

Modern word embedding methods

Väinö Yrjänäinen

8.12.2022

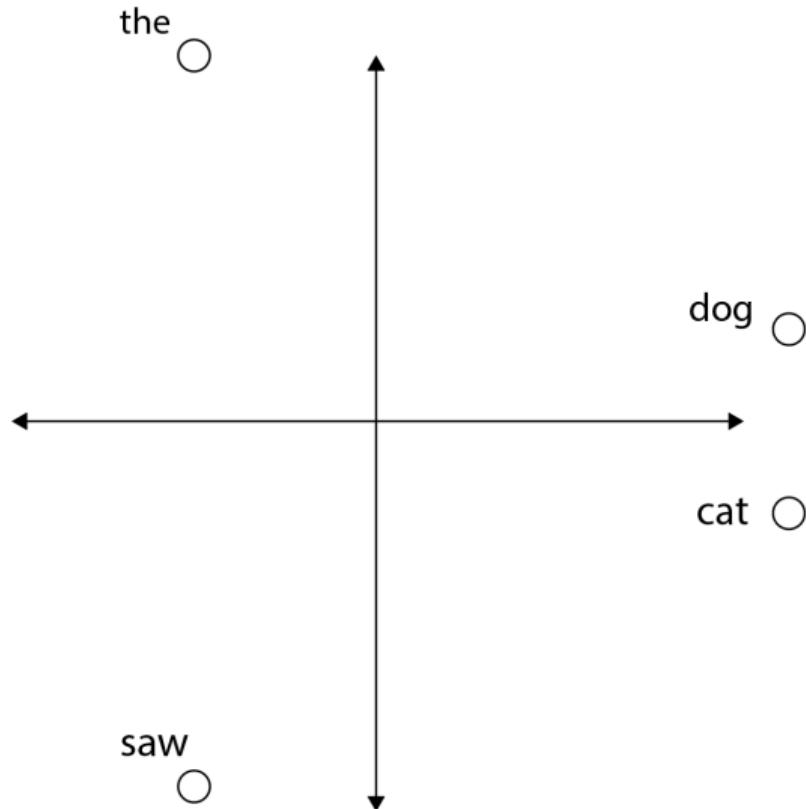
Analysis of large corpora?

