

Lecture Notes for **Neural Networks and Machine Learning**



A Practical Example of
Ethically Aware NLP



Logistics and Agenda

- Logistics
 - Post about your preferred lecture discussion or paper summary if you haven't yet!
- Last Time:
 - Ethical Guidelines
 - Case Studies
- Agenda
 - Paper Presentation:
 - ◆ Data (dis)contents
 - NLP Review
 - Extended Example



Paper Presentation

Data and its (dis)contents: A survey of dataset development and use in machine learning research

Amanda Lee Poolladi
Department of Linguistics
University of Washington

Indira Deborah Raji
Mozilla Foundation

Emily M. Bender
Department of Linguistics
University of Washington

Emily Denton
Google Research

Alex Hanna
Google Research

Abstract

Datasets have played a foundational role in the advancement of machine learning research. They form the basis for the models we design and deploy as well as our primary medium for benchmarking and evaluation. Furthermore, the ways in which we collect, construct and share these datasets inform the kinds of problems the field pursues and the methods employed in algorithm development. However, recent work from a breadth of perspectives has revealed the limitations of predominant practices in dataset collection and use. In this paper, we survey the many *researchers* raised about the way we collect and use data in machine learning and advocate that a more cautious and thorough understanding of data is necessary to address several of the practical and ethical issues of the field.

1 Introduction

The importance of datasets for machine learning research cannot be overstated. Datasets have been among the limiting factor for algorithmic development and scientific progress [Balevy et al., 2009, Sun et al., 2017], and a select few benchmark datasets have shaped some of the most significant developments in the field. Benchmark datasets have also played a critical role in informing the goals, values, and research agendas of the machine learning community [Bender and Miller, 2019].

In recent years, machine learning systems have been reported to achieve ‘super-human’ performance when evaluated on benchmark datasets, such as the GLUE benchmark for English natural understanding [Wang et al., 2019]. However, recent work has surfaced the shortcomings of such datasets as meaningful tests of human-like reasoning ability reveals that this appearance of progress may rest on faulty foundations.

As the machine learning field increasingly turned to data-driven approaches, the sort of skilled and methodical human annotation applied to dataset collection practices in earlier years was supplanted as ‘cheap and responsive to experts’, and a new trend of universal collection of increasingly large amounts of data from the Web, discussion increased reliance on non-expert crowdworkers, was seen as a boon to machine learning [Balevy et al., 2009, Deng et al., 2009]. These data practices tend to abstract away the human labor, subjective judgments and biases, and contingent contexts involved in dataset production. However, these details are important for assessing whether and how a dataset might be useful for a particular application, for enabling better, more systematic meta-analysis, and for acknowledging the significant difficulty required in constructing useful datasets. Enormous scale has been mobilized as beneficial to generality and objectivity, but all datasets have limitations and biases [Boyd and Crawford, 2012].

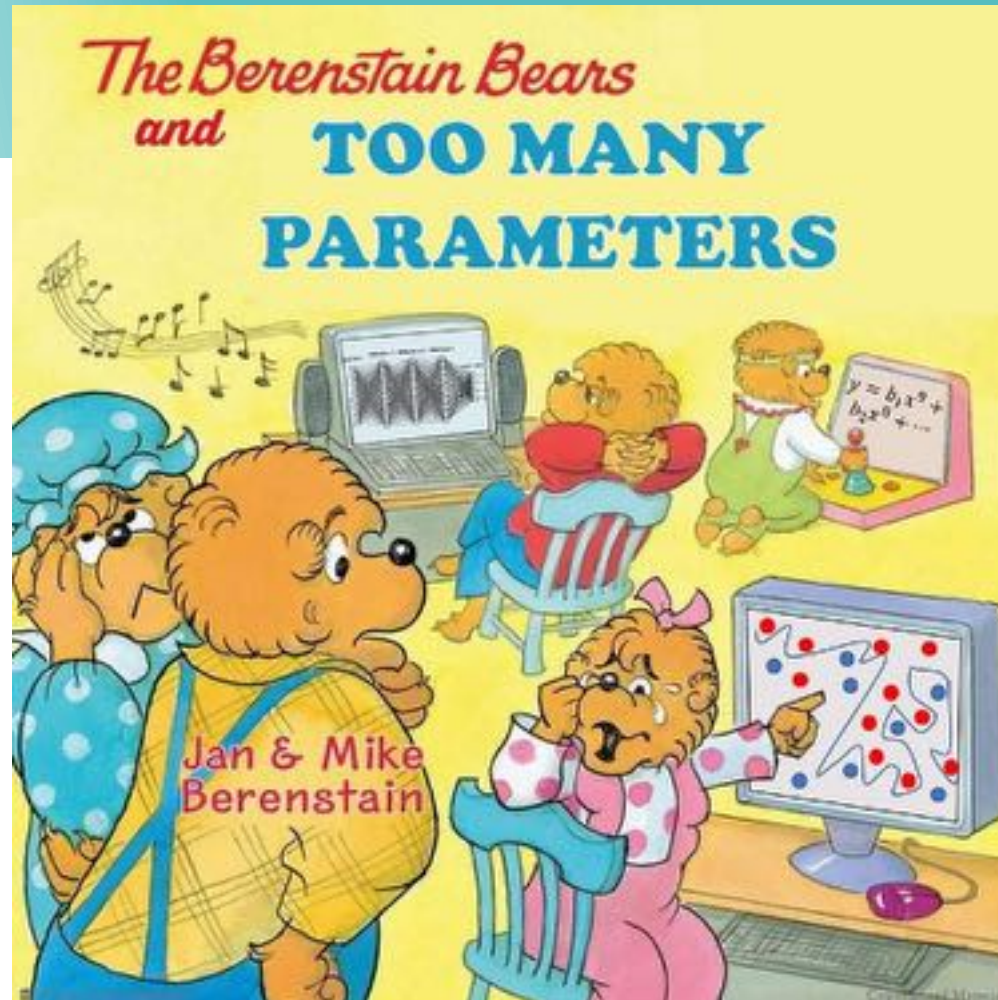
NeurIPS 2020 Workshop: ML Retrospectives, Surveys & Meta-analyses (ML-RSA), Virtual.

