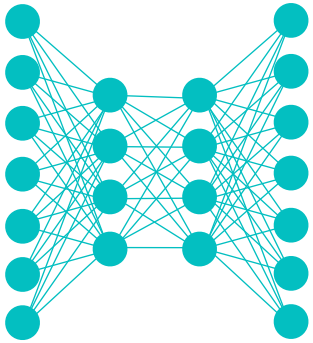
Lecture Notes for

Neural Networks and Machine Learning

A Practical Example of Ethically “Aware” NLP Practices



Logistics and Agenda

- Logistics
  - Lecture discussion assignments
  - Office hours posted
- Last Time:
  - Ethical Guidelines
  - Case Studies Intro
- Agenda
  - Final Case Study
  - NLP Review
  - Extended Example

Ethical Considerations in Military App.

- Ethical guidelines in combat
  - **One Interpretation**: Combat is not ethical because harm of individuals is incentivized
  - **Another**: Certain ethical guidelines can still be considered, accepting that the aim of combat is, in many instances, to create a weapon (and increase safety to various groups)
- These are both common (maybe valid) interpretations and it is up to you to decide if combat should be included in ethical guidelines
  - **My take**: combat should not be considered ethical, but is unavoidable in the presence of nefarious actors and therefore guidelines need to be in place
  - Many individuals will disagree with me on this and have excellent points, that I do not disagree with

AI Warfare

"Radical Shift In Warfighting": Real-Time Soldiers Fight Real Soldiers In Ukraine War; Emerge Victorious

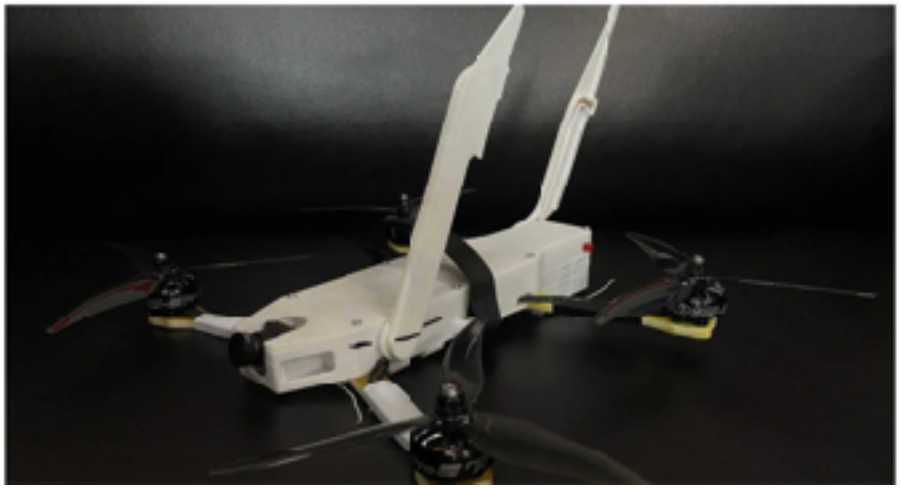
In another innovation in the drone war between Russia and Ukraine, Kyiv is using large UAVs with 'repeaters' to relay signals and control other latent kamikaze Unmanned Aerial Vehicles (UAVs) to strike Russian soldiers.

The employment of small suicide drones in anti-infantry roles is preventing Russian soldiers from coming to the frontlines or getting picked out by the UAVs. Russia's credible electronic warfare (EW), too, is catching up, and a top retired Ukrainian general admitted how Moscow keeps finding new ways to jam the UAVs.