mensal pathway into domestication. The questions of when and where dogs were first domesticated have taxed geneticists and archaeologists for decades. Genetic studies suggest a domestication process commencing over 25,000 years ago, in one or several wolf populations in either Europe, the high Arctic, or eastern Asia. In 2021, a literature review of the current evidence infers that the dog was domesticated in Siberia 23,000 years ago by ancient North Siberians, then later dispersed eastward into the Americas and westward across Eurasia. Breeds Dogs are the most variable mammal on earth with around 450 globally recognized dog breeds. In the Victorian era, directed human selection developed the modern dog breeds, which resulted in a vast range of phenotypes. Most breeds were derived from small numbers of founders within the last 200 years, and since then dogs have undergone rapid phenotypic change and were formed into today's modern breeds due to artificial selection imposed by humans. The skull, body, and limb proportions vary significantly between breeds, with dogs displaying more phenotypic diversity than can be found within the entire order of carnivores. These breeds possess distinct traits related to morphology, which include body size, skull shape, tail phenotype, fur type and colour. Their behavioural traits include guarding, herding, and hunting, retrieving, and scent detection. Their personality traits include hypersocial behavior, boldness, and aggression, which demonstrates the functional and behavioral diversity of dogs. As a result, present day dogs are the most abundant carnivore species and are dispersed around the world. The most striking example of this dispersal is that of the numerous modern breeds of European lineage during the Victorian era. Biology Anatomy Skeleton All healthy dogs, regardless of their size and type, have an identical skeletal structure with the exception of the number of bones in the tail, although there is significant skeletal variation between dogs of different types. The dog's skeleton is well adapted for running; the vertebrae on the neck and back have extensions for powerful back muscles to connect to, the long ribs provide plenty of room for the heart and lungs, and the shoulders are unattached to the skeleton allowing great flexibility. Compared to the dog's wolf-like ancestors, selective breeding since domestication has seen the dog's skeleton greatly enhanced in size for larger types as mastiffs and miniaturised for smaller types such as terriers; dwarfism has been selectively utilised for some types where short legs are advantageous such as dachshunds and corgis. Most dogs naturally have 26 vertebrae in their tails, but some with naturally short tails have as few as three. The dog's skull has identical components regardless of breed type, but there is significant divergence in terms of skull shape between types. The three basic skull shapes are the elongated dolichocephalic type as seen in sighthounds, the intermediate mesocephalic or mesaticephalic type, and the very short and broad brachycephalic type exemplified by mastiff type skulls. Senses A dog's senses include vision, hearing, smell, taste, touch, and sensitivity to Earth's magnetic field. Another study has suggested that dogs can see Earth's magnetic field. Coat The coats of domestic dogs are of two varieties: "double" being familiar with dogs (as well as wolves) originating from colder climates, made up of a coarse guard hair and a soft down hair, or "single", with the topcoat only. Breeds may have an occasional "blaze", stripe, or "star" of white fur on their chest or underside. Premature graying can occur in dogs from as early as one year of age; this is associated with impulsive behaviors, anxiety, fear of noise, and fear of unfamiliar people or animals. Tail There are many different shapes for dog tails: straight, straight up, sickle, curled, or corkscrew. As with many canids, one of the primary functions of a dog's tail is to communicate their emotional state, which can be crucial in getting along with others. In some hunting dogs the tail is traditionally docked to avoid injuries. Health Some breeds of dogs are prone to specific genetic ailments such as elbow and hip dysplasia, blindness, deafness, pulmonic stenosis, cleft palate, and trick knees. Two severe medical conditions significantly affecting dogs are pyometra, affecting unspayed females of all breeds and ages, and Gastric dilatation volvulus (bloat), which affects larger breeds or deep-chested dogs. Both of these are acute conditions and can kill rapidly. Dogs are also susceptible to parasites such as fleas, ticks, mites, hookworms, tapeworms, roundworms, and heartworms, which is a roundworm species that lives in the hearts of dogs. Several human foods and household ingestibles are toxic to dogs, including chocolate solids, causing theobromine poisoning, onions and garlic, causing thiosulphate, sulfoxide or disulfide poisoning, grapes and raisins, macadamia nuts, and xylitol. The nicotine in tobacco can also be dangerous to dogs. Signs of ingestion can include copious vomiting (e.g., from eating cigar butts) or diarrhea. Some other symptoms are abdominal pain, loss of coordination, collapse, or death. Dogs are also vulnerable to some of the same health conditions as humans, including diabetes, dental and heart disease, epilepsy, cancer, hypothyroidism, and arthritis. Lifespan The typical lifespan of dogs varies widely among breeds, but for most, the median longevity (the age at which half the dogs in a population have died and half are still alive) ranges from 10 to 13 years. The median longevity of mixed-breed dogs, taken as an average of all sizes, is one or more years longer than that of purebred dogs when all breeds are averaged. For dogs in England, increased body weight has been found to be negatively correlated with longevity (i.e., the heavier the dog, the shorter its lifespan), and mixed-breed dogs live on average 1.2 years longer than purebred dogs. Reproduction In domestic dogs, sexual maturity happens around six months to one year for both males and females, although this can be delayed until up to two years of age for some large breeds, and is the time at which female dogs will have their first estrous cycle. They will experience subsequent estrous cycles semiannually, during which the body prepares for pregnancy. At the peak of the cycle, females will become estrous, mentally and physically receptive to copulation. Because the ova survive and can be fertilized for a week after ovulation, more than one male can sire the same litter. Fertilization typically occurs two to five days after ovulation; 14-16 days after ovulation, the embryo attaches to the uterus and after seven to

Simple methods

- Counting words
 - Conceptually straightforward and computationally fast
 - Not very flexible
 - All words are treated as independent categories
- Manual analysis by humans
 - Not applicable to big data

Word embeddings



Word embeddings

$$\text{dog} = \begin{bmatrix} 0.23 \\ -1.47 \\ 1.01 \\ 1.33 \\ \dots \\ -0.56 \end{bmatrix}, \text{ cat} = \begin{bmatrix} 0.42 \\ -1.07 \\ 0.31 \\ 1.54 \\ \dots \\ 0.14 \end{bmatrix} \in \mathbb{R}^{100}$$

