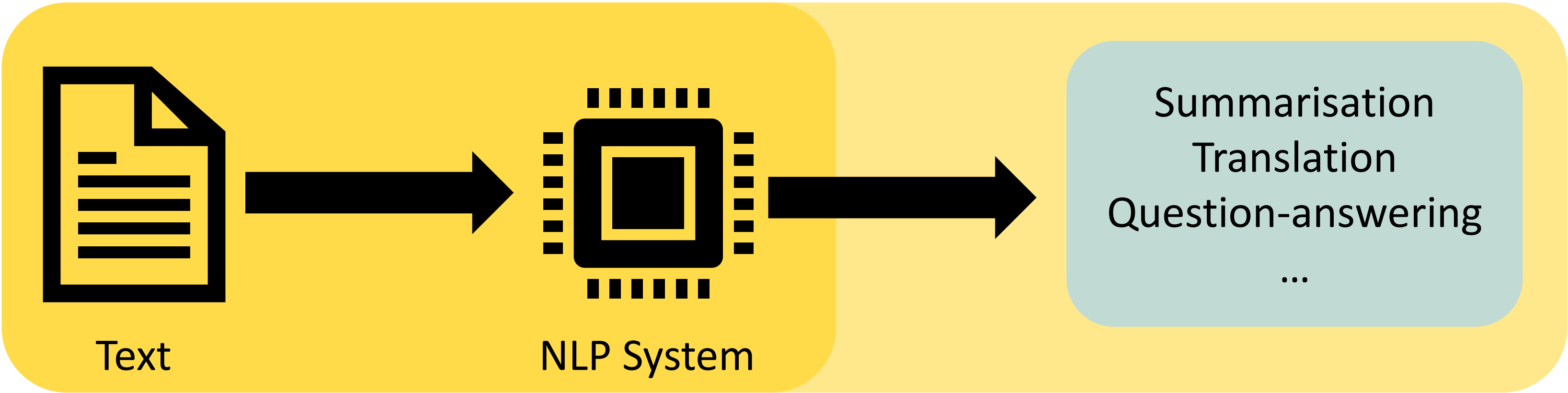
Word Embeddings

Hugh Mee Wong
NLP Group @ UU

Slides are either recycled or adapted from slides by Dong Nguyen

Natural language processing



How do we represent the meaning of words?

Agenda

- What are word embeddings?
- How do we learn word embeddings?
- How do we analyse word embeddings?

How do we represent the meaning of words?

Utrecht University

mouse 1 of 2 noun

ˈmau̇s

plural mice ˈmīs

Synonyms of *mouse* ›

Synonyms also do not take into account context

my big brother != my large brother

1 : any of numerous small rodents (as of the genus *Mus*) with pointed snout, rather small ears, elongated body, and slender tail

2 plural also **mouses :** a small mobile manual device that controls movement of the cursor and selection of functions on a computer display

3 : a timid person

4 : a dark-colored swelling caused by a blow

specifically : **BLACK EYE**

How do we know which sense to pick for a given context?

mouse 2 of 2 verb

Utrecht University

We want to capture similarity between words

- cat: dog, tiger, pet, cats
- book: novel, story, author, manuscript
- person: man, woman, child, self

And this is exactly what word embeddings will do for us!

Vector representations

- $a = [5, \ 5]$
- $b = [2, \ 1]$

These vectors are *two-dimensional* (2D)

