

Embeddings in Natural Language Processing

Machine Learning for Meaning Representation

Eva Maria Vecchi

Computational Linguistics Fall School 2019
IMS, University of Stuttgart

September 11, 2019

Neural nets as machine learning algorithm

- NNs can be both supervised and unsupervised algorithms, depending on flavour:
 - multi-layer perceptron (MLP) – supervised
 - RNNs, LSTMs – supervised
 - auto-encoder – unsupervised
 - self-organising maps – unsupervised

Neural networks: a motivation



How to recognise digits?

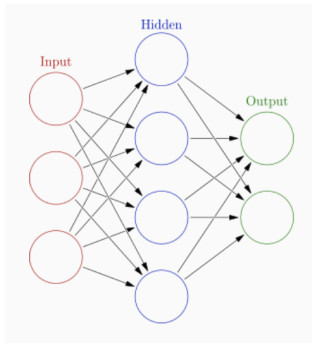
- Rule-based: a '1' is a vertical bar. A '2' is a curve to the right going down towards the left and finishing in a horizontal line...
- Feature-based: number of curves? of straight lines? directionality of the lines (horizontal, vertical)?
- Well, that's not gonna work...

Learning your own features

- We don't know what people pay attention to when recognizing digits (which features to use).
- Don't try to guess. Just let the system decide for you.
- A nice architecture to do this is the neural network:
 - Good for learning visual features.
 - Also good for learning latent linguistic features (remember SVD?)

Neural Nets

- A neural net is a set of interconnected neurons organised in 'layers'.
- Typically, we have one input layer, one output layer and a number of hidden layers in-between:

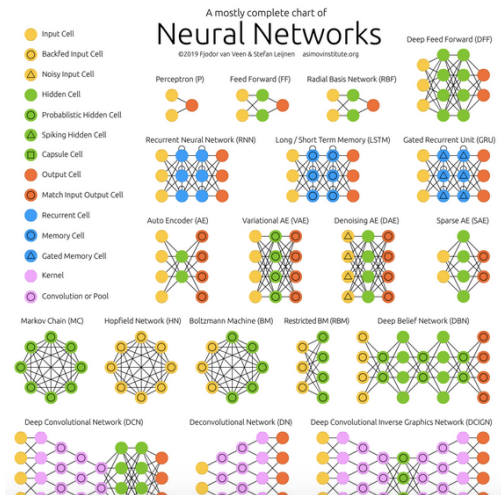


This is a multi-layer perceptron (MLP).

By Glosser.ca - Own work, Derivative of File:Artificial neural network.svg, CC BY-SA 3.0,

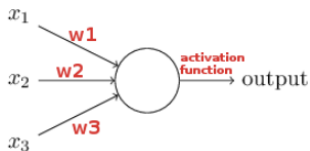
<https://commons.wikimedia.org/w/index.php?curid=24913461>

Neural network zoo



Go visit <http://www.asimovinstitute.org/neural-network-zoo/> – very cool!

The artificial neuron



- The output of the neuron (also called ‘node’ or ‘unit’) is given by:

$$a = \varphi \left(\sum_{j=0}^m w_j x_j \right)$$

where φ is the activation function.

- If this output is over a threshold, the neuron ‘fires’.

A (simplified) example

- Should you bake a cake? It depends on the following features:
 - Wanting to eat cake (0/+1)
 - Having a new recipe to try (0/+1)
 - Having time to bake (0/+1)

A (simplified) example

- Should you bake a cake? It depends on the following features:
 - Wanting to eat cake (0/+1)
 - Having a new recipe to try (0/+1)
 - Having time to bake (0/+1)
- How much weight should each feature have?
 - You like cake. Very much. *Weight: 0.8*
 - You need practice, as become a pastry chef is your professional plan B. *Weight: 0.3*
 - Baking a cake will take time away from your computational linguistics project, but you don't really care. *Weight: 0.1*

A (simplified) example

- We'll ignore φ for now, so our equation for the output of the neuron is:

$$a = \sum_{j=0}^m w_j x_j$$

- Assuming you want to eat cake (+1), you have a new recipe (+1) and you don't really have time (0), our output is:

$$0.8 * 1 + 0.3 * 1 + 0.1 * 0 = 1.1$$

- Let's say our threshold is 0.5, then the neuron will fire (output 1). You should definitely bake a cake.

