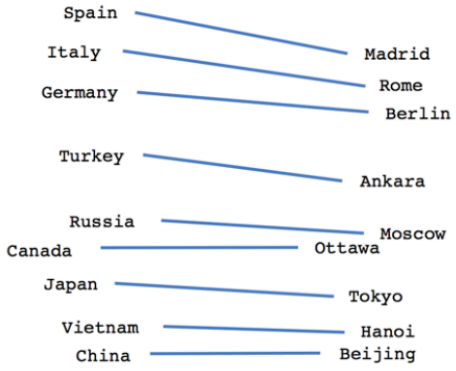
-- J. R. Firth



Male-Female



Verb tense



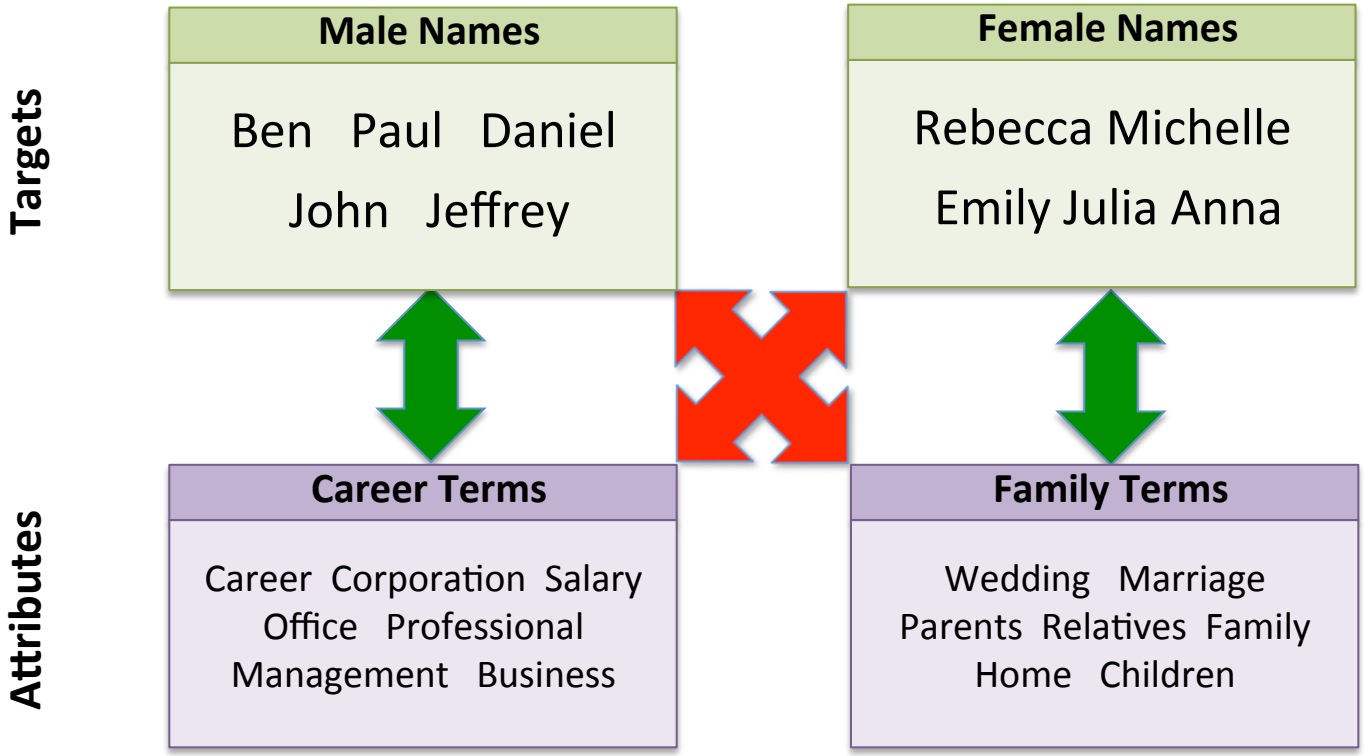
Country-Capital

Semantics derived automatically from language corpora necessarily contain human biases. Aylin Caliskan, Joanna J. Bryson, Arvind Narayanan. Science, 2017.