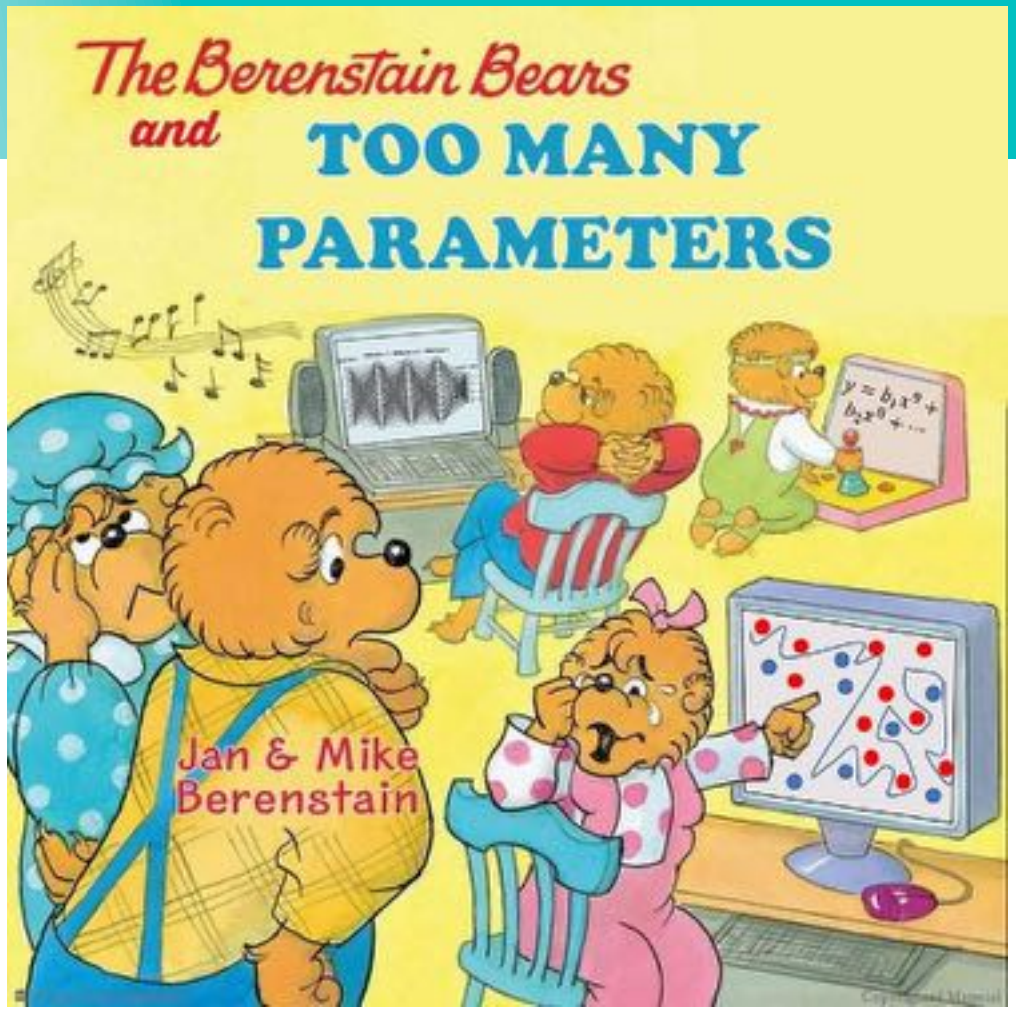
The developments follow previous analyses in the EurAsian Times, which reported on the escalation of the drone war and rapidly developing tactics ever since both countries began strapping commercial quadcopters with explosives. The technological advantage keeps shifting as Ukraine adopts a loitering munition-centric strategy for all its land warfare roles, including as substitutes for artillery.

'Mother' Drone Controlling Other Kamikaze UAVs?