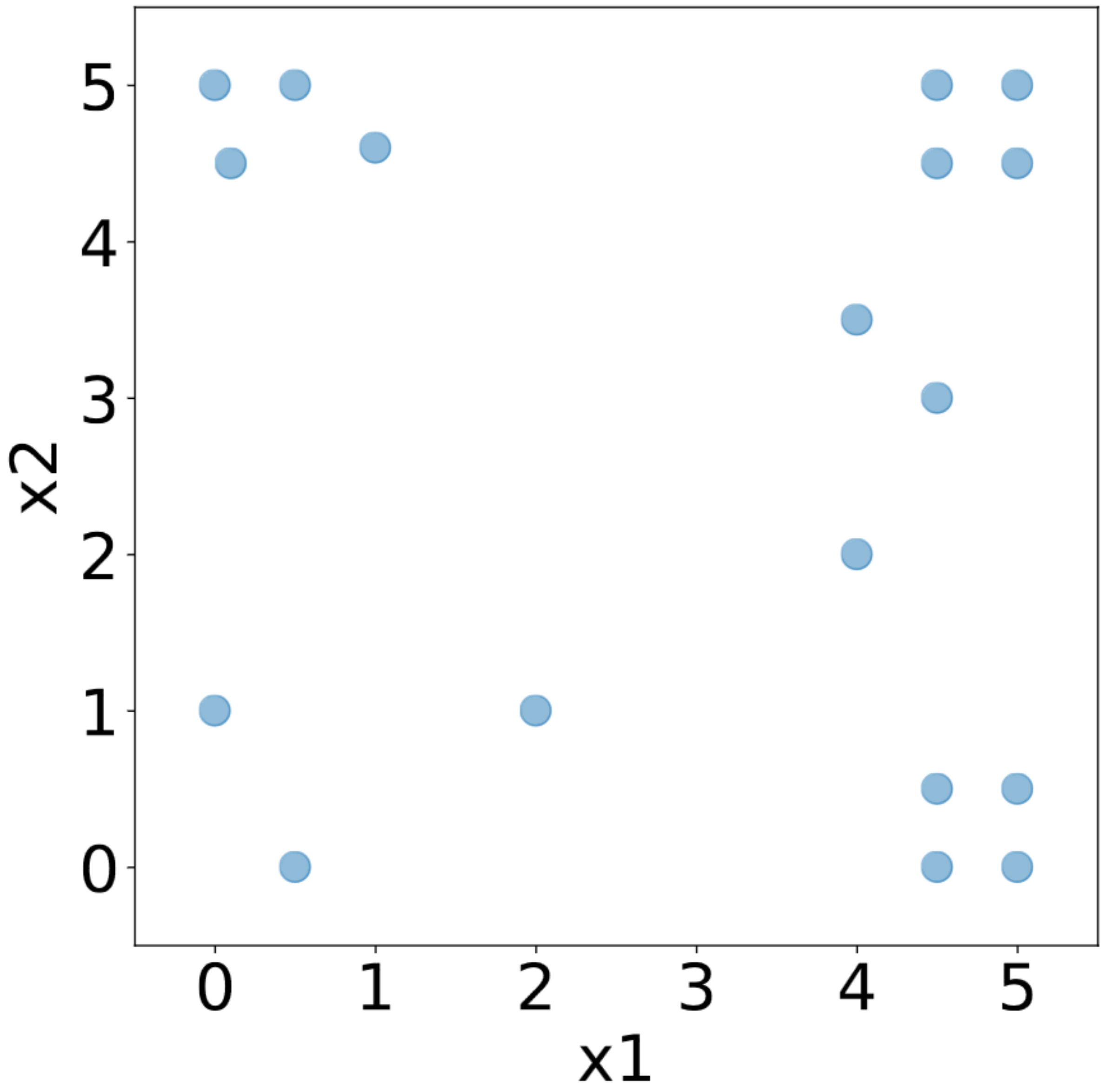


Figure by Dong Nguyen

Utrecht University

Vector representations

- $a = [5,\ 5,\ 2]$
- $b = [2,\ 1,\ 0]$

These vectors are *three-dimensional* (3D)

What if we represent words as vectors?

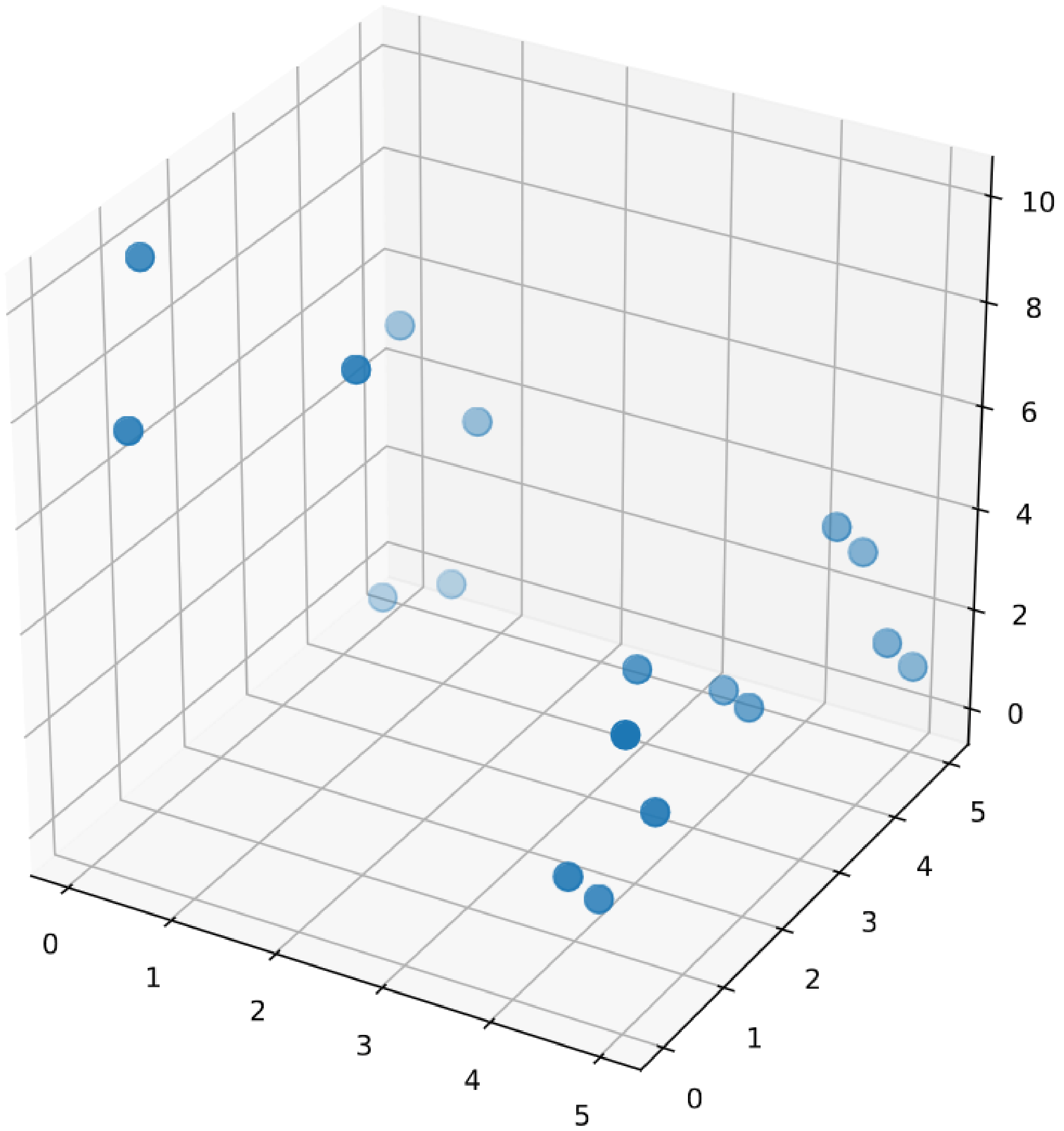
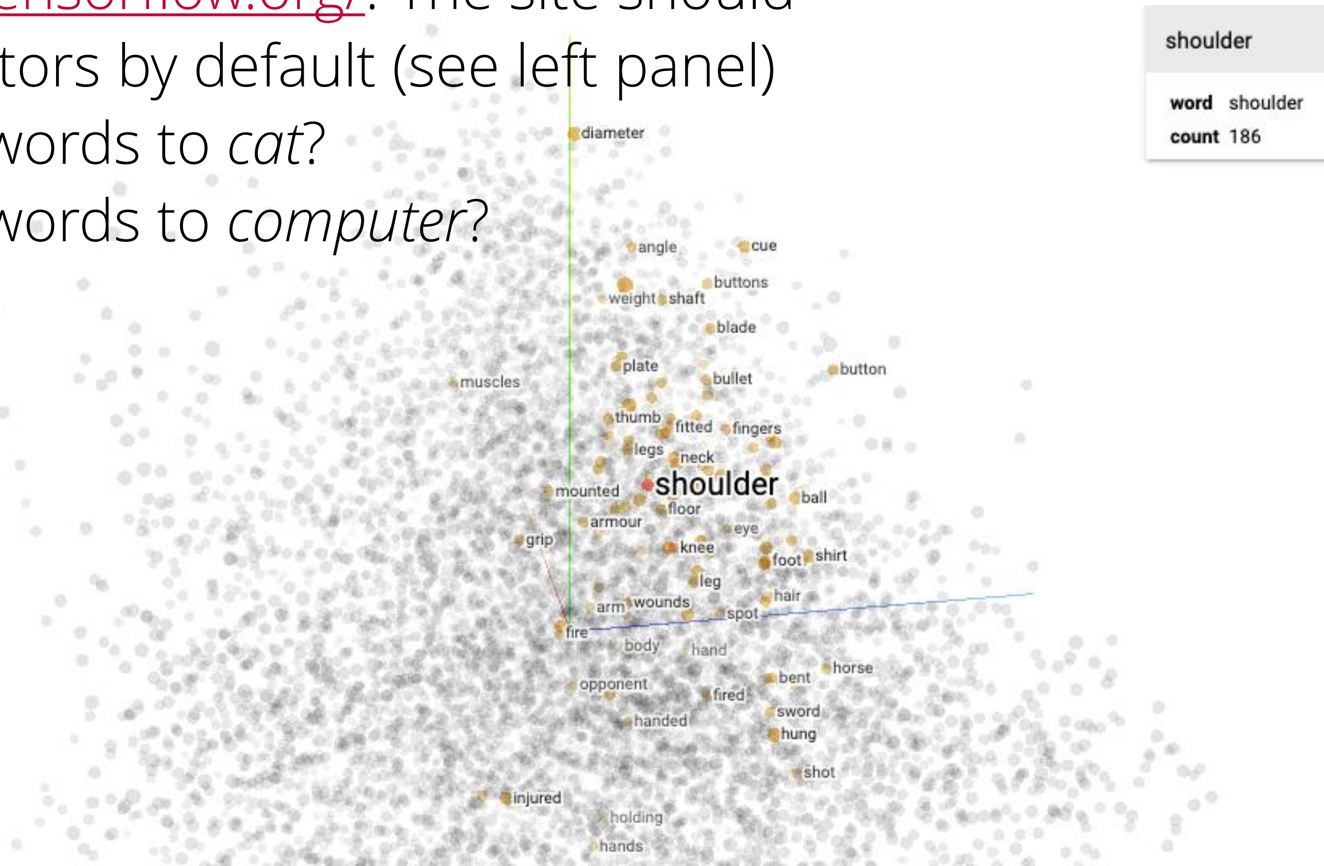