Word embeddings

- Distance
 - Cosine distance
 - How similar are words x, y
 - What is the most similar word to word x
- Analogies
 - $\text{king} - \text{man} + \text{woman} \approx \text{queen}$
- And more

Analogies

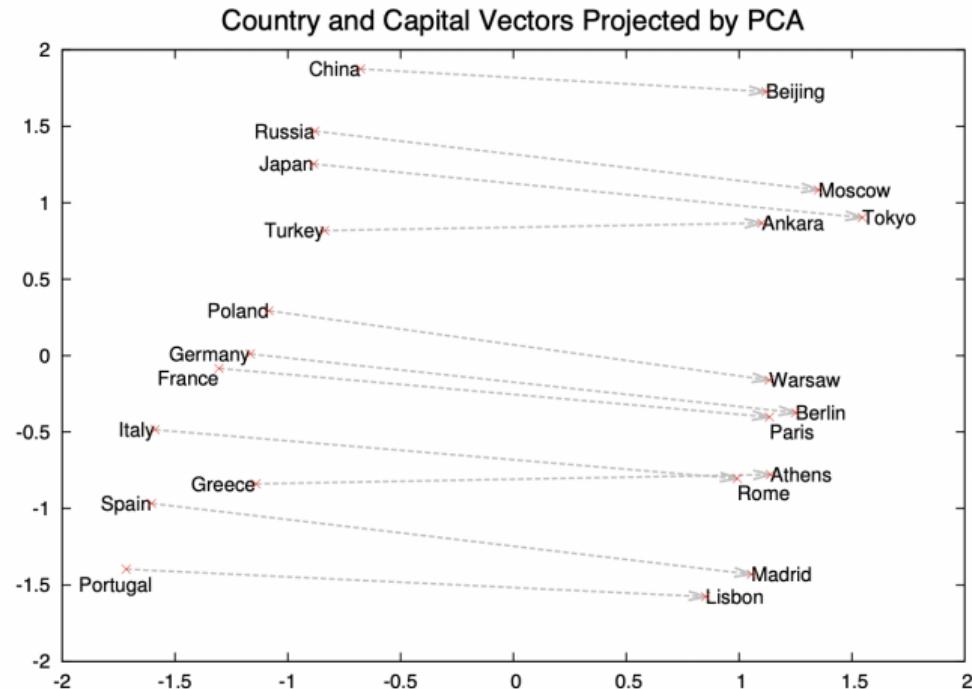


Figure 2: Two-dimensional PCA projection of the 1000-dimensional Skip-gram vectors of countries and their capital cities. The figure illustrates ability of the model to automatically organize concepts and learn implicitly the relationships between them, as during the training we did not provide any supervised information about what a capital city means.

Training word embeddings

Context window based methods

[...] as the oldest domesticated species, dogs' minds inevitably have been shaped by millennia of contact with humans [...]

Context window based methods

[...] as the oldest domesticated species, dogs' minds inevitably have been shaped by millennia of contact with humans [...]

Context window based methods

dog

{domesticated, species,
minds, inevitably}

Word2Vec – concept

- Continuous bag-of-words CBOW
Does the word x appear, given that its context window is y_1, y_2, \dots, y_{ws} ?
- Skip-gram
Do words x and y appear in the same word window?

Continuous bag-of-words

[...] as the oldest domesticated species, [???] minds inevitably have been shaped by millennia of contact with humans [...]

[...] as the oldest domesticated species, [???] minds inevitably have been shaped by millennia of contact with humans [...]

[...] as the oldest domesticated species, [???] minds inevitably have been shaped by millennia of contact with humans [...]

[...] as the oldest domesticated species, [???] minds inevitably have been shaped by millennia of contact with humans [...]

[...] as the oldest domesticated species, [???] minds inevitably have been shaped by millennia of contact with humans [...]

Word2Vec - mathematical definition

- Each word is assigned a word vector w and a context vector c
 - $w, c \in \mathbb{R}^D$, where D is the *dimensionality* of the embedding
- Conditional probabilities are functions of these vectors
 - $P(x | y_1, \dots, y_w) = f(w_x, c_{y_1}, \dots, c_{y_w})$

Skip-gram

Do words x and y appear in the same word window?

$$p(x \text{ and } y \text{ co-occur}) = \sigma(w_x^T c_y)$$

where $\sigma(\cdot)$ is the logistic function.