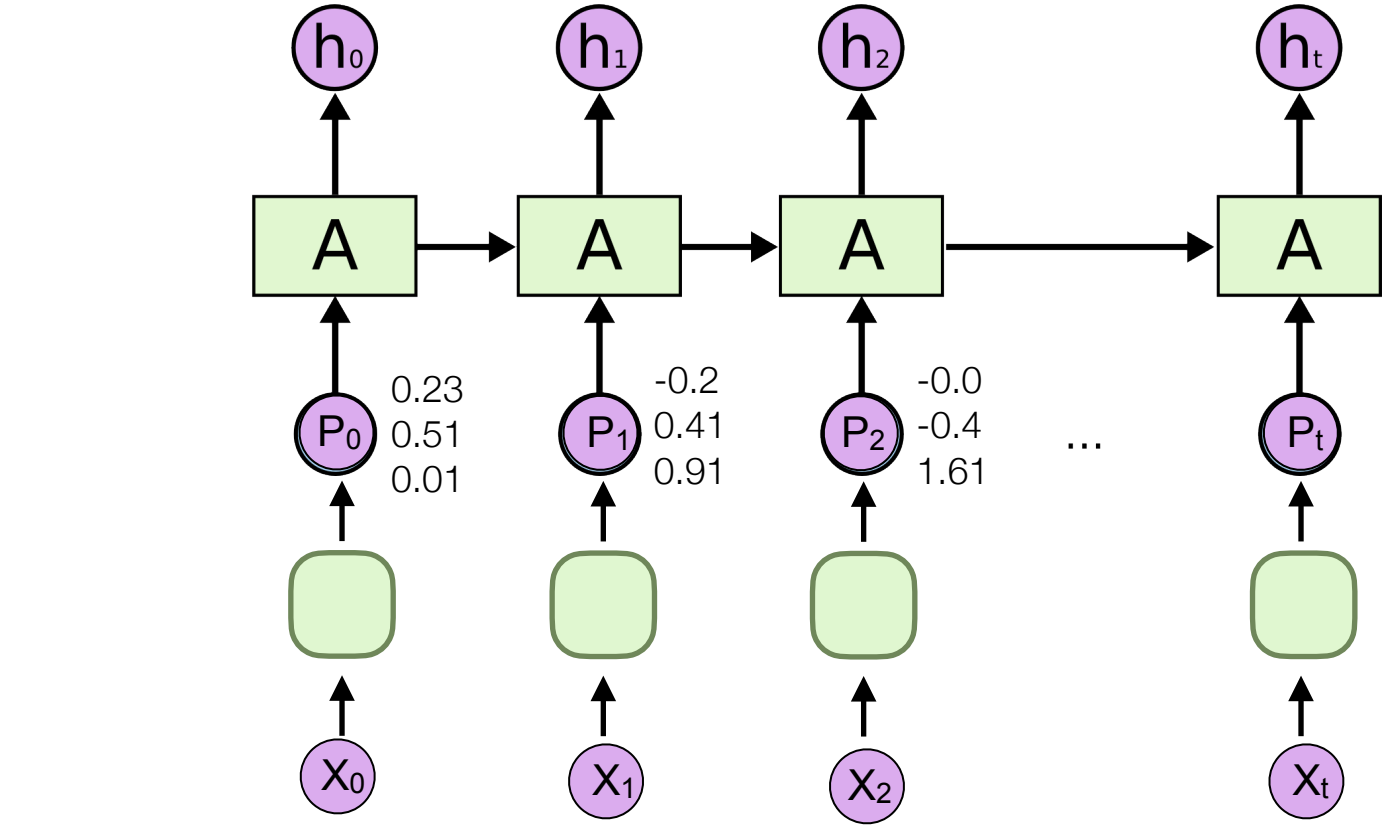
Izvestia journalist Dmitry Zimenkin posted a video of an interview with a Russian soldier on his Telegram channel, where the latter reported the new drone tactics. The soldier, identified with the call sign "Screw," said Ukraine uses "a big flying queen and a flock of her little drones," enemy in the Severskiy direction.