<https://arxiv.org/pdf/2012.05345.pdf>

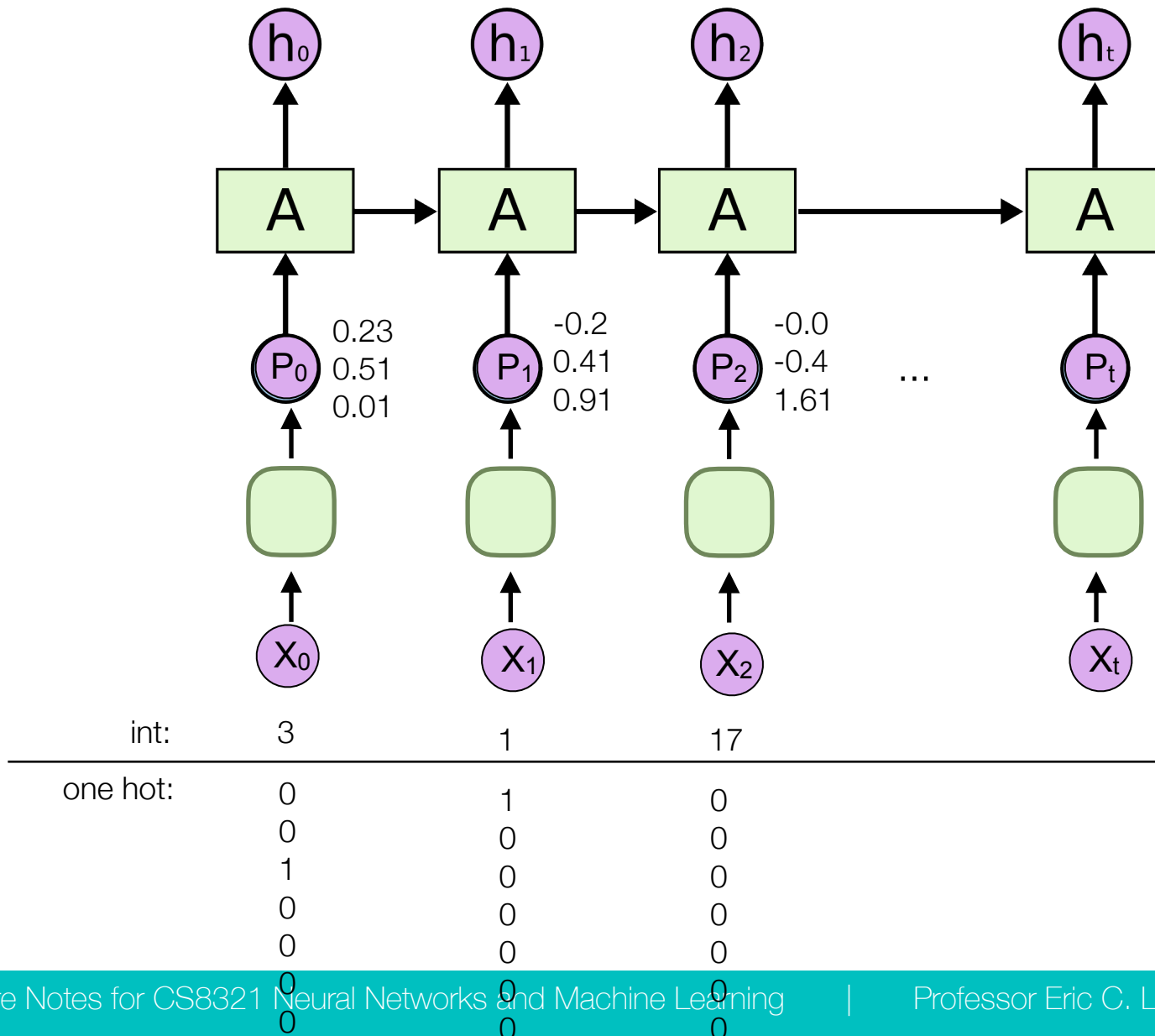
NeurIPS 2020 Workshop: ML Retrospectives, Surveys & Meta-analyses (ML-RSA), Virtual.