From threshold to bias

- We can write $\sum_{j=0}^m w_j x_j$ as the dot product $\vec{w} \cdot \vec{x}$
- We usually talk about bias rather than threshold – which is just a way to move the value to the other side of our inequality:
 - if $\vec{w} \cdot \vec{x} > t$, then 1 (fire) else 0
 - if $\vec{w} \cdot \vec{x} - t > 0$, then 1 (fire) else 0
- The bias is a ‘special neuron’ in each layer, with a connection to all other units in that layer.

But hang on...

- Didn't we say we didn't want to encode features? Those inputs look like features...
- Right. In reality, what we will be inputting are not human-selected features but simply a vectorial representation of our input.
- Typically, we have one neuron per value in the vector.
- Similarly, we have a vectorial representation of our output (which could be as simple as a single neuron representing a binary decision).

Neural nets and word meaning

Instead of creating the co-occurrence matrix and then reducing its dimensions. . .

Why not learn the compressed representations directly from the data?

Word embeddings

- give words from a vocabulary as input to a (feed-forward) neural network
- embed them as vectors into a lower dimension space
- fine-tune through back-propagation
- \rightarrow yields word embeddings as the weights of the first layer, usually referred to as *Embedding Layer*
- The objective is to create word representations (embeddings) that are good at predicting the surrounding context.

Distributional vs. Distributed Representation

Distributional Representation

- captures linguistic distribution of each word in form of a high-dimensional numeric vector
- typically based on co-occurrence counts (aka “count” models)
- based on distributional hypothesis: similar distribution $\tilde{\text{similar}}$ meaning (similar distribution = similar representation)

Distributional vs. Distributed Representation

Distributional Representation

- captures linguistic distribution of each word in form of a high-dimensional numeric vector
- typically based on co-occurrence counts (aka “count” models)
- based on distributional hypothesis: similar distribution $\tilde{\text{similar}}$ meaning (similar distribution = similar representation)

Distributed Representation

- **sub-symbolic**, compact representation of words as dense numeric vector
- meaning is captured in different dimensions and it is used to predict words (aka “predict” models)
- similarity of vectors corresponds to similarity of the words
- aka **word embeddings**

Methods to train word embeddings

- First and most used: [word2vec](#) (see below)
- [FastText](#): similar to word2vec but trained on character n-grams instead of words
- [GloVe](#): Global Vectors - first uses co-occurrence matrix, calculates ratios of probabilities; trained with log-bilinear regression model
- [ELMo](#), [BERT](#), [Flair](#): Contextualized word embeddings
- among many others...

word2vec

Mikolov et al, 2013

Framework for learning word embeddings; main idea:

- takes words from a very large corpus of text as input (unsupervised)
- learn a vector representation for each word to predict between every word and its context
- fully connected feed-forward neural network with one hidden layer
- Two main algorithms:
 - **Continuous Bag of Words (CBOW)**: predicts center word from the given context (sum of surrounding words vectors)
 - **Skip-gram**: predicts context taking the center word as input

Center word and context

- Embedding models consider the history (previous words) **and** the future (following words) of a center word¹
- The number of words considered is called the **window size** (standard size = 5)
- **Importance of window size**
 - “**Australian scientists discover stars with** telescopes.”
 - **context window size 2 center word context window size 2**
 - Note: different meaning of “stars” with and without telescope

¹Unlike language models, who only look at past words for predictions

CBOW

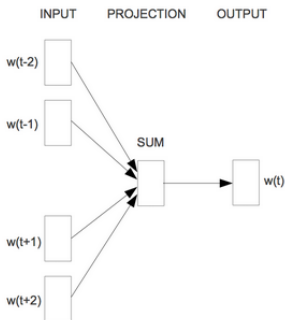


Figure 4: Continuous bag-of-words (Mikolov et al., 2013)

- It uses continuous representations whose order is of no importance
- CBOW can be seen as a precognitive language model
- Objective function similar to a language model

$$J_{\theta} = \frac{1}{T} \sum_{t=1}^T \log p(w_t : | : w_{t-n}, \dots, w_{t-1}, w_{t+1}, \dots, w_{t+n})$$