NLP Embeddings Review

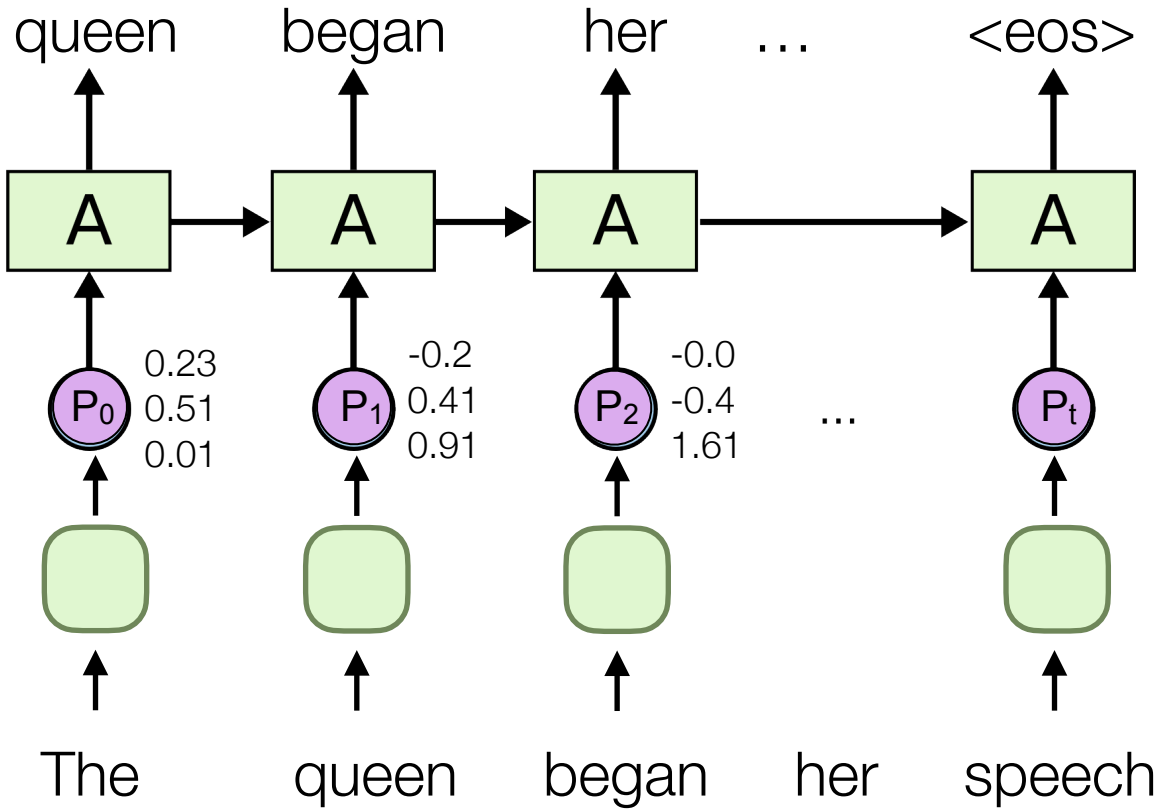


Word Embeddings Review



Word Embeddings: Training Review

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



GloVe Review

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

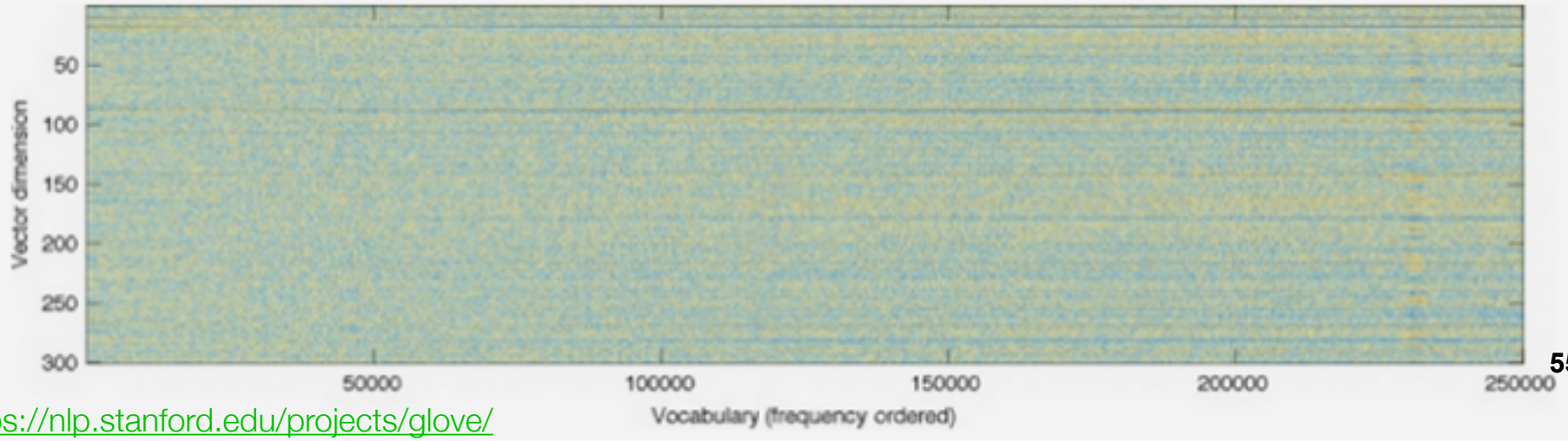
3. litoria

4. leptodactylidae

5. rana

7. eleutherodactylus

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

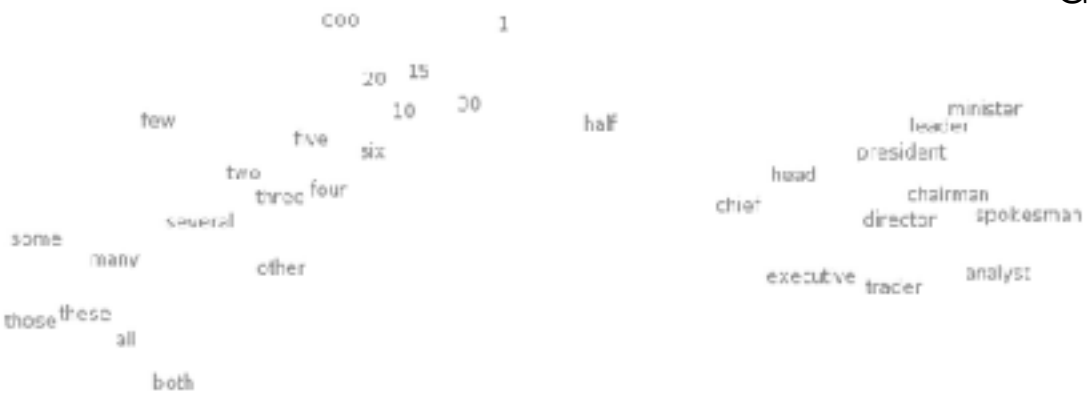


https://nlp.stanford.edu/projects/glove/

Word Embeddings: proximity

GloVe Review

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	BAUD
GREECE	KALI	SEGA	BROWNISH	CRIMPED	CARATS
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)