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.



Greenwald, Anthony G., Debbie E. McGhee, and Jordan LK Schwartz. "Measuring individual differences in implicit cognition: the implicit association test." Journal of personality and social psychology 74.6 (1998): 1464.

IAT Results Reproduced in Word Embeddings

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

N = # participants
N_T = # target words
N_A = # attribute words
d = effect size
p = p-value

Semantics derived automatically from language corpora contain human-like biases. Aylin Caliskan, Joanna J. Bryson, Arvind Narayanan. Science, 2017.

Rudinger 2017

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:

$$s(w, A, B) = \text{mean}_{a \in A} \cos(\vec{w}, \vec{a}) - \text{mean}_{b \in B} \cos(\vec{w}, \vec{b})$$

→ "measures the association of *w* with the attribute."

$$s(X, Y, A, B) = \sum_{x \in X} s(x, A, B) - \sum_{y \in Y} s(y, A, B)$$

→ "measures the differential association of the two sets of target words with the attribute."

p-value **(Permutation test):**

$$\Pr_i[s(X_i, Y_i, A, B) > s(X, Y, A, B)]$$

Effect Size:

$$\frac{\text{mean}_{x \in X} s(x, A, B) - \text{mean}_{y \in Y} s(y, A, B)}{\text{std-dev}_{w \in X \cup Y} s(w, A, B)}$$

Semantics derived automatically from language corpora contain human-like biases. Aylin Caliskan, Joanna J. Bryson, Arvind Narayanan. Science, 2017.

Rudinger 2017

Gender Bias in Word Embedding Correlates with Real-World Gender Bias

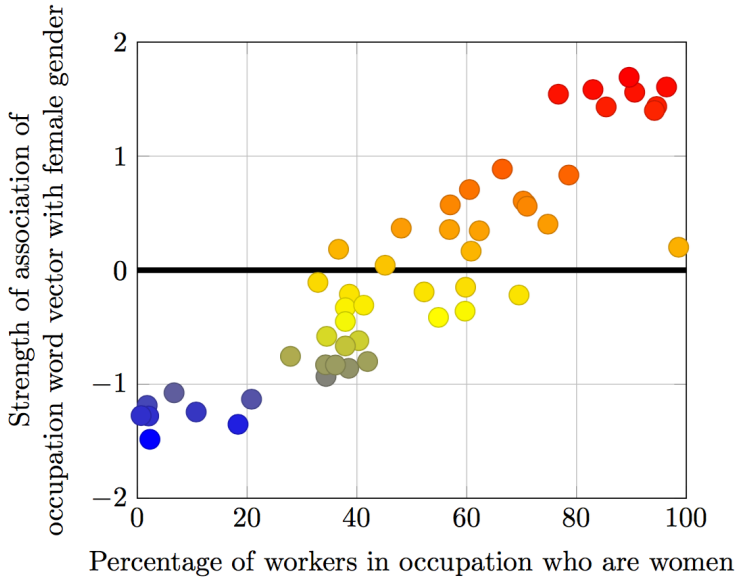


Figure 1: Occupation-gender association. Pearson's correlation coefficient $\rho = 0.90$ with p-value $< 10^{-18}$.

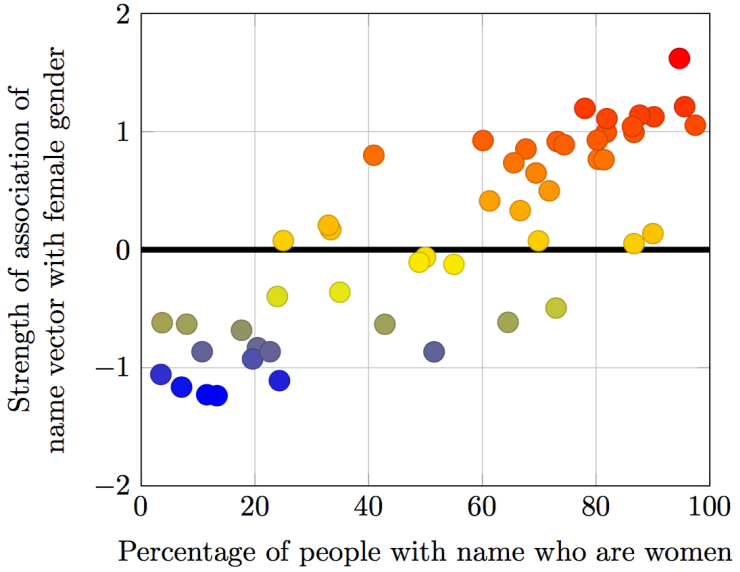


Figure 2: Name-gender association. Pearson's correlation coefficient $\rho = 0.84$ with p-value $< 10^{-13}$.