NLP Embeddings Review

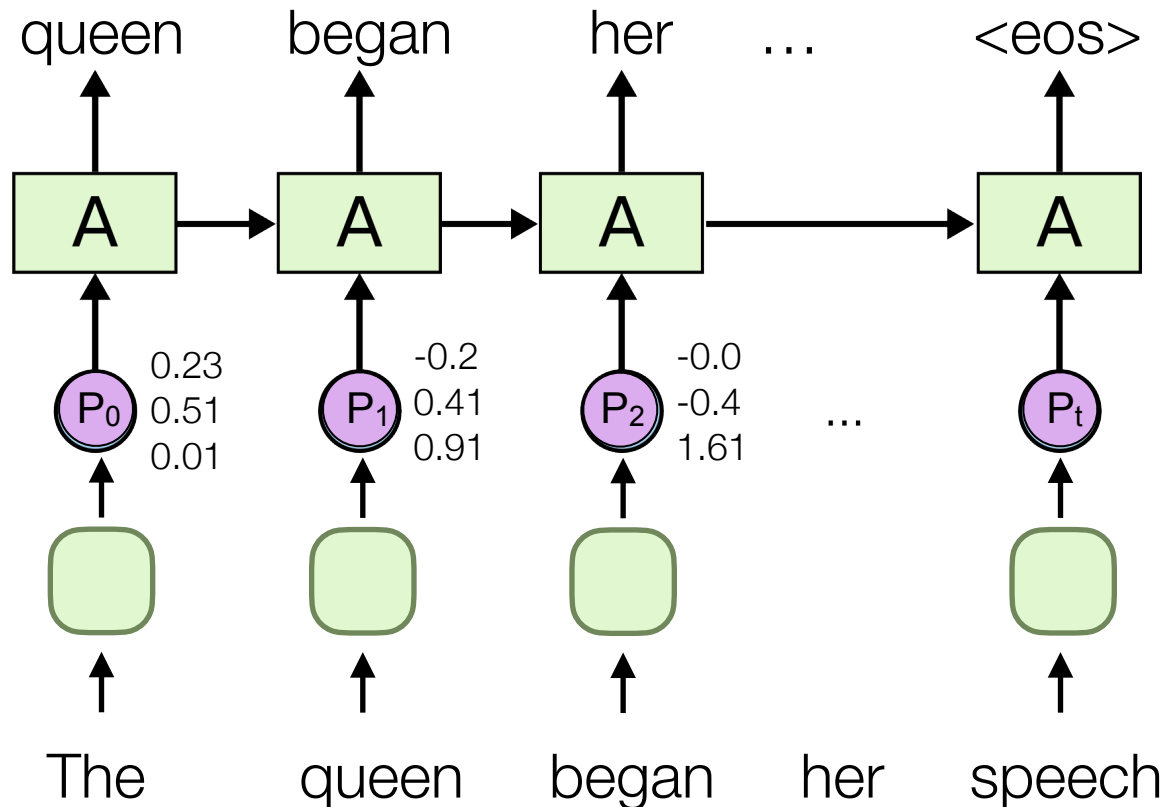


Word Embeddings (like Wide/Deep)



Word Embeddings: Training

- many training options exist
 - a popular option, next word prediction



Word Embeddings

GloVe

Global Vectors for Word Representation

Highlights

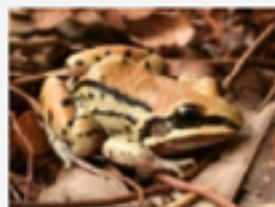
1. Nearest neighbors

The Euclidean distance (or cosine similarity) between two word vectors provides an effective method for measuring the linguistic or semantic similarity of the corresponding words. Sometimes, the nearest neighbors according to this metric reveal rare but relevant words that lie outside an average human's vocabulary. For example, here are the closest words to the target word *frog*:

0. *frog*
1. frogs
2. toad
3. *litoria*
4. *leptodactylidae*
5. *rana*
6. lizard
7. *eleutherodactylus*