Skip-gram numerical example

$$w_{\text{dog}} = \begin{bmatrix} 0.23 \\ -1.47 \\ 1.01 \\ 1.33 \\ \dots \\ -0.56 \end{bmatrix}, c_{\text{cat}} = \begin{bmatrix} 0.42 \\ -1.07 \\ 0.31 \\ 1.54 \\ \dots \\ 0.14 \end{bmatrix}$$

Skip-gram numerical example

Do the words *dog* and *cat* appear in the same word window?

$$p(\text{dog and cat co-occur}) = \sigma \left(\begin{bmatrix} 0.23 & -1.47 & 1.01 & 1.33 & \dots & -0.56 \end{bmatrix} \begin{bmatrix} 0.42 \\ -1.07 \\ 0.31 \\ 1.54 \\ \dots \\ 0.14 \end{bmatrix} \right)$$
$$= \sigma(6.2325) = 0.998$$

Skip-gram numerical example

$$w_{\text{dog}} = \begin{bmatrix} 0.23 \\ -1.47 \\ 1.01 \\ 1.33 \\ \dots \\ -0.56 \end{bmatrix}, c_{\text{university}} = \begin{bmatrix} -2.03 \\ 0.07 \\ 1.01 \\ -0.34 \\ \dots \\ 0.89 \end{bmatrix}$$

Skip-gram numerical example

Do the words *dog* and *university* appear in the same word window?

$$p(\text{dog and university co-occur}) = \sigma \left(\begin{bmatrix} 0.23 & -1.47 & 1.01 & 1.33 & \dots & -0.56 \end{bmatrix} \begin{bmatrix} -2.03 \\ 0.07 \\ 1.01 \\ -0.34 \\ \dots \\ 0.89 \end{bmatrix} \right) = \sigma(-0.13) = 0.4675$$

Continuous bag-of-words

Do words x and y appear in the same word window?

$$p(x \text{ occurs given } C = y_1, \dots, y_{ws}) = \sigma(w_x^T \sum_{y \in C} c_y)$$

where C is the context window, and $\sigma(\cdot)$ is the logistic function.

Negative sampling – idea

- In the data, we can easily find instances where words do co-occur
 - “the dog saw the cat”
 - Eg. (dog,saw), (dog,cat) and (the, cat) co-occur
- We do not have direct examples of words not co-occurring
- We need negative examples, otherwise we will just have all vectors be the same
- Solution: pick at random

Negative sampling

- For each word x in the data
 - Sample ns negative samples y'_1, y'_{ns} from the empirical distribution of word types
 - Add a term $1 - p(x \text{ and } y' \text{ co-occur})$ in the likelihood

Negative sampling – putting it all together

Say we only have the data point *dog co-occurs with cat*. We draw a negative sample *university* from the empirical distribution of words in the data. The likelihood is

$$\begin{aligned} p(\text{data}) &= \underbrace{p(\text{dog and cat co-occur})}_{\text{positive sample}} \underbrace{(1 - p(\text{dog and university co-occur}))}_{\text{negative sample}} \\ &= 0.998 \cdot (1 - 0.4675) = 0.531435 \end{aligned}$$

Negative sampling – putting it all together

For a real dataset, the log likelihood is

$$\log p(\mathcal{D} \mid w, c) = \sum_{i=1}^N \left(\underbrace{\sum_{y \in C_i} \log \sigma(w_{x_i}^T c_y)}_{\text{positive samples}} + \underbrace{\sum_{y \in C_{ns}} \log(1 - \sigma(w_{x_i}^T c_y))}_{\text{negative samples}} \right)$$

where $\mathcal{D} = (x_1, \dots, x_N)$ is the dataset indexed by $i \in \{1, \dots, N\}$, C_i is the context window at i and C_{ns} is a randomized context window generated from the empirical distribution of words.