Figure by Dong Nguyen

Exercise (5 min)

- Go to https://projector.tensorflow.org/. The site should load 'Word2Vec 10K' vectors by default (see left panel)
- What are the 5 nearest words to *cat*?
- What are the 5 nearest words to *computer*?



Utrecht University

How do we represent words as vectors?

Utrecht University

One-hot encodings

We want a representation that efficiently captures similarity between words

So we need to do better!

Idea: map each word to a unique identifier (ID)

- Vector representation: all zeros, except 1 at the ID position
- High number of dimensions
- Related words have distinct vectors

cat	0	0	1	0	0	0	0
dog	0	0	0	0	1	0	0
car	0	0	0	0	0	0	1

$\text{cat} \mapsto 3$

$\text{dog} \mapsto 5$

$\text{car} \mapsto 7$

Utrecht University

Figure by Dong Nguyen

Distributional Hypothesis

You shall know a word by the company it keeps.

J.R. Firth (1957)

- Words that occur in similar contexts tend to have similar meanings
- *Distributional*: frequency/pattern of how words appear in different contexts

some believe that	**wampos**	scales have medicinal qualities
approach to fighting	**wampos**	(and general wildlife) trafficking
even though	**wampos**	scales are made of exactly the

wampos = pangolin

img source: imageBROKER / Alamy

Utrecht University

Example by Dong Nguyen

Word vectors based on co-occurences

word-document matrix
documents as context

	doc_1	doc_2	doc_3	doc_4	doc_5	doc_6	doc_7
cat	5	2	0	1	4	0	0
dog	7	3	1	0	2	0	0
car	0	0	1	3	2	1	1

word-word matrix
neighbouring words as context

	cat	dog	car	bike	book	house	tree
cat	0	3	1	1	1	2	3
dog	3	0	2	1	1	3	1
car	0	0	1	3	2	1	1

Figures by Dong Nguyen

Word vectors based on co-occurences

- Also called count-based methods
- Vectors are sparse: lots of zeros
- There are many variants

	doc_1	doc_2	doc_3	doc_4	doc_5	doc_6	doc_7
cat	5	2	0	1	4	0	0
dog	7	3	1	0	2	0	0
car	0	0	1	3	2	1	1

	cat	dog	car	bike	book	house	tree
cat	0	3	1	1	1	2	3
dog	3	0	2	1	1	3	1
car	0	0	1	3	2	1	1

Figures by Dong Nguyen

Word embeddings

cat	0.52	0.48	-0.01	···	0.28
dog	0.32	0.42	-0.09	···	0.78

- These vectors are
  - Short: typically 50-1024 dimensions
  - Dense: mostly non-zero values
- Effective for many NLP tasks
- Individual dimensions not very interpretable

Figure by Dong Nguyen

https://semantle.com

SEMANTLE

The nearest word has a similarity of **67.85**, the tenth-nearest has a similarity of **42.93** and the thousandth nearest word has a similarity of **24.86**

Game #933

Enter a word... | Guess

Hint | Give Up

Play Junior | Play Archive

FAQ

How to play?

The objective is to guess the secret word.

Each guess must be a single word. Semantle will inform you how semantically similar your guess is to the secret word.

Unlike other word games, this game is not about spelling; it's about meaning. We calculate this meaning using artificial intelligence (specifically word2vec technology).

Agenda

- What are word embeddings?
- How do we learn word embeddings?
- How do we analyse word embeddings?

How do we learn word embeddings?

How do we learn word embeddings?

Word2Vec
GloVe
fastText
BERT

cat	0.52	0.48	-0.01	...
dog	0.32	0.42	-0.09	...

Training data

- How can we train a model to learn the meaning of words?
- Which data can we use for supervised learning?