3. *litoria*



4. *leptodactylidae*

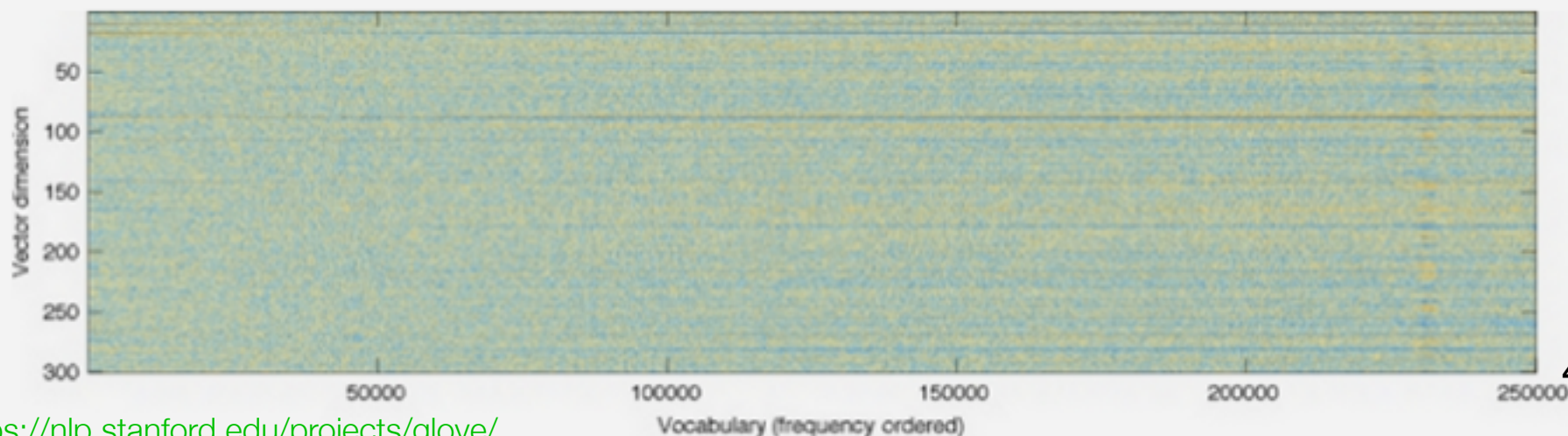


5. *rana*



7. *eleutherodactylus*

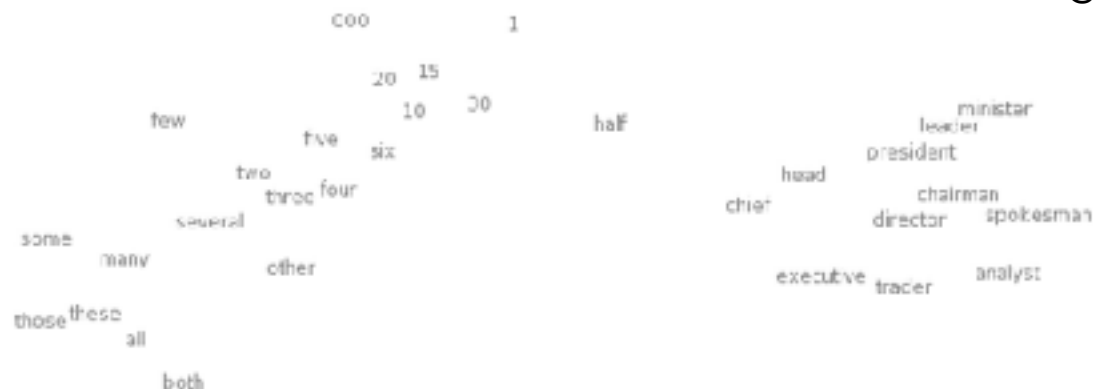
GloVe produces word vectors with a marked banded structure that is evident upon visualization:



Word Embeddings: proximity

GloVe

Global Vectors for Word Representation



t-SNE visualizations of word embeddings. Left: Number Region; Right: Jobs Region. From Turian *et al.* (2010), see complete image.

FRANCE	JESUS	XBOX	REDDISH	SCRATCHED	MEGABITS
AUSTRIA	GOD	AMIGA	GREENISH	NAILED	OCTETS
BELGIUM	SATI	PLAYSTATION	BLuish	SMASHED	MB/S
GERMANY	CHRIST	MSX	PINKISH	PUNCHED	BIT/S
ITALY	SATAN	IPOD	PURPLISH	POPPED	DAUD
GREECE	KALI	SEGA	BROWNISH	CRIMPED	CARATE
SWEDEN	INDRA	PSNUMBER	GREYISH	SCRAPED	KBIT/S
NORWAY	VISHNU	HD	GRAYISH	SCREWED	MEGAHERTZ
EUROPE	ANANDA	DREAMCAST	WHITISH	SECTIONED	MEGAPIXELS
HUNGARY	PARVATI	GEFORCE	SILVERY	SLASHED	GBIT/S
SWITZERLAND	GRACE	CAPCOM	YELLOWISH	RIPPED	AMPERES

What words have embeddings closest to a given word? From Collobert *et al.* (2011)

<http://colah.github.io/posts/2014-07-NLP-RNNs-Representations/>

The **chairman** called the **meeting** to order.

The **director** called the **conference** to order.

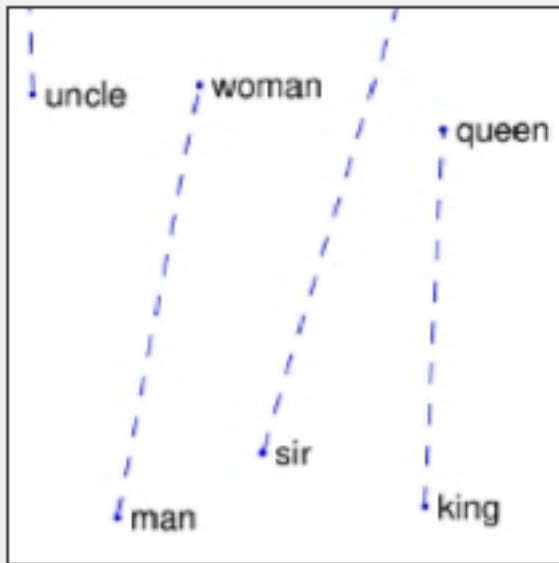
The **chief** called the **council** to order.



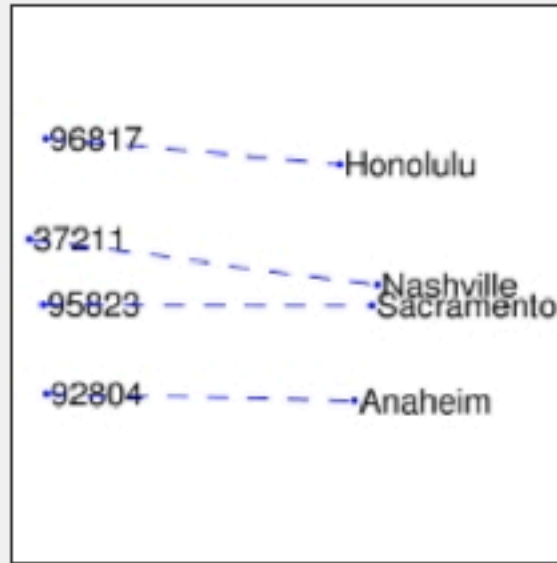
Word Embeddings: Analogy

GloVe

Global Vectors for Word Representation



man - woman



city - zip code

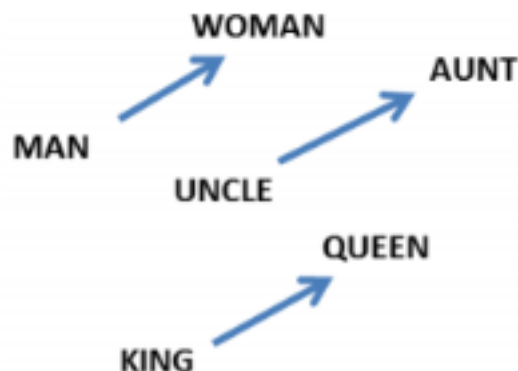


comparative - superlative

each vector difference **might** encode analogy



Word Embeddings: Analogy?