Semantics derived automatically from language corpora contain human-like biases. Aylin Caliskan, Joanna J. Bryson, Arvind Narayanan. Science, 2017.

Rudinger 2017

Word Embedding Factual Association Test (WEFAT)

W := set of taget words, e.g., W={technician, accountant, therapist, mechanic, hairdresser...}

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

Compute statistic:

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

Each target word has corresponding real-world association p_w. E.g., if *w*="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).

Semantics derived automatically from language corpora contain human-like biases. Aylin Caliskan, Joanna J. Bryson, Arvind Narayanan. Science, 2017.

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.

Semantics derived automatically from language corpora contain human-like biases. Aylin Caliskan, Joanna J. Bryson, Arvind Narayanan. Science, 2017.

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

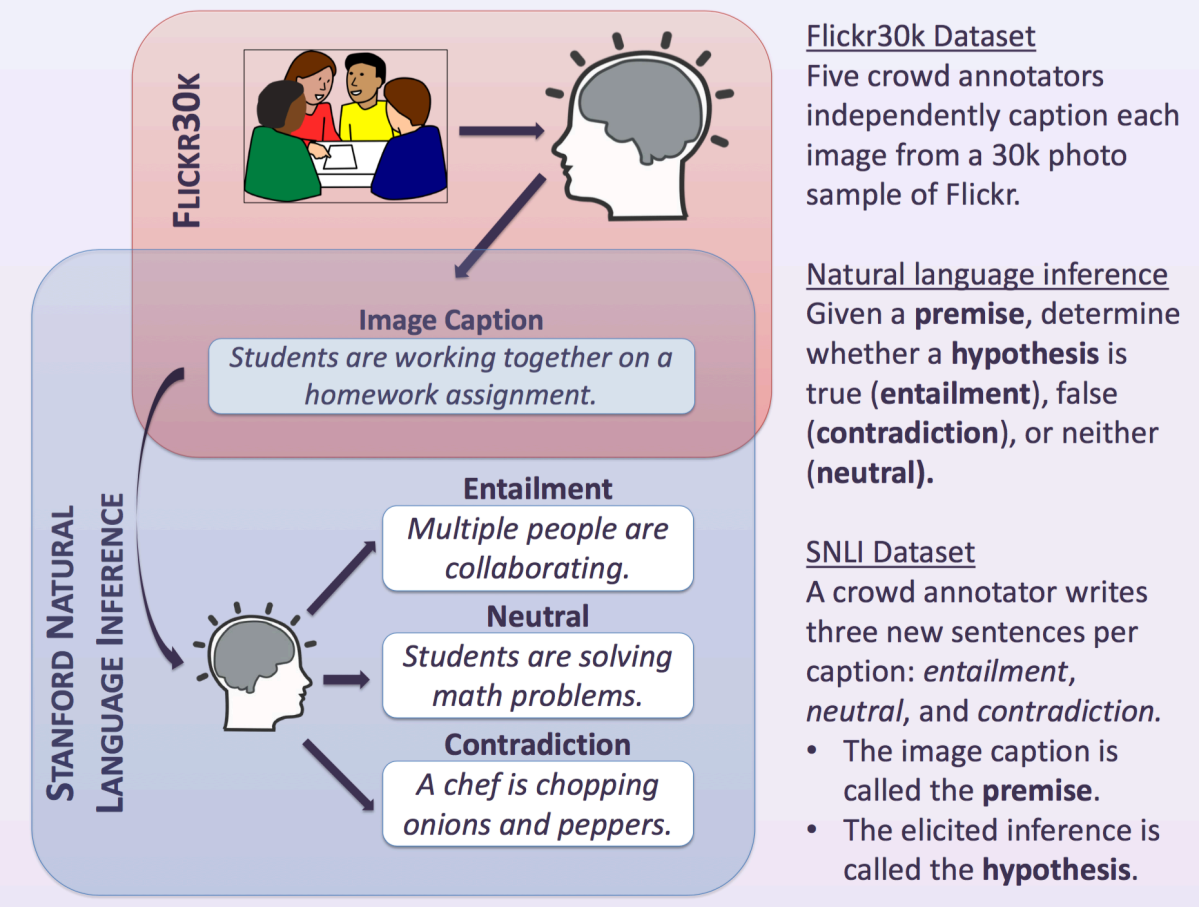


Image: Rudinger and May, 2017

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. EMNLP, 2015.

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:

$$\underset{y}{\text{argmax}}\ \text{PMI}(\text{“woman”}, y)$$

Rachel Rudinger*, Chandler May*, and Benjamin Van Durme. "Social Bias in Elicited Natural Language Inferences.” Workshop on Ethics in NLP. EACL, 2017.

Gender Stereotypes in SNLI

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

ENTAILMENT	**women**	scarves† ladies‡ womens‡ wemon‡ females‡ woman‡ affection dressing chat smile†
	men	mens‡ guys‡ guitars cowboys† remove dock dudes workers‡ computers‡ boxers
	girls	cheerleaders‡ females‡ girl‡ dancers children‡ smile practice dance‡ outfits laughing
	boys	males‡ children‡ boy‡ kids‡ four‡ fighting† exercise play‡ pose fun
NEUTRAL	**women**	actresses‡ gossip‡ husbands‡ womens‡ nuns† bridesmaids† gossiping‡ ladies‡ strippers purses
	men	lumberjacks mens‡ supervisors thieves‡ homosexual roofers reminisce† contractors groomsmen engineers‡
	girls	fifteen‡ slumber† gymnasts‡ cheerleading‡ bikinis† sisters‡ cheerleaders‡ daughters‡ selfies† teenage‡