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

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

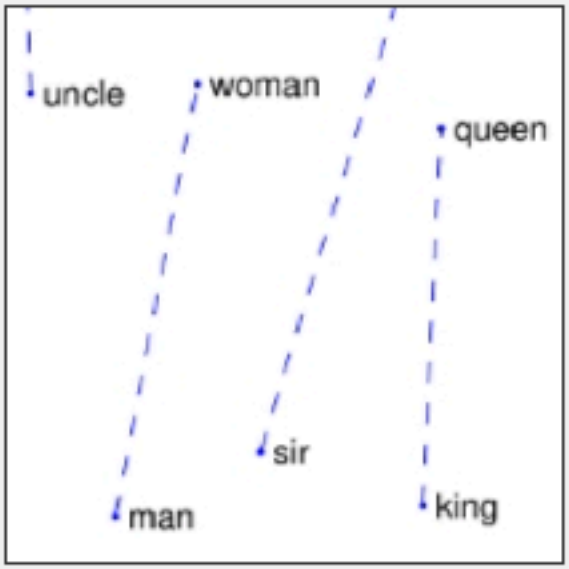
The **chief** called the **council** to order.

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

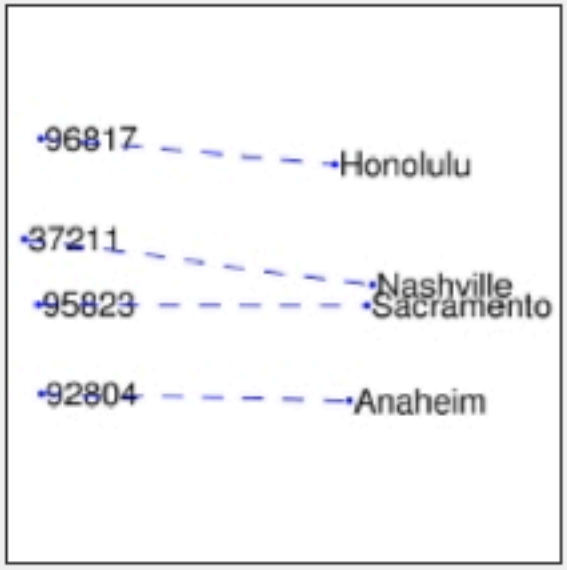
Word Embeddings: Analogy

GloVe Review

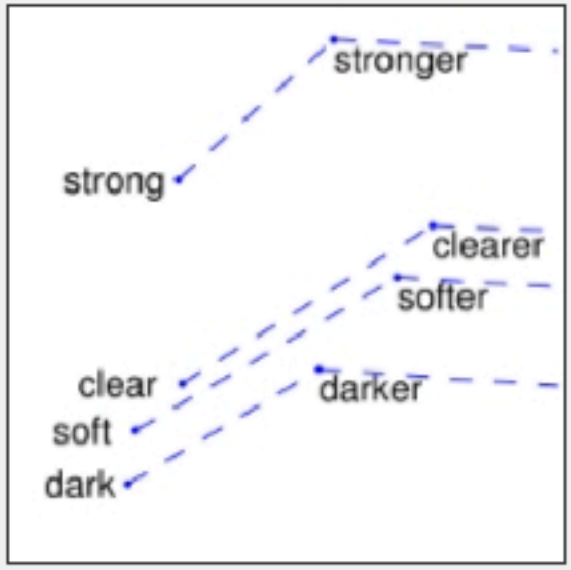
Global Vectors for Word Representation



man - woman



city - zip code



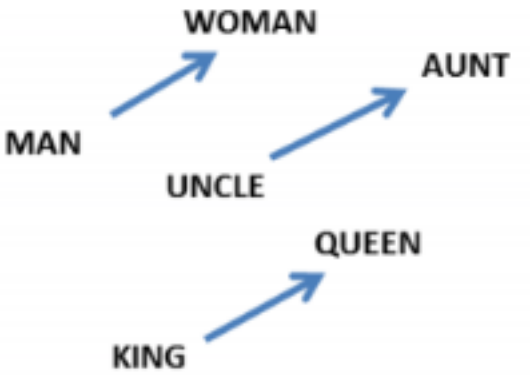
comparative - superlative

each vector difference **might** encode analogy

https://nlp.stanford.edu/projects/glove/ http://colah.github.io/posts/2014-07-NLP-RNNs-Representations/

Word Embeddings: Analogy?