Examples

- Train a neural network to predict the next word
- Train a neural network to predict the missing word

Use the text itself as training data!
A form of *self-supervision*.

Word2Vec

"Context": distributional hypothesis!

- Target word: w_0
- Context words: $\{w_{-2}, w_{-1}, w_1, w_2\}$
- Context window: 2

the	cute	**cat**	sat	on	the	warm	mat
w_{-2}	w_{-1}	w_0	w_1	w_2	w_3	w_4	w_5

Word2Vec

Two different tasks

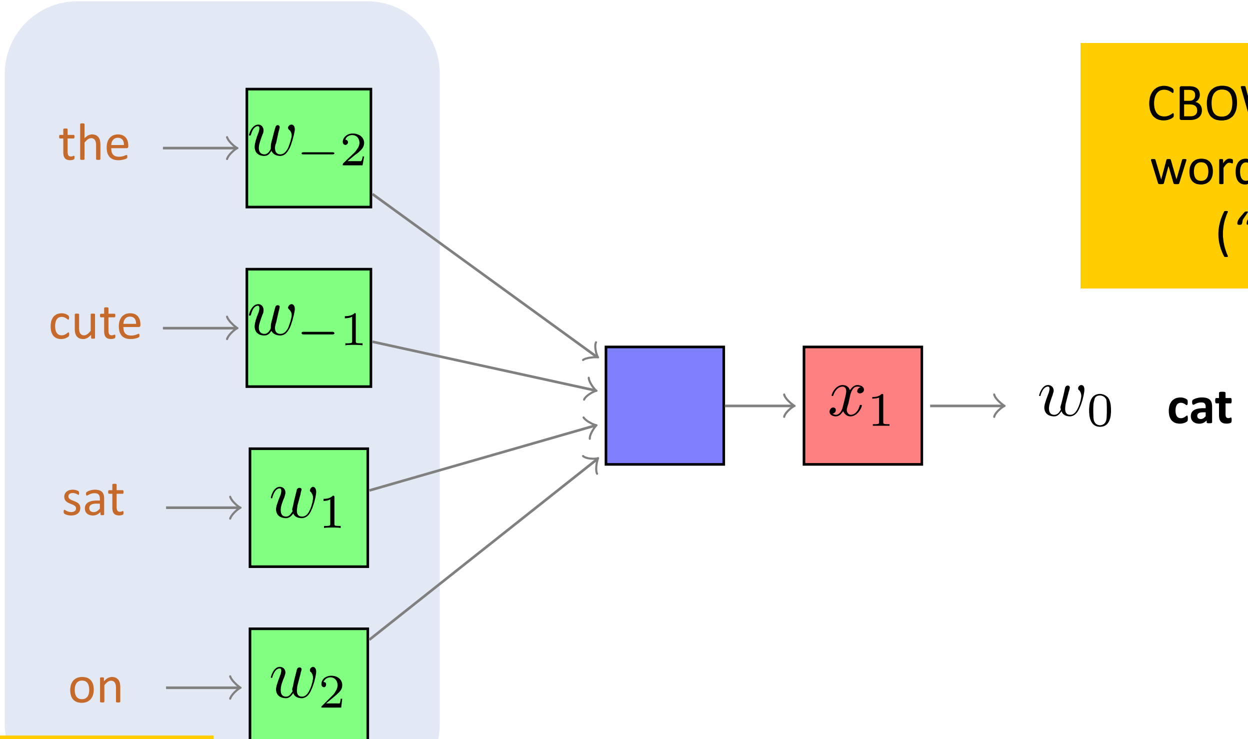
- Continuous bag-of-words (CBOW)
- Skip-gram

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space. In Proceedings of Workshop at ICLR, 2013

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed Representations of Words and Phrases and their Compositionality. In Proceedings of NIPS, 2013

the	cute	**cat**	sat	on	the	warm	mat
w_{-2}	w_{-1}	w_0	w_1	w_2	w_3	w_4	w_5

Word2Vec: Continuous Bag-of-Words (CBOW)



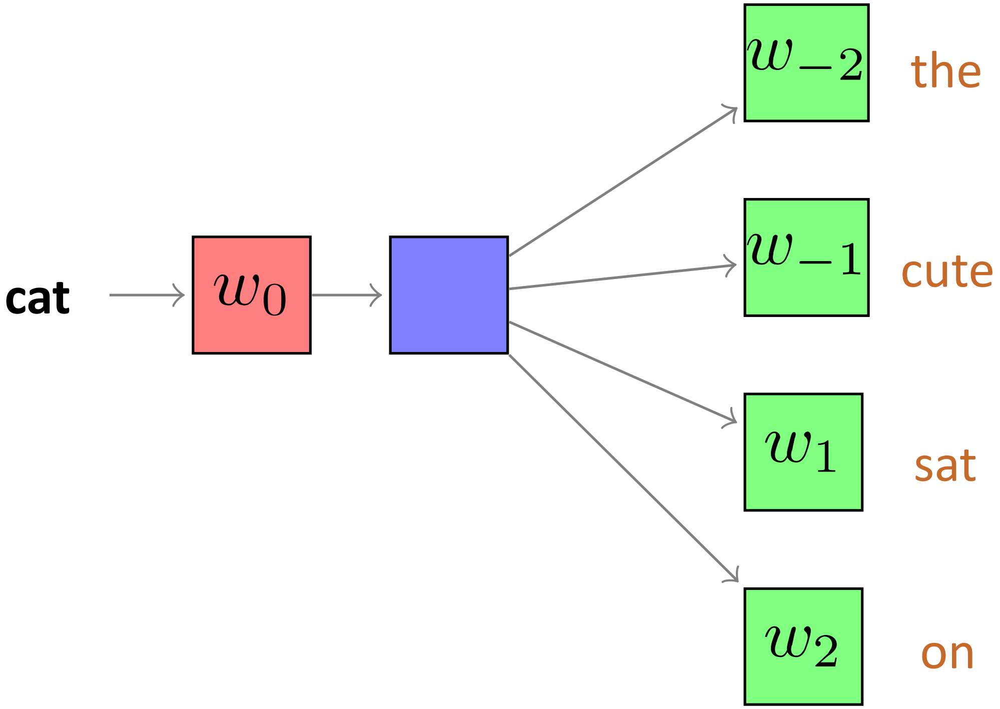
CBOW predicts the target word from context words (*“fill in the blanks”*)