Skip-gram

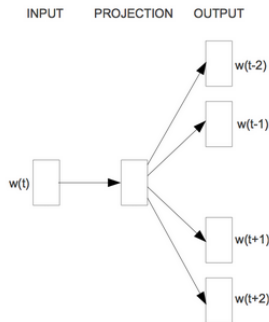


Figure 5: Skip-gram (Mikolov et al., 2013)

- Instead of using the surrounding words to predict the centre word as with CBOW, skip-gram uses the centre word to predict the surrounding words
- objective thus sums the log probabilities of the surrounding n words to the left and to the right of the target word w_t

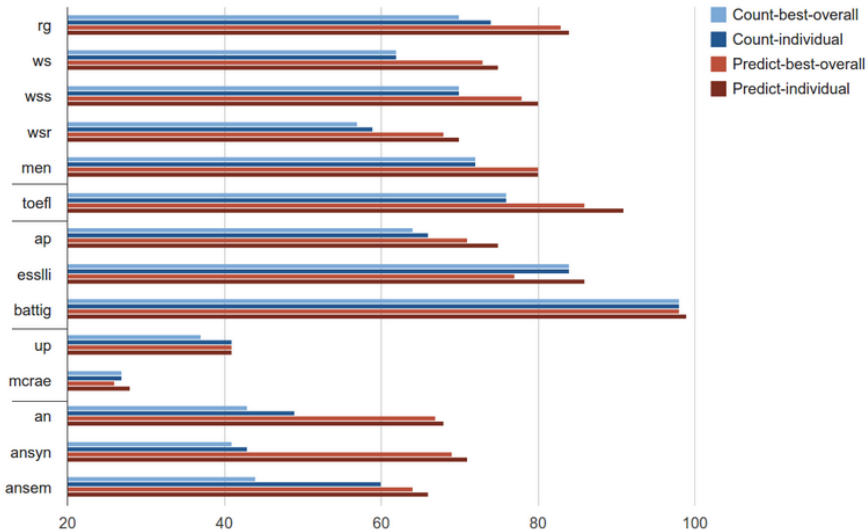
$$J_{\theta} = \frac{1}{T} \sum_{t=1}^T \sum_{-n \leq j \leq n, j \neq 0} \log p(w_{t+j} : | : w_t)$$

Don't count, predict!

Baroni et al., ACL 2014

Don't count, predict!

Baroni et al., ACL 2014



Evaluation: Analogy

$$\overrightarrow{\text{king}} - \overrightarrow{\text{man}} + \overrightarrow{\text{woman}}$$

Evaluation: Analogy

$$\vec{king} - \vec{man} + \vec{woman}$$

1. queen
 2. monarch
 3. princess
 4. kings ...
- Word Analogy Task: a is to b , as c is to ?
 - How many word analogies can the trained embeddings predict correctly?

Analogies can be funny...

<http://graceavery.com/word2vec-fish-music-bass>

Evaluation: Topics

$$\overrightarrow{\text{hungry}} + \overrightarrow{\text{monster}}$$

Evaluation: Topics

$$\overrightarrow{\text{hungry}} + \overrightarrow{\text{monster}}$$

- monsters, beast, ravenous, creature, monstrous, starving, famished, hunger, thirsty, cannibal, ravening, ...
- Pretty good, but a lot of emphasis on (food-like) hunger

Evaluation: Topics

$$\overrightarrow{\text{hungry}} + \overrightarrow{\text{monster}}$$

- monsters, beast, ravenous, creature, monstrous, starving, famished, hunger, thirsty, cannibal, ravening, ...
- Pretty good, but a lot of emphasis on (food-like) hunger

$$\overrightarrow{\text{hungry}} + \overrightarrow{\text{monster}} - \overrightarrow{\text{food}}$$

- Refine topic by removing too generically food-related words
- monstrous, monsters, beast, ravenous, ogre, **monster-like**, **three-headed**, **child-eating**, creature, mad, **bloodthirsty**, frightened, ...