Different word embeddings

Probabilistic word embeddings

- Word2Vec can be formulated as a probabilistic language model
 - Same likelihood
 - Adds a prior on the parameters, such as $w_x \sim \mathcal{N}(0, \lambda_0 I)$ for all word types $x \in W$
 - Different priors are essential
- Estimated via MAP or variational inference

Bernoulli embeddings (Rudolph et al 2016)

$$\begin{aligned}\log p(w, c \mid \mathcal{D}) &= \log p(\mathcal{D} \mid w, c) \\ &\quad + \sum_{x \in W} \lambda_1 \|w_{x,t}\|^2 \\ &\quad + \sum_{x \in W} \lambda_0 \|c_x\|^2\end{aligned}$$

Probabilistic word embeddings – advantages

- Straightforward way of including prior knowledge
 - Regularize models for data with natural divisions to subsets, for example regularize models that model different time periods
 - Incorporate additional information about the words, such as dictionary information
 - Enforce certain properties of the embedding, such as sentiment aspects
- Uncertainty estimation

Dynamic probabilistic word embeddings

- Different set of word vectors for each time period $t \in \{1, \dots, T\}$
- Random walk prior over the word vectors

$$w_{x,1} \sim \mathcal{N}(0, \lambda_0 I)$$

$$w_{x,t} \sim \mathcal{N}(w_{x,t-1}, \lambda_1 I) \text{ for } t \in \{2, \dots, T\}$$

- Spherical prior on context vectors

$$c_x \sim \mathcal{N}(0, \lambda_0 I)$$

Dynamic probabilistic word embeddings

$$\begin{aligned}\log p(w, c \mid \mathcal{D}) &= \log p(\mathcal{D} \mid w, c) \\ &\quad + \sum_{x \in W} \sum_{t=1}^T \lambda_1 \|w_{x,t+1} - w_{x,t}\|^2 + \sum_{x \in W} \sum_{t=1}^T \lambda_0 \|w_{x,t}\|^2 \\ &\quad + \sum_{x \in W} \lambda_0 \|c_x\|^2\end{aligned}$$

Applications

Dynamic probabilistic word embeddings

words with largest drift (Senate)			
IRAQ	3.09	coin	2.39
tax cuts	2.84	social security	2.38
health care	2.62	FINE	2.38
energy	2.55	signal	2.38
medicare	2.55	program	2.36
DISCIPLINE	2.44	moves	2.35
text	2.41	credit	2.34
VALUES	2.40	UNEMPLOYMENT	2.34

Figure 10: Detecting words whose usage has changed a lot over time. Rudolph et al (2018).

Polarizing words

PELP
democrat, abortion,
announce, maine, republican,
gun, illegal, republicans,
breaks, stimulus, taxes, immigration,
kentucky, accounting, wyoming
Bernoulli
hubbert, islanders, rickover,
pottawatomi, gaspee, mastercard,
morgenthau, compean, 205106150,
fairtax, vertical, peaking,
follette, isna, hubberts

Table 2: Top 15 words w with the largest distance between $\rho_{w,R}$ and $\rho_{w,D}$. PELP model above, reference model Bernoulli Embeddings below.