Start with one-hot vectors!

the	cute	**cat**	sat	on	the	warm	mat
w_{-2}	w_{-1}	w_0	w_1	w_2	w_3	w_4	w_5

Utrecht University

Word2Vec: Skip-Gram



Skip-Gram predicts the context words from the center word

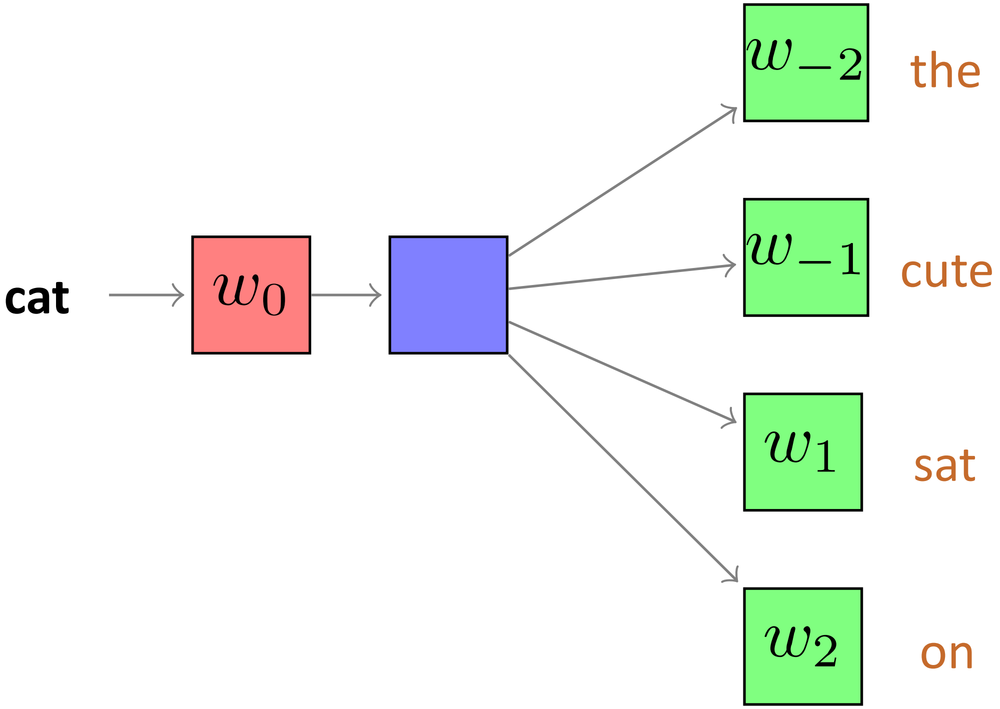
the	cute	**cat**	sat	on	the	warm	mat
w_{-2}	w_{-1}	w_0	w_1	w_2	w_3	w_4	w_5

Word2Vec: Skip-Gram (example)

the	cute	**cat**	sat	on	the	warm	mat
w_{-2}	w_{-1}	w_0	w_1	w_2	w_3	w_4	w_5

(intentionally left blank)

Word2Vec: Skip-Gram



Nice trick: negative sampling

1. Create sets containing (target, context)-pairs of positive samples and negative samples
2. Train a logistic regression model to distinguish between the positive and negative samples
3. The resulting weights are the embeddings

Positive samples:
(cat, sat)
(cat, cute)

Negative samples:
(cat, electricity)
(cat, beer)

the	cute	**cat**	sat	on	the	warm	mat
w_{-2}	w_{-1}	w_0	w_1	w_2	w_3	w_4	w_5

Word2Vec: some observations

- Operates on a local level
- Cannot deal with unseen words: *wampos*?

the	cute	**cat**	sat	on	the	warm	mat
w_{-2}	w_{-1}	w_0	w_1	w_2	w_3	w_4	w_5

Utrecht University

Global Vectors (GloVe)

GloVe also has the problem of not being able to deal with unseen words

- Creates a word-word co-occurrence matrix for all words in the document
- Values are normalised
- Training objective: learn embeddings v and w such that $v \cdot w = \log(P(v \text{ and } w \text{ co-occuring}))$

	cat	dog	car	bike	book	house	tree
cat	0	3	1	1	1	2	3
dog	3	0	2	1	1	3	1
car	0	0	1	3	2	1	1

Figure by Dong Nguyen

GloVe: Global Vectors for Word Representation. Pennington et al., EMNLP 2015

fastText

CBOW: context → target
SG: target → context

- An extension of Word2Vec
- Words are represented by a bag of n-grams
  - apple (with $n = 3$) → $\langle \text{ap}, \text{ app}, \text{ ppl}, \text{ ple}, \text{ le}\rangle$
  - $v_{\text{apple}} = v_{\langle\text{ap}} + v_{\text{app}} + v_{\text{ppl}} + v_{\text{ple}} + v_{\text{le}\rangle} + v_{\langle\text{apple}\rangle}$
- Generally used with Skip-Gram, but CBOW possible