Evaluation: Descriptions

2. Choose the most appropriate **verb** from the list in the boxes for each of the situations below. Use each of the words only **once**.

sidle-	amble	totter	trudge
stagger	strut	march	lurch

- a) A baby just learning to walk totter
- b) A drunk man walking down the street lurch
- c) A weary farmer returning home through the mud trudge
- d) Two teenagers guiltily approaching someone sidle
- e) Someone who has just been shot stagger
- f) A lazy walk in the country amble
- g) A model walking down the catwalk strut
- h) Someone going to the manager of a hotel to make a strong complaint march

Choose the most appropriate word for each space from the bracketed words:

Evaluation: Descriptions

- **baby learning to walk**
 - amble, **totter**, trudge, stagger, march, strut, sidle, lurch
- **drunk man walking down street**
 - stagger, amble, trudge, march, trotter, sidle, strut, **lurch**
- **weary farmer returning home through mud**
 - **trudge**, stagger, amble, lurch, march, totter, strut, sidle
- **two teenagers guiltily approaching someone**
 - **sidle**, amble, stagger, totter, trudge, lurch, march, strut
- **someone who has just been shot**
 - **stagger**, march, trudge, sidle, lurch, amble, strut, totter
- **a lazy walk in the country**
 - **amble**, trudge, stagger, march, totter, sidle, strut, lurch

The unreasonable effectiveness ...

- In 2015, Andrej Karpathy wrote a blog entry which became famous: *The unreasonable effectiveness of Recurrent Neural Networks*².
- How a simple model can be unbelievably effective.

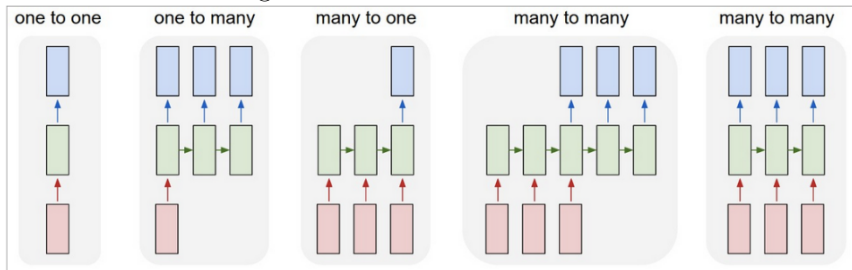
²<https://karpathy.github.io/2015/05/21/rnn-effectiveness/>

Recurrence

- Feedforward NNs which take a vector as input and produce a vector as output are limited.
- Putting recurrence into our model, we can now process *sequences* of vectors, at each layer of the network.

Architectures

What might these architectures be used for?



<https://karpathy.github.io/2015/05/21/rnn-effectiveness/>

Language Modeling

- A language model (LM) is a model that computes the probability of a sequence of words, given some previously observed data.
- LMs are used widely, for instance in predictive text on your smartphone:

Today, I am in (my/bed/Munich/Ulaanbaatar).

The Markov assumption

- Let's assume the following sentence:

I am in Rome.

- We are going to use the chain rule for calculating its probability:

$$P(A_n, \dots, A_1) = P(A_n | A_{n-1}, \dots, A_1) \cdot P(A_{n-1}, \dots, A_1)$$

- For our example,

$$P(I, am, in, Rome) = P(Rome | in, am, I) \cdot P(in | am, I) \cdot P(am | I) \cdot P(I)$$

The Markov assumption

- The problem is, we cannot easily estimate the probability of a word in a long sequence.
- There are too many possible sequences that are not observable in our data or have very low frequency:

$$P(\text{Rome}|\text{in}, \text{am}, \text{I}, \text{today}, \text{but}, \text{yesterday}, \text{there}, \dots)$$

- So we make a simplifying Markov assumption:

$$P(\text{Rome}|\text{in}, \text{am}, \text{I}) \approx P(\text{Rome}|\text{in})(\text{bigram})$$

or

$$P(\text{Rome}|\text{in}, \text{am}, \text{I}) \approx P(\text{Rome}|\text{in}, \text{am})(\text{trigram})$$

- That is, we are not taking into account *long-distance dependencies* in language.
- Trade-off between accuracy of the model and trainability.