Polarizing words over time

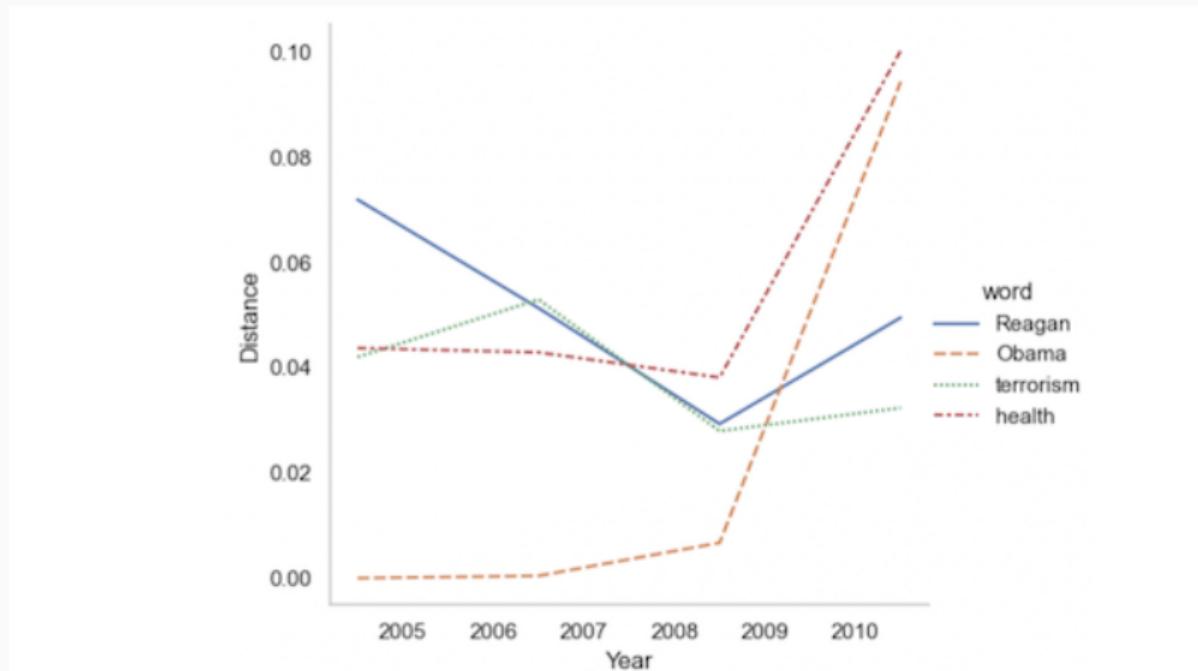


Figure 12: Partisan differences over time in the US Congress dataset.

Incorporating additional information

- Probabilistic word embeddings can be prior knowledge about the embeddings
- For example, if have a list of synonym pairs, we can inject this prior knowledge into the embeddings
 - This will make embeddings more closely resemble human evaluation of similarity
- As the dataset grows larger, there will be more word types
 - There will always be words that benefit from regularization

Incorporating additional information – better correspondence with human evaluation

Model	50M	200M
PELP	.509	.515
Dict2Vec	.512	.513
Bernoulli	.444	.465
word2vec	.445	.461

Table 5: Word similarity performance with the 50M and the 200M token partitions of the English Wikipedia, $D=100$, $\lambda_0, \lambda_1 = 1.0, 7.0$. Averages over all evaluation sets.

Uncertainty estimation

- Many applications want to draw conclusions about the data
- Random noise can be mistaken for results
- Methods to estimate embedding uncertainty
 - Probabilistic word embeddings
 - **Bootstrap**

Bootstrapping word embeddings (Antoniak and Mimno 2018)

- Divide corpus into documents that can be resampled
- Generate N bootstrap resamples
- Train the word embedding separately on all N resamples
- Calculate quantity of interest for all N word embeddings