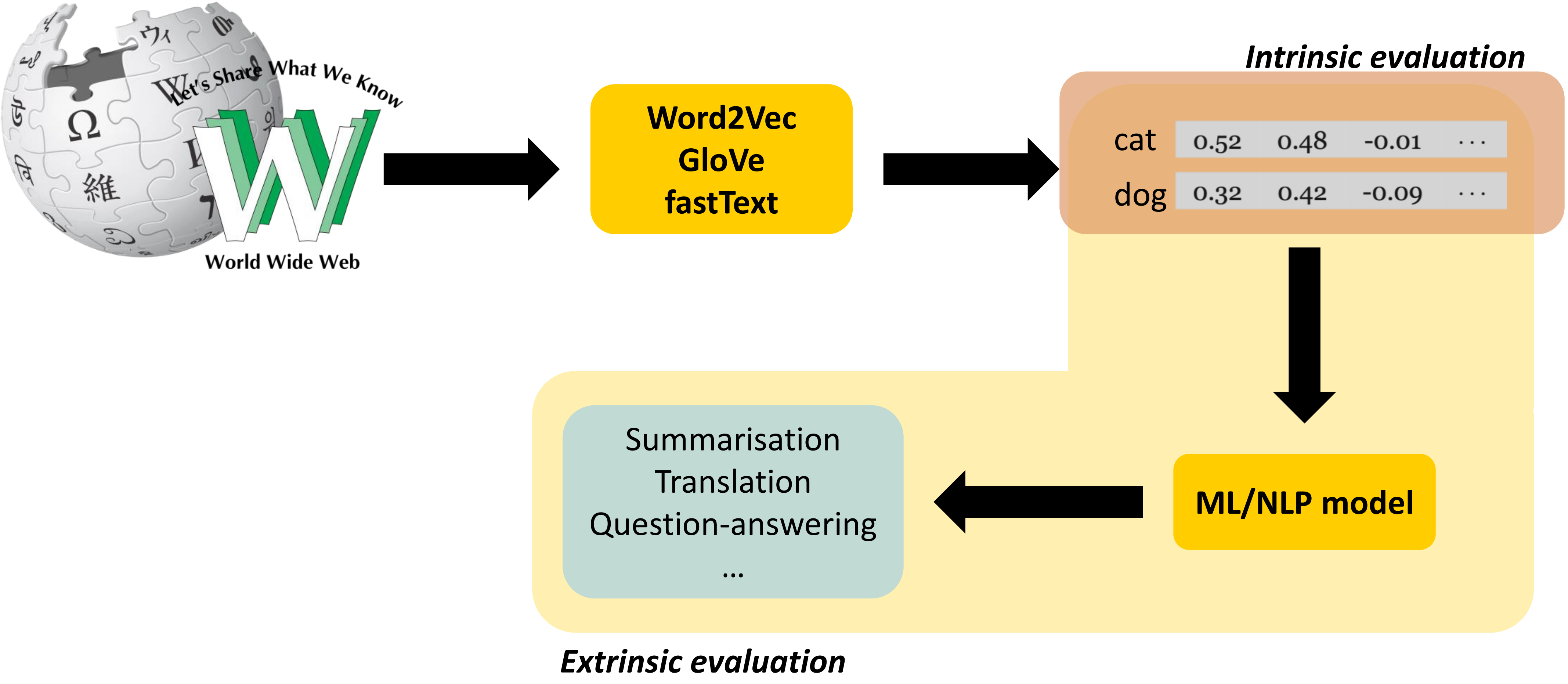
Utrecht University

Agenda

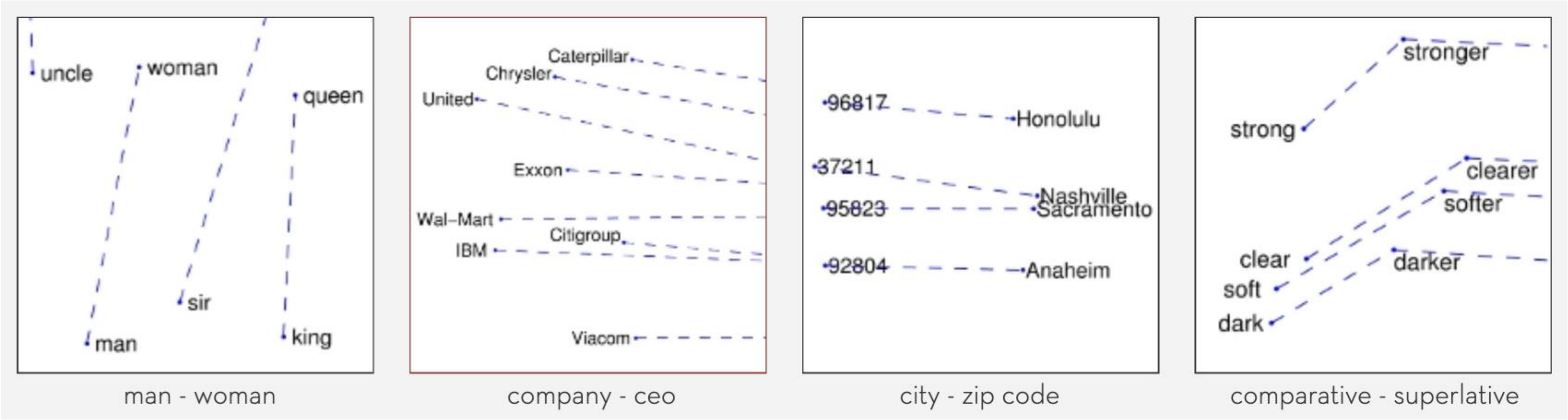
- What are word embeddings?
- How do we learn word embeddings?
- How do we analyse word embeddings?

Analysing word embeddings

Natural language processing



Analogies: GloVe



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

Analogies: Word2Vec

Warning
This generally works if we allow "cheating": exclude the original vector from the results
https://blog.esciencecenter.nl/king-man-woman-king-9a7fd2935a85

- $\text{king} - \text{man} = [4,\ 2] - [1,\ 6] = [3,\ -4]$
- $\text{king} - \text{man} + \text{woman} = [3,\ -4] + [4,\ 8] = [7,\ 4]$

Why are embeddings generally not precisely like this?