GloVe Review



From Mikolov *et al.* (2013a)

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

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

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

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

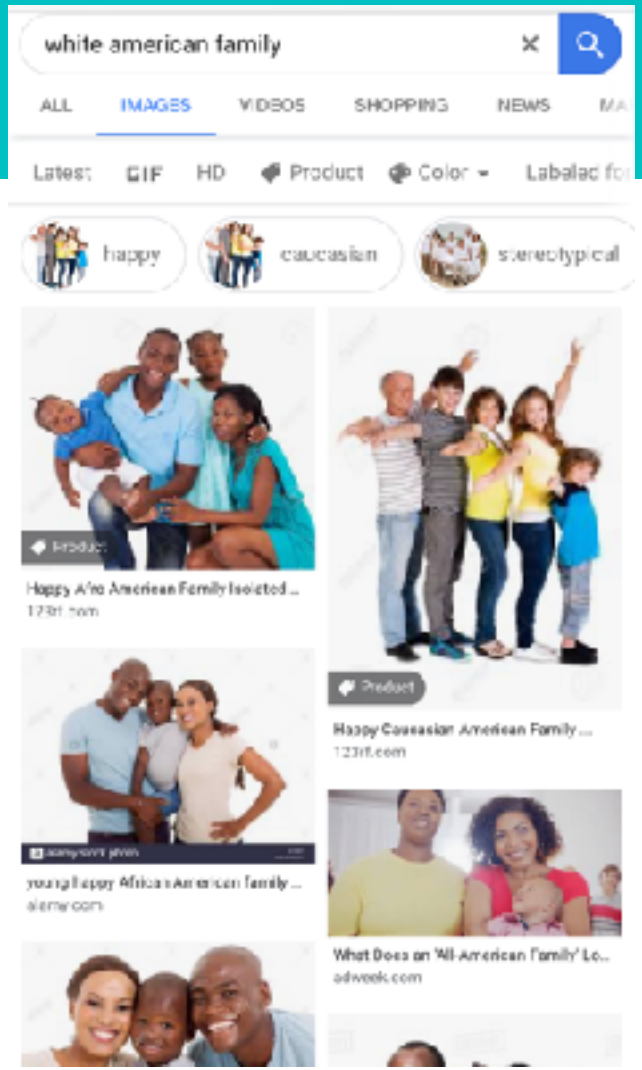
1. maestro	2. skipper	3. protege
4. philosopher	5. captain	6. architect
7. financier	8. warrior	9. broadcaster
10. magician	11. fighter pilot	12. boss