Bootstrapping word embeddings

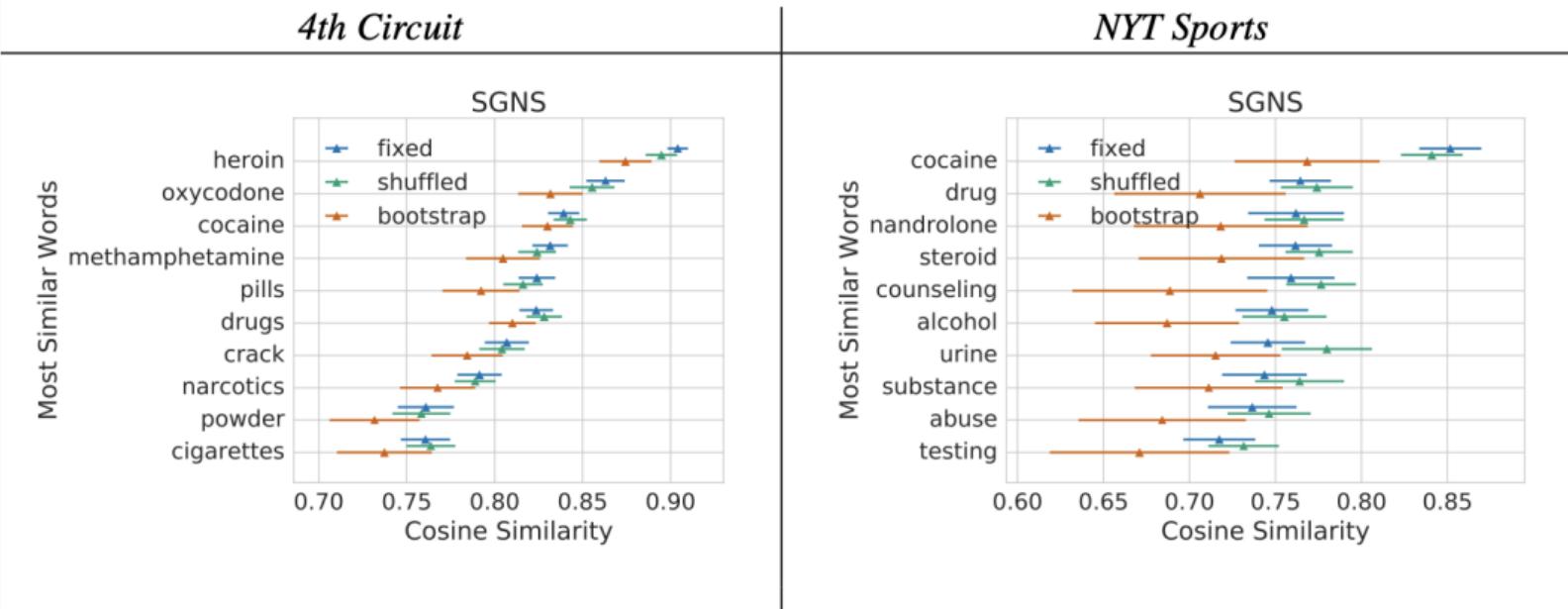


Figure 13: Most similar words to 'marijuana' in two corpora (Antoniak and Mimno 2018)

Bootstrapping word embeddings application (Garg et al 2018)

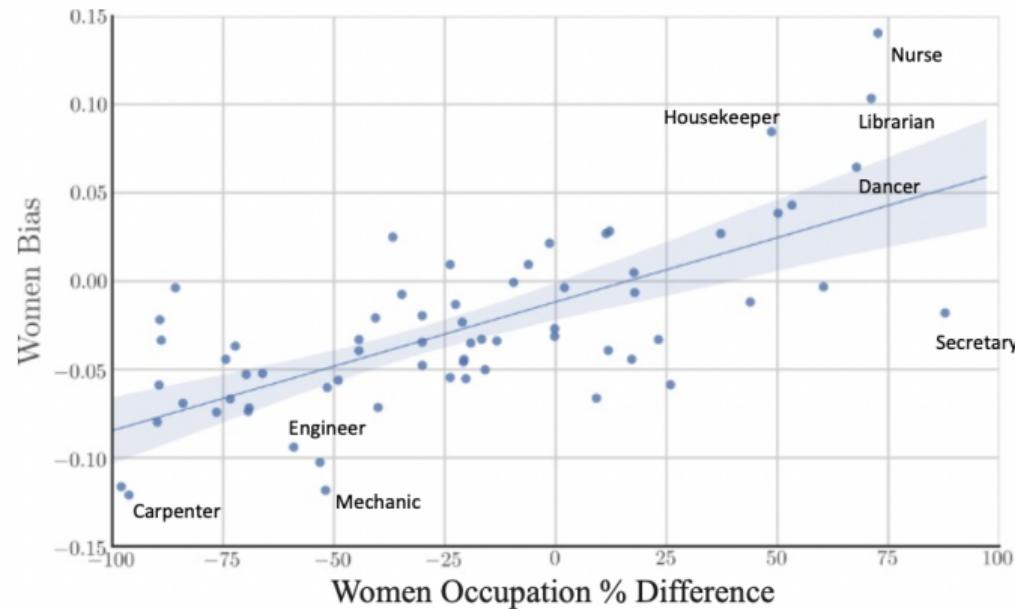


Fig. 1. Women's occupation relative percentage vs. embedding bias in Google News vectors. More positive indicates more associated with women on both axes. $P < 10^{-10}$, $r^2 = 0.499$. The shaded region is the 95% bootstrapped confidence interval of the regression line. In this single embedding, then, the association in the embedding effectively captures the percentage of women in an occupation.

Summary

- Word2vec was seminal work and its variants are still used to this day
- Probabilistic embeddings enable the incorporation of prior knowledge and uncertainty estimation
- Bootstrap can be used to estimate embedding uncertainty if the data is small