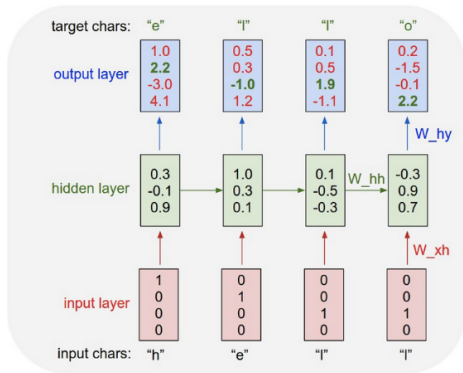
LMs as a generative model

- In your smartphone, the LM does not just calculate a sentence probability, it suggests the next word to what you're writing.
- Given the sequence *I am in*, for each word w in the vocabulary, the LM can calculate:

$$P(w|in, am, I)$$

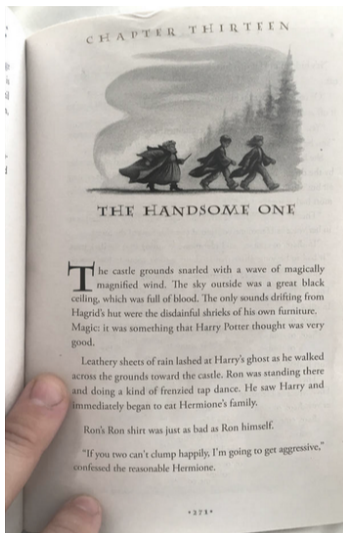
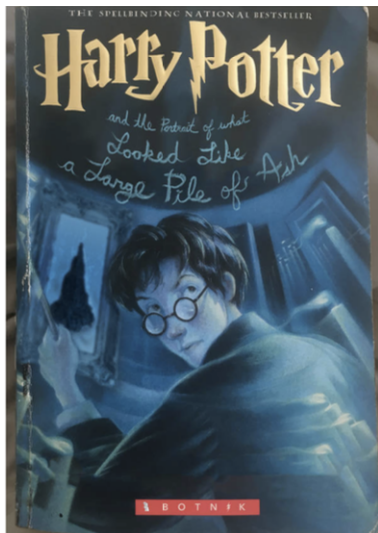
- The word with the highest probability is returned

Language modeling with RNNs



- The *sequence* given to the RNN is equivalent to the n-gram of a language model.
- Given a word or character, it has to predict the next one.

Example: Rewriting Harry Potter



<http://www.botnik.org/content/harry-potter.html>

Types of recurrent NNs

- **RNNs (Recurrent Neural Networks)**: the original version. Simple architecture but does not have much memory.
- **LSTMs (Long Short-Term Memory Networks)**: an RNN able to remember and forget selectively.
- **GRUs (Gated Recurrent Units)**: a variation on LSTMs.

Trends and future directions

Subword-level embeddings

- Word embeddings have been augmented with subword-level information for many applications
 - e.g. named entity recognition, part-of-speech tagging, dependency parsing, and language modelling
- Often use a CNN or a BiLSTM
 - input: characters of a word
 - output: a character-based word representation
- Character n-grams features have been shown to be more powerful than composition functions over individual characters
- Even smaller!
 - subword units based on *byte-pair encoding* perform well for machine translation and entity typing
 - easily learned, but no real advantage over character-based representations for most tasks

OOV handling

- Main problem with pre-trained embeddings: unable to deal with out-of-vocabulary (OOV) words
- One solution: subword-level embeddings – some success
- Recent approaches aim to generate OOV embedding on-the-fly
 - e.g. initialize the embedding of OOV words as the sum of their context words, then rapidly refine only the OOV embedding with a high learning rate (Herbelot & Baroni, 2017)

Evaluation

- Evaluation of pre-trained embeddings remain a contentious issue
 - word similarity and analogy datasets have been shown to only correlate weakly with downstream performance
- The RepEval Workshop at ACL 2016 exclusively focused on better ways to evaluate pre-trained embeddings
- So far, best way to evaluate: extrinsic evaluation on downstream tasks

Multi-sense embeddings

- Common criticism: embeddings are unable to capture **polysemy**
- Most approaches for learning multi-sense embeddings solely evaluate on word similarity
- *However*, strong results in Neural Machine Translation → models are expressive enough to contextualize and disambiguate words in context
- Yet, still need to understand if and how models are disambiguating, and how to improve if necessary

Phrases and multi-word expressions

- Problem: Embeddings fail to capture the meanings of phrases and multi-word expressions
 - Eg. *kick the bucket*, *work hard*, *take a seat*, etc.
- Some attempts to build phrase embeddings or better learn compositional and non-compositional phrases
- explicitly modelling phrases has so far not shown significant improvements on downstream tasks that would justify the additional complexity
- a better understanding of how phrases are modelled in neural networks would allow methods that augment the capabilities of our models to capture compositionality and non-compositionality of expressions

Bias

- Word embeddings trained on, e.g., Google News articles exhibit female/male gender stereotypes to a disturbing extent (Bolukbasi et al., 2016)
- Bias in models becoming big issue in the field
- What other biases are captured in embeddings, and how best to remove bias?