$$W(\text{"woman"}) - W(\text{"man"}) \simeq W(\text{"aunt"}) - W(\text{"uncle"})$$

$$W(\text{"woman"}) - W(\text{"man"}) \simeq W(\text{"queen"}) - W(\text{"king"})$$

$$\vec{\text{man}} - \vec{\text{woman}} \approx \vec{\text{computer programmer}} - \vec{\text{homemaker}}$$

From Mikolov *et al.*
(2013a)

Trained on
New York Times



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 |

Bolukbasi et al., NeurIPS 2016

<https://arxiv.org/pdf/1607.06520.pdf>

<https://nlp.stanford.edu/projects/glove/>



Tolga Bolukbasi¹, Kai-Wei Chang², James Zou², Venkatesh Saligrama^{1,2}, Adam Kalai²¹Boston University, 8 Saint Mary's Street, Boston, MA²Microsoft Research New England, 1 Memorial Drive, Cambridge, MA

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

Step 1: Identify gender subspace. Inputs: word sets W , defining sets $D_1, D_2, \dots, D_n \subset W$ as well as embedding $\{\vec{w} \in \mathbb{R}^d\}_{w \in W}$ and integer parameter $k \geq 1$. Let

$$\mu_i := \sum_{w \in D_i} \vec{w} / |D_i|$$

be the means of the defining sets. Let the bias subspace B be the first k rows of $\text{SVD}(\mathbf{C})$ where

$$\mathbf{C} := \sum_{i=1}^n \sum_{w \in D_i} (\vec{w} - \mu_i)^T (\vec{w} - \mu_i) / |D_i|.$$

Step 2a: Hard de-biasing (neutralize and equalize). Additional inputs: words to neutralize $N \subseteq W$, family of equality sets $\mathcal{E} = \{E_1, E_2, \dots, E_m\}$ where each $E_i \subseteq W$. For each word $w \in N$, let \vec{w} be re-embedded to

$$\vec{w} := (\vec{w} - \vec{w}_B) / \|\vec{w} - \vec{w}_B\|.$$

For each set $E \in \mathcal{E}$, let

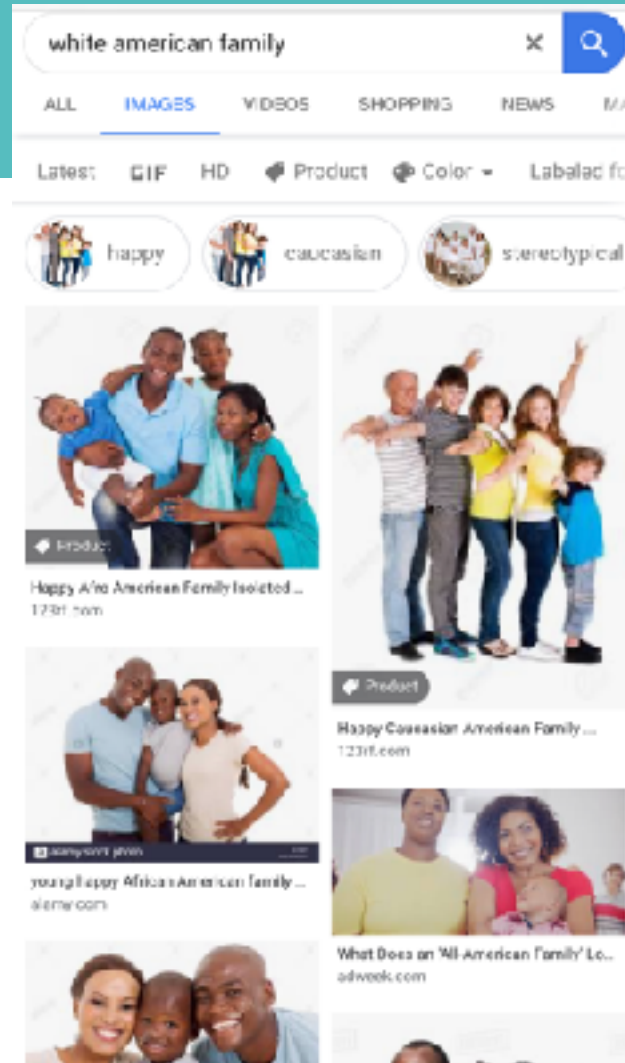
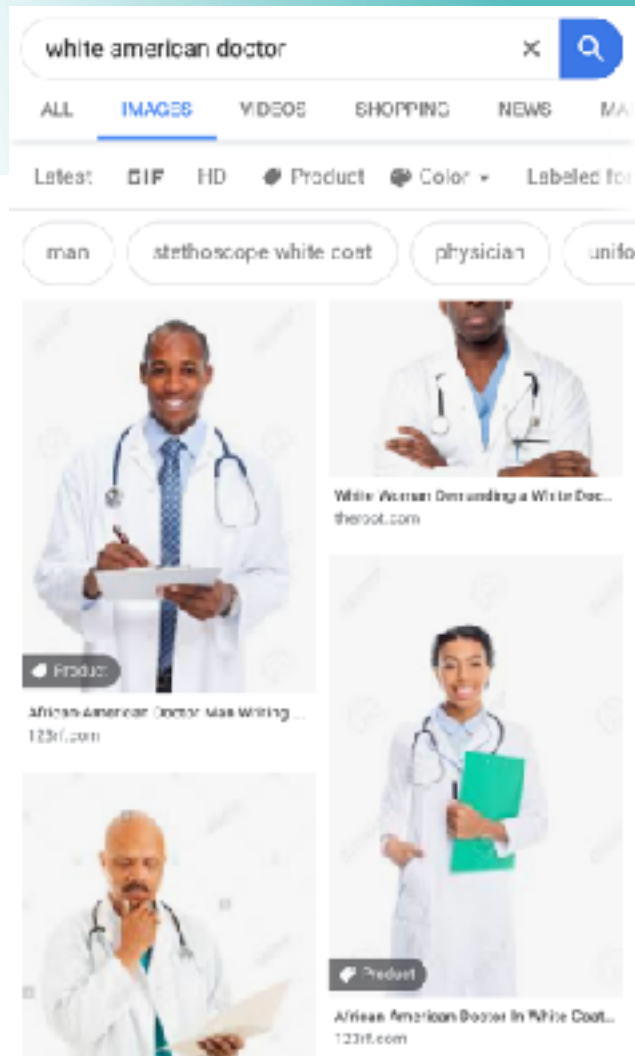
$$\mu := \sum_{w \in E} w / |E|$$