	boys	skip† sons‡ brothers‡ twins‡ muddy trunks† males† league‡ cards recess†
CONTRADICTION	**women**	womens† wemon bikinis‡ ladies‡ towels females‡ politics dresses‡ discussing men‡
	men	dudes mens‡ motel‡ gossip surfboards wives caps sailors floors helmets
	girls	sking‡ boys‡ 50 brothers sisters dolls† pose opposite phones hopscotch
	boys	girls‡ sisters‡ sons bunk homework† males coats beds† guns professional

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*),

Social Bias in Elicited Natural Language Inferences

Rudinger*, May*, and Van Durme, 2017

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

Rachel Rudinger*, Chandler May*, and Benjamin Van Durme. "Social Bias in Elicited Natural Language Inferences." Workshop on Ethics in NLP. EACL, 2017.

Explicit Bias from Elicitation in SNLI

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

Premise: An African American man looking at some butchered meat that is hanging from a rack outside a building. **Hypothesis (contr.)**: A black man is in jail
Premise: New sport is being played to show appreciation to the kids who can not walk. **Hypothesis (entail.)**: People are playing a sport in honor of crippled people.
Premise: 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. **Hypothesis (entail.)**: The woman is wearing slutty shorts.
Premise: adult with red boots and purse walking down the street next to a brink wall. **Hypothesis (neutr.)**: A whore looking for clients.
Premise: Several Muslim worshipers march towards Mecca. **Hypothesis (neutr.)**: The Muslims are terrorists.
Premise: A man dressed as a woman and other people stand around tables with checkered tablecloths and a ladder. **Hypothesis (neutr.)**: The man is a transvestite.

Rachel Rudinger*, Chandler May*, and Benjamin Van Durme. "Social Bias in Elicited Natural Language Inferences." Workshop on Ethics in NLP. EACL, 2017.

Sources of Bias in SNLI

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

Rachel Rudinger*, Chandler May*, and Benjamin Van Durme. "Social Bias in Elicited Natural Language Inferences." Workshop on Ethics in NLP. EACL, 2017.

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.