Temporal dimension

- Word meanings are subject to continuous change
- We can consider the temporal dimension and the diachronic nature of words
- Useful to reveal laws of semantic change, model temporal word analogy or relatedness, and capture dynamics of semantic relations.

Lack of theoretical understanding

- Little work on gaining a better theoretical understanding of the word embedding space and its properties
 - e.g. that summation captures analogy relations
- Some insights explored in Arora et al. (2016), Gittens et al. (2017), Mimno & Thompson (2017)

Task and domain-specific embeddings

- Downside of pre-trained embeddings: news data for training often different than data for tasks
 - also hard to come by millions of unlabelled docs in most target domains
- Some attempts to adapt pre-trained embeddings to capture characteristics of target domain, and retain relevant existing knowledge
 - e.g. Lu & Zheng (2017): regularized skip-gram model to learn cross-domain embeddings
- Or, use existing knowledge from semantic lexicons to augment pre-trained embeddings with relevant information
 - e.g. retro-fitting (Faruqui et al., 2015), injecting prior knowledge like monotonicity (You et al., 2017) and word similarity (Niebler et al., 2017), etc.

Transfer learning

- Aim to create *contextualized* word vectors (rather than adapting them)
 - augment word embeddings with embeddings based on hidden states of models pre-trained for certain tasks
 - e.g. machine translation or language modeling
- **Bidirectional Encoder Representations from Transformers (BERT)**
 - introduced by Google AI in 2018
 - first deeply *bidirectional*, *unsupervised* language representation, pre-trained using only a plain text corpus
 - stunning results on numerous NLP tasks

Embeddings for multiple languages

- Goal to create multilingual word embeddings
- Methods being developed that learn cross-lingual representations with as few parallel data as possible
- Some work also aims to learn multilingual embeddings *without* parallel data
- Note issue of training for low-resource languages

Machine Learning, Embeddings, and the World

Machine Learning and the World

“Society is now at a crucial juncture in determining how to deploy AI-based technologies in ways that promote rather than hinder democratic values such as freedom, equality, and transparency.”

Artificial Intelligence and Life in 2030

https://ai100.stanford.edu/sites/default/files/ai_100_report_0831fn1.pdf

Brave New World

- “As cars will become better drivers than people, city-dwellers will own fewer cars, live further from work, and spend time differently, leading to an entirely new urban organization.”
- “Though quality education will always require active engagement by human teachers, AI promises to enhance education at all levels, especially by providing personalization at scale.”
- “As dramatized in the movie *Minority Report*, predictive policing tools raise the specter of innocent people being unjustifiably targeted. But well-deployed AI prediction tools have the potential to actually remove or reduce human bias.”

Cambridge Analytica

- *The* ML scandal of the last three years ...
- Used millions of Facebook profiles to (allegedly) influence US elections, Brexit referendum, and many more political processes around the world
- Provided user-targeted ads after classifying profiles into psychological types
- Closed and reopened under the name *Emerdata*

Palantir Technologies

- Named after Lord of the Rings' *Palantír* (all-seeing eye)³
- Two projects: *Palantir Gotham* (for defense and counter-terrorism) and *Palantir Metropolis* (for finance)
- Billion-dollar company accumulating data from every possible source, and making predictions from that data

³Forbes' *How A 'Deviant' Philosopher Built Palantir, A CIA-Funded Data-Mining Juggernaut*

Predictive policing

- RAND Corporation: a think tank originally created to support US armed forces
- RAND Report on predictive policing:⁴
 - *“Predictive policing – the application of analytical techniques, particularly quantitative techniques, to identify promising targets for police intervention and prevent or solve crime – can offer several advantages to law enforcement agencies. Policing that is smarter, more effective, and more proactive is clearly preferable to simply reacting to criminal acts. Predictive methods also allow police to make better use of limited resources.”*

⁴https://www.rand.org/pubs/research_briefs/RB9735.html

ML and predicting

- ML algorithms are fundamentally about *predictions*
- Must ask ourselves:
 - What is the quality of those predictions?
 - Do we even want to make those predictions?
- If the possible futures of an individual become part of the representation of that individual *here* and *now*, what does it mean for the way they are treated by institutions?