$$\nu := \mu - \mu_B$$

$$\text{For each } w \in E, \quad \vec{w} := \nu + \sqrt{1 - \|\nu\|^2} \frac{\vec{w}_B - \mu_B}{\|\vec{w}_B - \mu_B\|}$$



Practical Example in NLP



ConceptNet

en **cooking dinner**

An English term in ConceptNet 5.7

Source: Open Mind Common Sense contributors

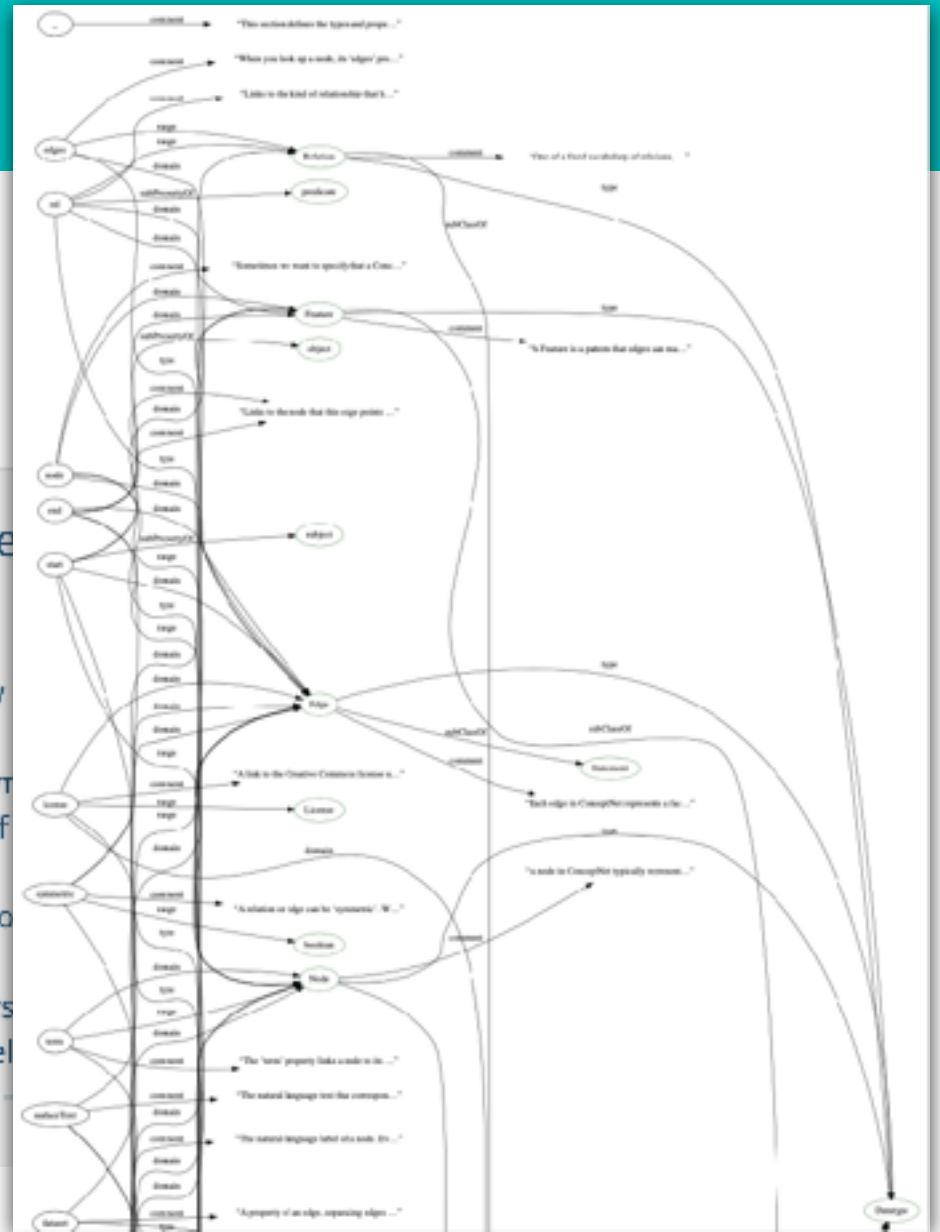
[View this term in the API](#)

cooking dinner is a
subevent of...