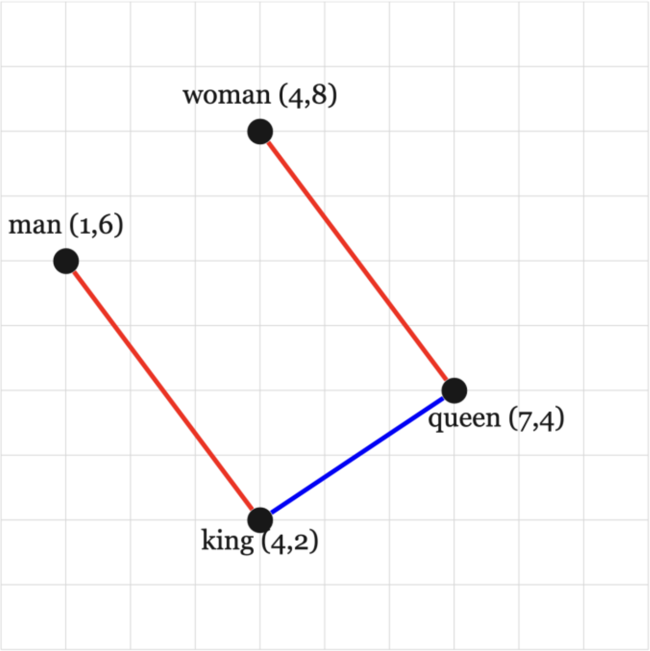


Figure by Dong Nguyen

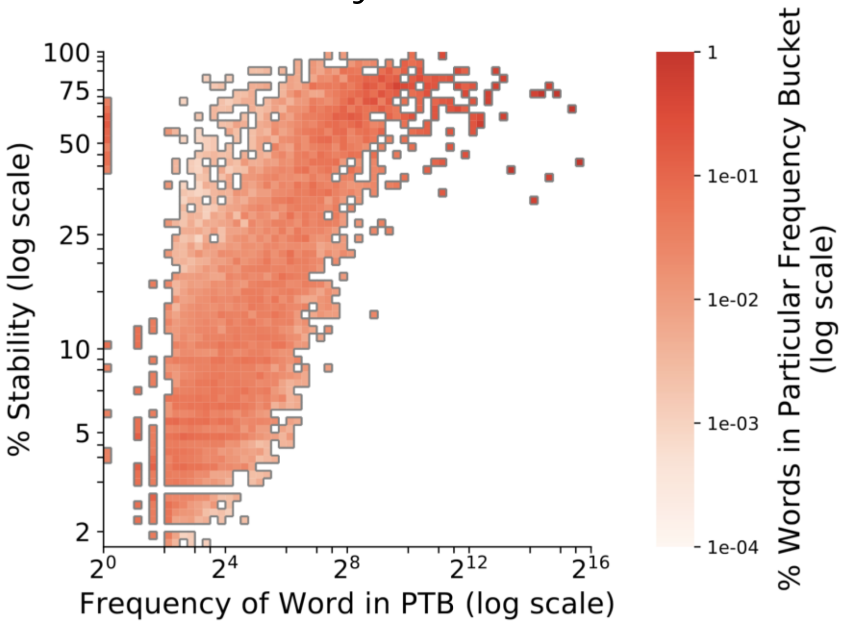
Factors influencing training

- Corpus size
- Corpus diversity
- Presence/absence of documents
- Context window (size)
- Frequency of occurrence
- Model architecture (e.g. CBOW vs Skip-Gram)
- …

Utrecht University

Stability of embeddings

- Measuring stability: look at the overlap between nearest neighbours in embedding space
- Word2Vec: lower frequency words have lower stability and higher frequency words higher



Factors Influencing the Surprising Instability of Word Embeddings, Wendlandt et al., NAACL 2018

Analogies: Word2Vec

- $\text{king} - \text{man} = [4,\ 2] - [1,\ 6] = [3,\ -4]$
- $\text{king} - \text{man} + \text{woman} = [3,\ -4] + [4,\ 8] = [7,\ 4]$

In what way can such analogies be troublesome?

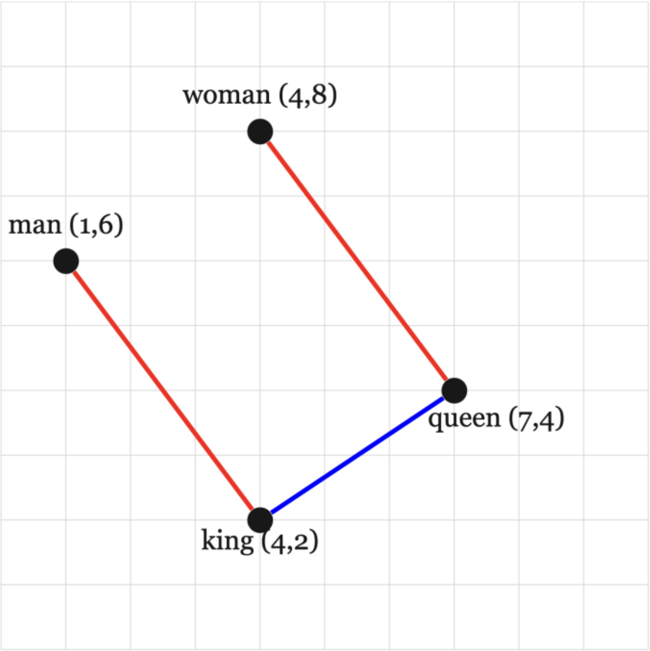


Figure by Dong Nguyen

Biases in word embeddings

- Using word embeddings to study societal trends
- Training data might contain biased language (gender bias, racial bias, …)

You shall know a word by the company it keeps.

Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings, Bolukbasi, et al. NIPS 2016

Semantics derived automatically from language corpora contain human-like biases, Caliskan, Bryson, Narayanan, Science 2017