NLP and predicting

- The above examples have to do with real people's data
- What does it all have to do with language and words?
- Language gives us direct access to some of our conceptual structure. It models *how we see the world*.

The revelation ...

**MIT
Technology
Review**

Topics

Space

How Vector Space Mathematics Reveals the Hidden Sexism in Language

As neural networks tease apart the structure of language, they are finding a hidden gender bias that nobody knew was there.

by **Emerging Technology from the arXiv**

Jul 27, 2016

Women are ... *what?!?*

Bolukbasi et al., NIPS, 2016

man : king :: woman : x

x = queen

Paris : France :: Tokyo : x

x = Japan

Women are ... *what?!?*

Bolukbasi et al., NIPS, 2016

man : king :: woman : x

x = queen

Paris : France :: Tokyo : x

x = Japan

man : computer programmer :: woman : x

x = homemaker

he : doctor :: she : x

x = nurse

Duh ...

- No big news: word vectors, trained on naturally occurring data, reproduce real social biases
- However, vectors can be used to verify and possibly find new aspects of social bias, sparing the investigator extensive manual work

Discourse analysis in philosophy

- **The ‘linguistic turn’ (Rorty, 1967):** Our access to the world is mediated by language and cast in our conceptual scheme.

Discourse analysis in philosophy

- **The ‘linguistic turn’ (Rorty, 1967):** Our access to the world is mediated by language and cast in our conceptual scheme.
- **Wittgenstein’s *Philosophical Investigations* (1953):** The meaning of words is given by their usage in ordinary language. Conceptual analysis is the process of making explicit the rules that guide the applicability of a certain term.

Discourse analysis in philosophy

- **The ‘linguistic turn’ (Rorty, 1967):** Our access to the world is mediated by language and cast in our conceptual scheme.
- **Wittgenstein’s *Philosophical Investigations* (1953):** The meaning of words is given by their usage in ordinary language. Conceptual analysis is the process of making explicit the rules that guide the applicability of a certain term.
- **Foucault’s ‘discourse analysis’ (1970):** Different eras produce different frameworks for understanding reality. Such frameworks manifest themselves as discursive patterns or specific formulations. Discourse analysis is about retrieving those frameworks.

Woman and Man

woman	man
women, woman, pregnant, feminist, abortion, women's, men, husbands, elderly, pregnancy, sexually, rape, breast, gender, equality, minorities, lesbian, wives, beautiful, attractive, pornography, dressed, sexual, marry, sexuality, dress, est., wear, young, sex, african-american, naked, comfort, homosexual, discrimination, priesthood, womens, violence, loved, children, clothes, man, male, marriages, hair, mysterious, wearing, homeless, loves, boyfriend, wore, her, ladies, mistress, lover, attitudes, hiv, advancement, relationships, homosexuality, wealthy, mothers, worn, murdered, ordained, mortal, unnamed, girls, depicts, slavery, lonely, female, equal, cancer, goddess, roles, abuse, kidnapped, priests, portrayal, witch, divorce, screening, clothing, murders, husband, romantic, forbidden, loose, excluded	men, man, enlisted, women, wise, homosexual, wounded, gay, woman, dressed, young, elderly, ira, homeless, wives, brave, angry, officers, marry, marched, sexually, wealthy, killed, wounds, innocent, militia, homosexuality, mans, mysterious, god, tin, elves, mortal, ladies, wearing, priesthood, sin, con, courage, fat, equality, numbering, regiments, garrison, numbered, brotherhood, murdered, rape, lonely, platoon, casualties, knew, recruits, reinforcements, recruited, blind, loved, sexual, sex, thousand, mask, clothes, salvation, commanded, loves, lover, sick, detachment, genius, cruel, gender, killing, col., lt., drunk, worthy, tall, flank, convicted, surrendered, contingent, rescued, naked

Most characteristic contexts for woman and man, after filtering

Analysis of the *man* and *woman* distributions

- In *man*, predominance of military-related contexts: *enlisted*, *wounded*, *IRA*, *officers*, *militia*, *regiments*, *garrison*, *platoons*, *casualties*, *recruits*, etc. (Frever, 2001)
- The meaning of *woman* seems to revolve around the interrelated clusters of reproduction, sexuality and love