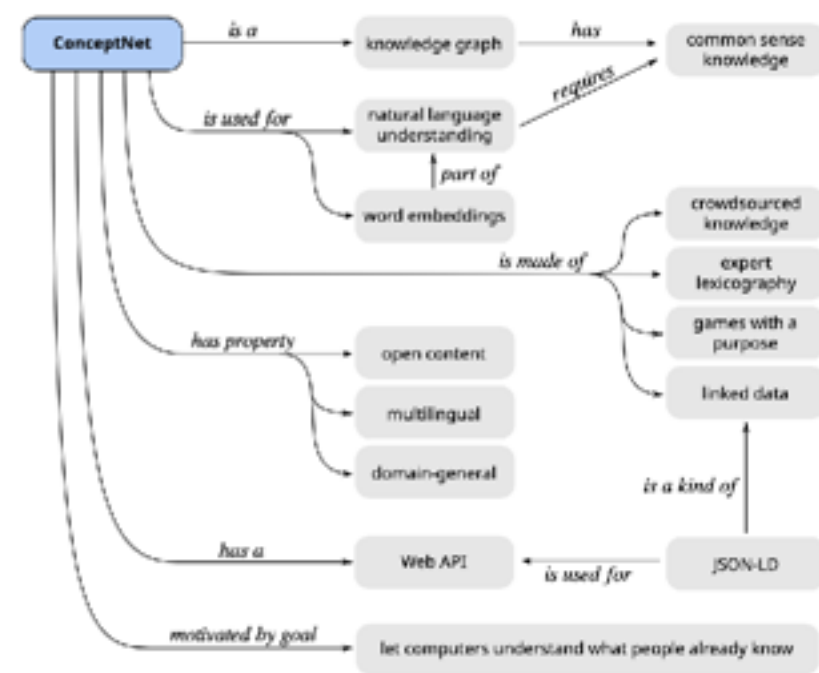
cooking dinner
for...

- en boiling water →
- en it burns →
- en preheat the oven →
- en taste the food →
- en boil salt water →
- en boil water →
- en brown the hamburger →
- en chop a vegetable →
- en defrost →
- en a fire →
- en the fire alarm might go off →

- en feeding a family
- en TO EAT →
- en entertaining com
- en feeding yourself
- en anyone →
- en avoiding fast food
- en being a cook →
- en caring for others
- en cheering yourself
- en creative people -
- en eating →



ConceptNet Numberbatch



- Create with a Knowledge Graph (from multiple sources with relations like *UsedFor*, *PartOf*, etc.)
- Based on this KG, perturb existing embeddings (like GloVe) to optimize:

$$\Psi(Q) = \sum_{i=1}^n \left[\alpha_i \|q_i - \hat{q}_i\|^2 + \sum_{(i,j) \in E} \beta_{ij} \|q_i - q_j\|^2 \right]$$

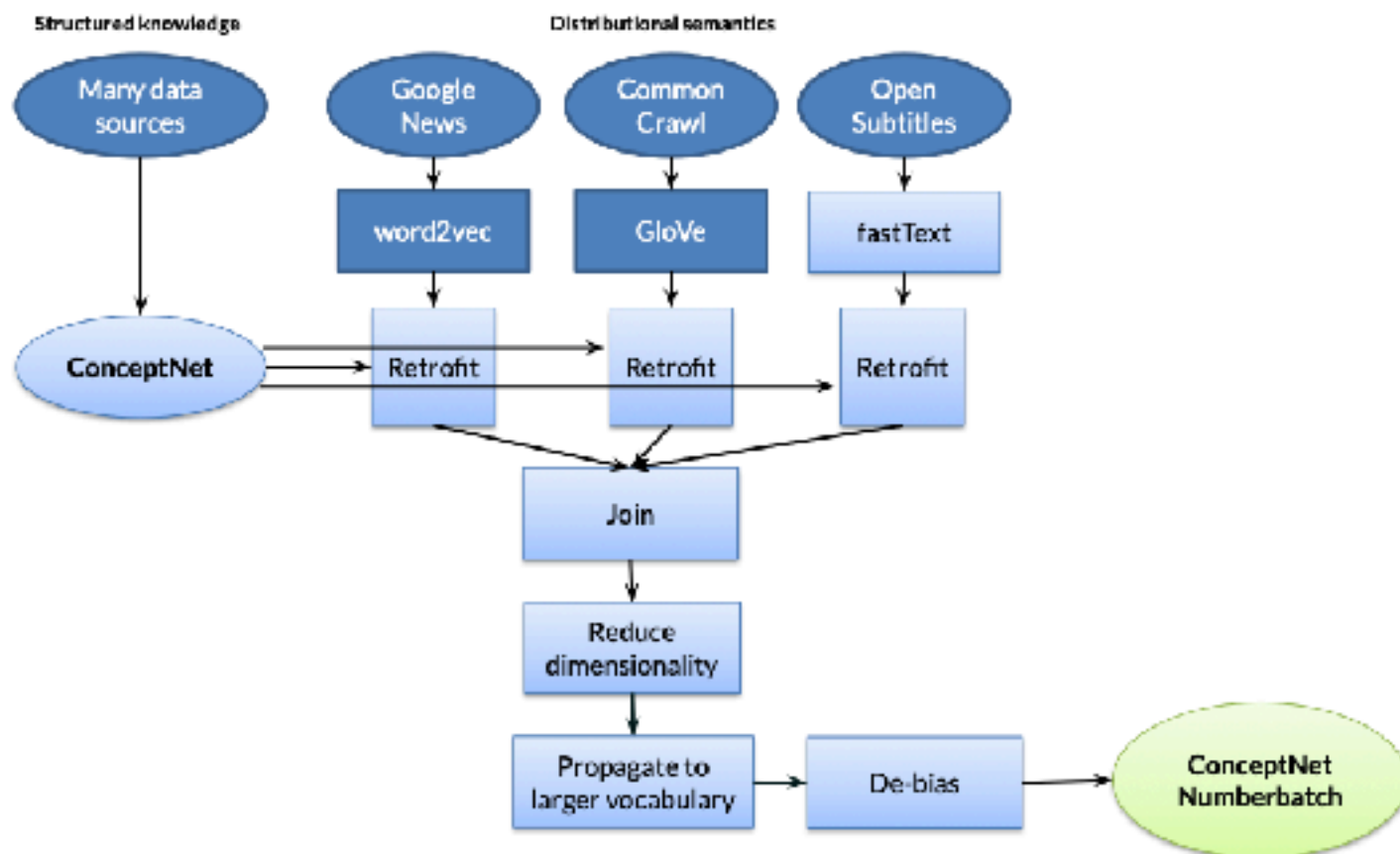
↑ ↑ ↖
 new embed old embed neighbors from KG