Trained on New York Times

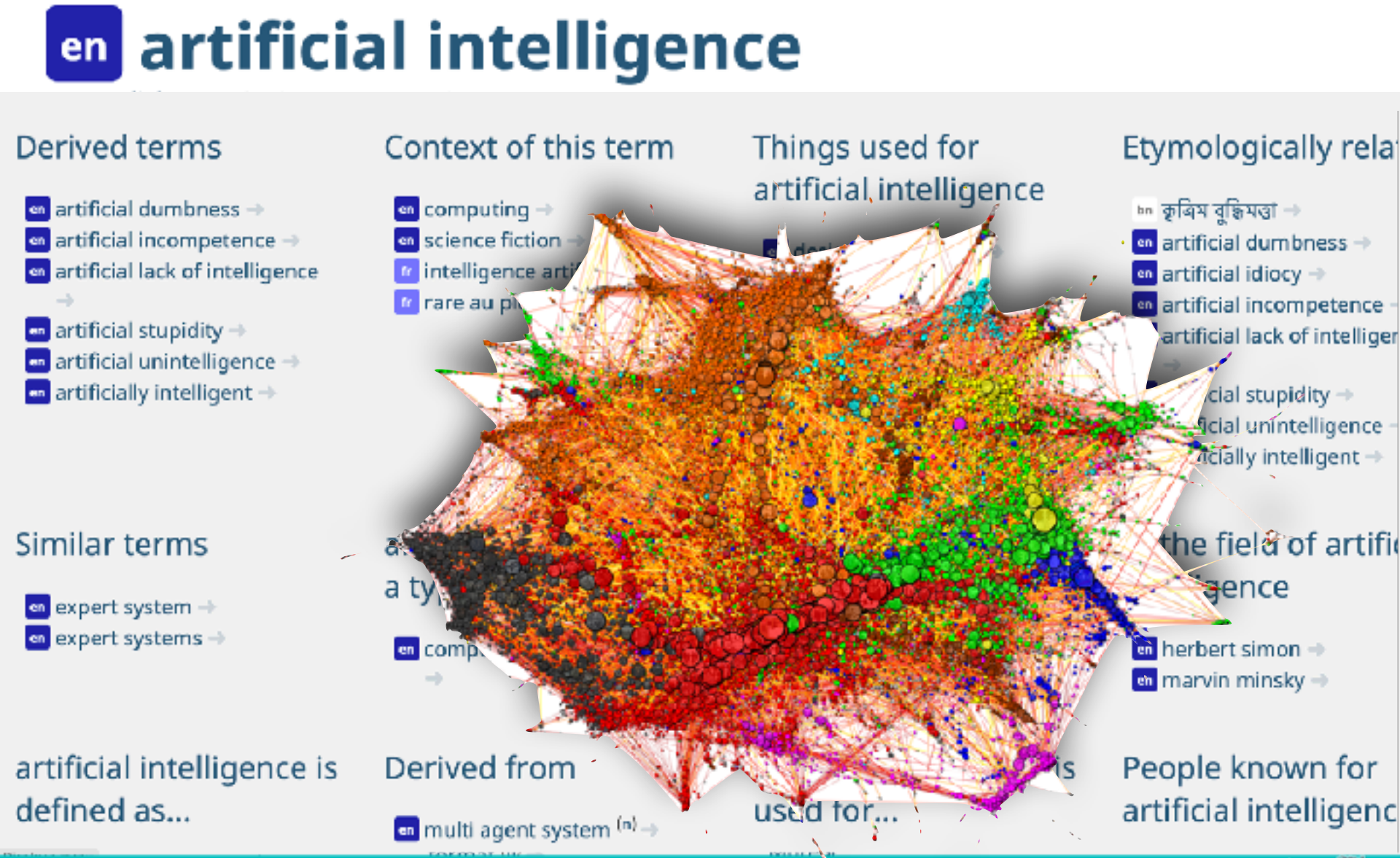
Bolukbasi et al., NeurIPs 2016
https://arxiv.org/pdf/1607.06520.pdf

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

Practical Example in NLP



ConceptNet


en artificial intelligence
Derived terms
artificial dumbness
artificial incompetence
artificial lack of intelligence
artificial stupidity
artificial unintelligence
artificially intelligent
Context of this term
computing
science fiction
Things used for
artificial intelligence
Etymologically rela
কৃত্রিম বুদ্ধিমত্তা
artificial dumbness
artificial idiocy
artificial incompetence
Similar terms
expert system
expert systems
herbert simon
marvin minsky
artificial intelligence is
defined as...
Derived from
multi agent system
People known for
artificial intelligenc

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:

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

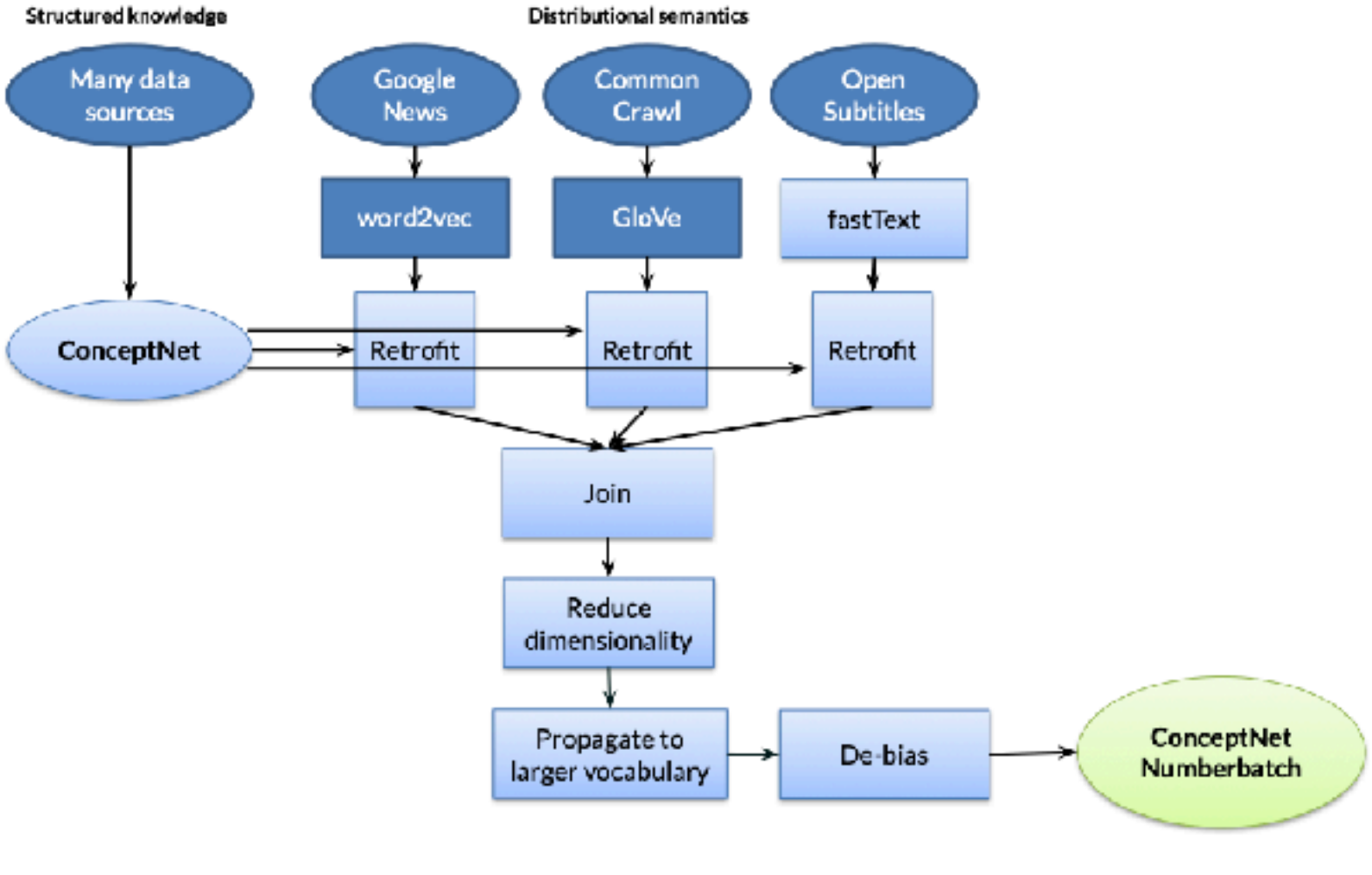
new embed — old embed — neighbors from KG

(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)

ConceptNet 5.5: An Open Multilingual Graph of General Knowledge, Speer et al., 2017

Building ConceptNet Numberbatch



LUMINOSO

Aside: Transparency in Research

ConceptNet is all you need

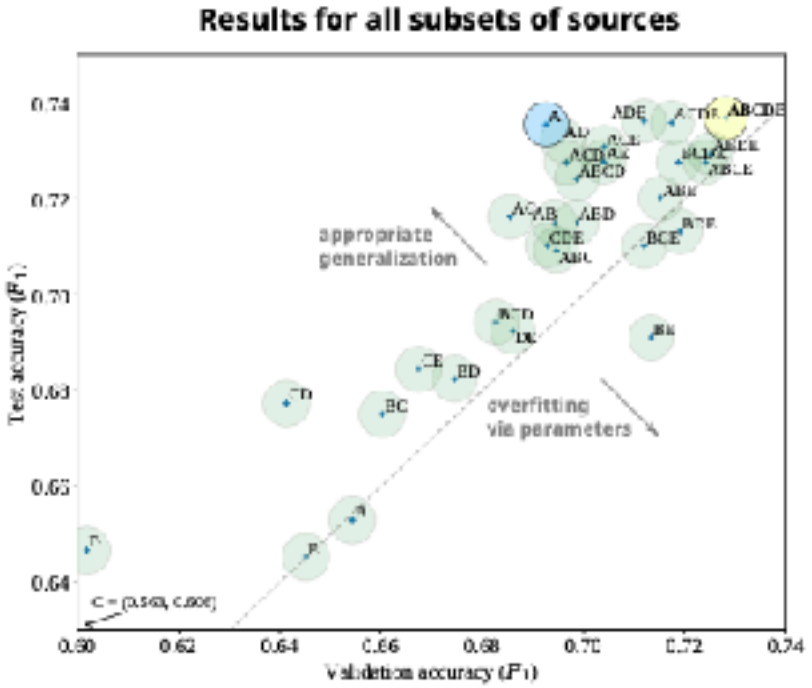
Our full classifier used the linear combination of 5 types of input features shown above. This point is labeled **ABCDE** on the graph to the right. The other points are ablated versions of the classifier, trained on subsets of the five sources.

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:

$$\text{sim}(term_1, art) - \text{sim}(term_2, art) > 0.0961$$

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.



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.

https://blog.conceptnet.io/2018/06/naacl2018-poster.pdf

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 becuase I was asking "does this system work well for everyone". It's a good question, but there's a more important 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 Practices

Next Time:
Transfer Learning

Reading: Chollet Article