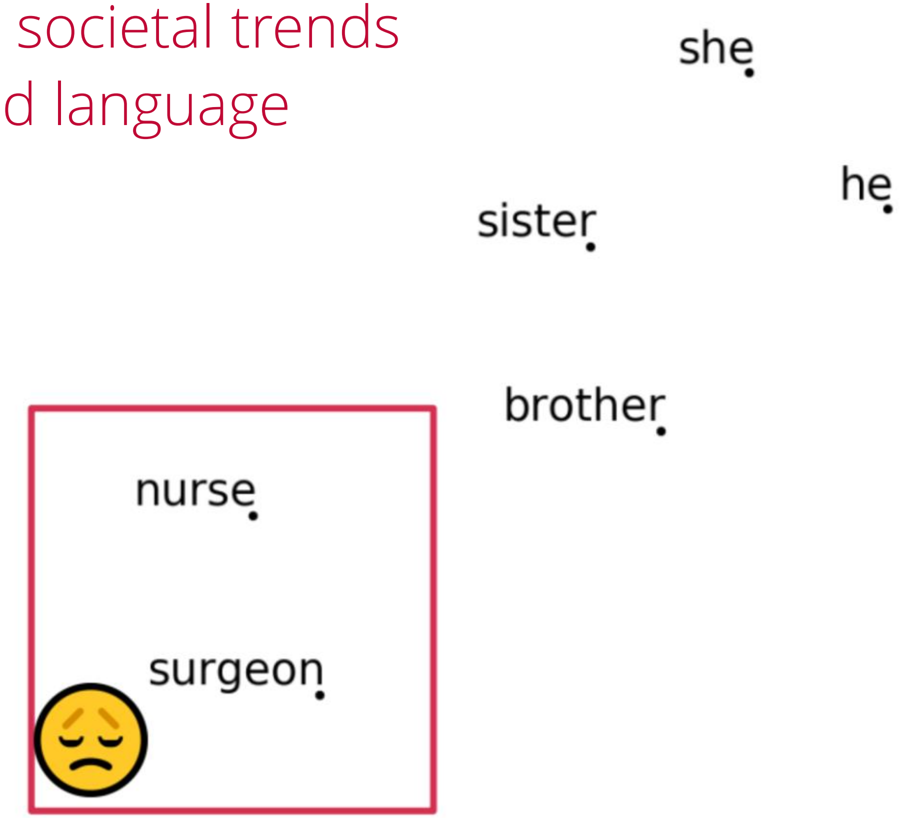


Figure by Dong Nguyen

Utrecht University

Biases in word embeddings

"I had tried building an algorithm for sentiment analysis based on word embeddings […]. When I applied it to restaurant reviews, I found it was ranking Mexican restaurants lower. The reason was not reflected in the star ratings or actual text of the reviews. It's not that people don't like Mexican food. The reason was that the system had learned the word 'Mexican' from reading the Web."

http://blog.conceptnet.io/posts/2017/conceptnet-numberbatch-17-04-better-less-stereotyped-word-vectors/

Excerpt found by Dong Nguyen

Implicit Association and Word-Embedding Association

- The Implicit Association Test (IAT) is based on response time: quicker with **John** to $\langle \text{male names}, \text{career} \rangle$ than to $\langle \text{male names}, \text{family} \rangle$
- The Word-Embedding Association Test (WEAT): cosine similarity analogous to IAT reaction time

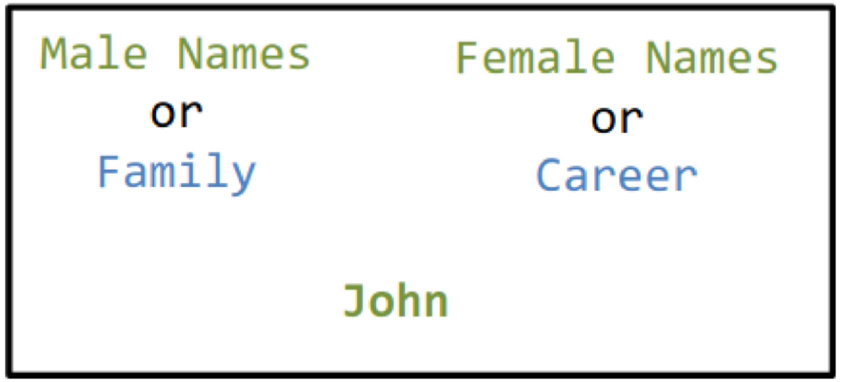


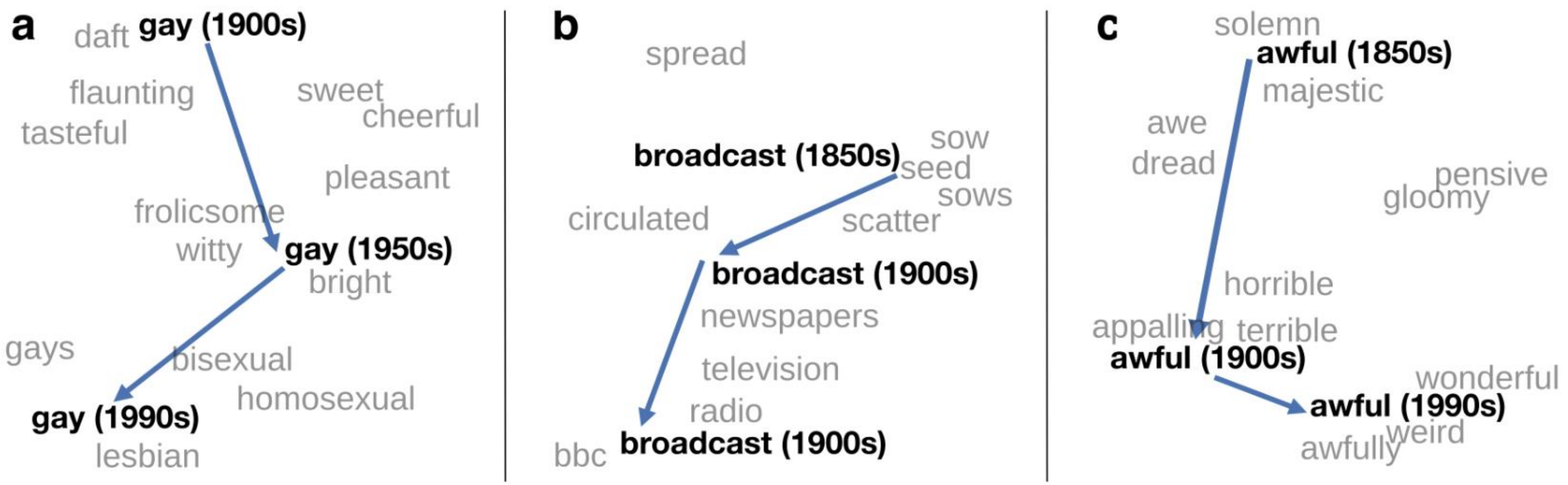
Figure by Dong Nguyen

Semantics derived automatically from language corpora contain human-like biases, Caliskan, Bryson, Narayanan, Science 2017

Utrecht University

Studying semantic change

- Using word embeddings to study societal trends



Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change, Hamilton et al., ACL 2016

Semantic change in social media

“Lit”

CBS News @CBSNews

New York City's Rockefeller Center Christmas tree lit up for the holidays Wednesday night 🎄🗽

The College Board @CollegeBoard

Good luck to all the AP students taking their AP Chemistry, AP Spanish Lit, AP German, and AP Psychology Exams today! ⚗️📖🇩🇪🧠

Lewis Hamilton @LewisHamilton

Roscoe's birthday party last night was lit 🔥

Utrecht University

Example by Dong Nguyen

Agenda

- What are word embeddings?
- How do we learn word embeddings?
- How do we analyse word embeddings?