(keep similar to original) (make similar according to other knowledge)

- Easy to optimize the objective by averaging neighbors in the ConceptNet KG
- Multiple embeddings achieved by merging through “retrofitting” which projects onto a shared matrix space (with SVD)



Building ConceptNet Numberbatch



Aside: Transparency in Research

ConceptNet is all you need

Our full classifier used the linear combination of 5 types of input features shown above. This point is labeled **ABCDE** on the graph to the right. The other points are ablated versions of the classifier, trained on subsets of the five sources.

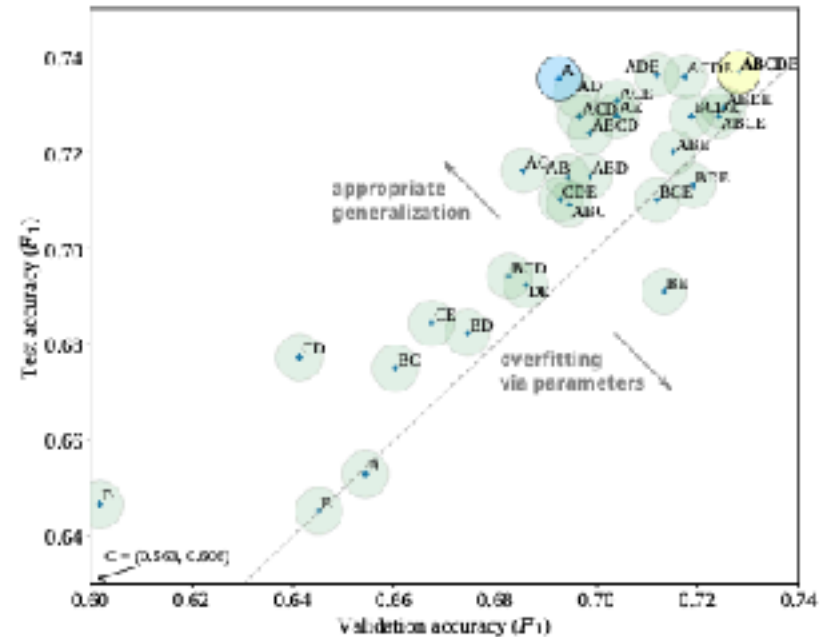
We found that the single feature of ConceptNet similarity (**A**) performed just as well on the test data as the full classifier, despite its lower validation accuracy.

This one-feature classifier could be more simply described as a heuristic over cosine similarities of ConceptNet embeddings:

$$\text{sim}(\text{term}_1, \text{attr}) - \text{sim}(\text{term}_2, \text{attr}) > 0.0961$$

It seems that the test data contained distinctions that can already be found by comparing ConceptNet embeddings, and that more complex features may have simply provided an opportunity to overfit to the validation set by parameter selection.

Results for all subsets of sources



This graph shows the validation and test accuracy of classifiers trained on subsets of the five sources of features. Ellipses indicate standard error of the mean, assuming that the data is sampled from a larger set.



ConceptNet Numberbatch

As a kid, I used to hold marble racing tournaments in my room, rolling marbles simultaneously down plastic towers of tracks and funnels. I went so far as to set up a bracket of 64 marbles to find the fastest marble. I kind of thought that running marble tournaments was peculiar to me and my childhood, but now I've found out that marble racing videos on YouTube are a big thing! Some of them even have **overlays as if they're major sporting events**.

In the end, there's nothing special about the fastest marble compared to most other marbles. It's just lucky. If one ran the tournament again, the marble champion might lose in the first round. But the one thing you could conclude about the fastest marble is that it was no *worse* than the other marbles. A bad marble (say, a misshapen one, or a plastic bead) would never luck out enough to win.

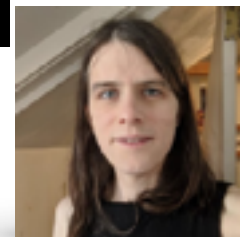
In our paper, we tested 30 alternate versions of the classifier, including the one that was roughly equivalent to this very simple system. We were impressed by the fact that it performed as well as our real entry. And this could be because of the inherent power of ConceptNet Numberbatch, or it could be because it's the lucky marble.

-Robyn Speer
<http://blog.conceptnet.io>





How to Make a Racist AI without Really Trying



Robyn Speer, 2017

<http://blog.conceptnet.io/posts/2017/how-to-make-a-racist-ai-without-really-trying/>

Debiasing: Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings

Bolukbasi et al., NeurIPs 2016

<https://arxiv.org/pdf/1607.06520.pdf>

ConceptNet 5.5: An Open Multilingual Graph of General Knowledge

Speer et al., AAAI 2017

<https://arxiv.org/pdf/1612.03975.pdf>



Rachael Tatman @rctatman · 18h

I first got interested in ethics in NLP/ML because I was asking "does this system work well for everyone". It's a good question, but there's a more important one:

Who is being harmed and who is benefiting from this system existing in the first place?



Lecture Notes for **Neural Networks and Machine Learning**

Ethically Aware NLP



Next Time:
CNN Visualization
Reading: Chollet Article