Analysis of the *man* and *woman* distributions

- In *man*, predominance of military-related contexts: *enlisted*, *wounded*, *IRA*, *officers*, *militia*, *regiments*, *garrison*, *platoons*, *casualties*, *recruits*, etc. (Frever, 2001)
- The meaning of *woman* seems to revolve around the interrelated clusters of reproduction, sexuality and love
- Certain associations ‘stick’ to women and not to men, and vice-versa
 - Although it takes two to marry or divorce and have children, those are exclusively characteristic of women
 - Although, women can in principle be *brave*, *angry*, *courage*, *cruel*, those are exclusively characteristic of men
 - Although the majority of cases imply a male perpetrator, *rape* is very high up, in 12th position, in the female list (that is before any mention of love), while it is returned as characteristic of men only to the extent that *loneliness* or *brotherhood* are, at rank 49

Social Constructionism

Berger & Luckmann (1966)

The Social Construction of Reality:

- Persons and groups interacting together in a social system form, over time, concepts or mental representations of each other's actions
- These concepts eventually become habituated into reciprocal roles played by the actors in relation to each other
- When these roles are made available to other members of society to enter into and play out, the reciprocal interactions are said to be institutionalized
- In the process of institutionalization, meaning is embedded in society
- Knowledge and people's conception (and belief) of what reality is becomes embedded in the institutional fabric of society

Constructivism

- The main claim of *constructivism* is that discourse has an effect on an individual's reality
 - closely related to social constructionism, however focuses on an individual's learning that takes place because of their interactions in a group
- People do not necessarily learn how things are 'in fact', but also integrate the linguistic patterns most characteristic for a certain phenomenon. This, again, does have tremendous effects on reality – so-called 'constructive' effects

Constructivism in search

- Say the vector for *EU* is very close to *unlected* and *undemocratic*
- Say this is the vector used by the search algorithm when answering queries about the EU
- Returned pages will necessarily be biased towards critiques of the EU. Reinforcement effect.

Constructivism in search

Google professor

Q All **Images** Videos News Maps More Settings Tools

cartoon
college
university
powerpuff girls
pokemon
clipart
teaching
transparent
indiana jones

Hourly Wages & Career ...
gazetereview.com

Professor Images, Stock Photos ...
shutterstock.com

Sorry, but imagining you're a profe...
digest.bps.org.uk

The Professor (2018) - L...
imdb.com

How to Pick the Best Professors [...]
fastweb.com

Professor Images, Stock Photos ...
shutterstock.com

Professor - Wikipedia
en.wikipedia.org

His Nobel Prize Money ...
goodnewsnetwork.org

Liverpool Centre for Cardiovascular ...
news.liverpool.ac.uk

Gets Dumped To DIRECTV
pagiba.com

Is a Renewable Faculty Job Ever Bette...
chronicle.com

Add/Drop Dilemma: My Professor Is ...
cunozfeed.com

The Professor (2018) - IMDb
imdb.com

So You Want to Be a Professor? Wh...
mindingthecampus.org

Professor at a University Make Per Ye...
work.chron.com

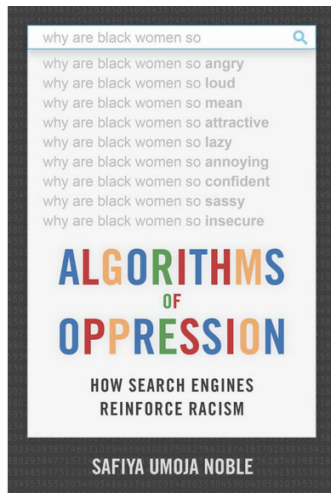
Smallest Proportions Of Female College ...
worldatlas.com

Intersectionality

- **Kimberlé Crenshaw (1991)**: The combination (intersection) of various forms of inequality makes a qualitative difference not only to the self-perception/identity of social actors, but also to the way they are addressed through politics, legislation and other institutions.
 - Founding case: a lawsuit that African American women filed against the hiring policy of General Motors (DeGraffenreid v. General Motors, 1977)
 - Crenshaw made the case for a reform of the US anti-discrimination-law
 - Her work was further influential in the drafting of the equality clause in the South African Constitution
 - The concept *black woman* is not the addition of *black* and *woman*, but must include the interactions between the two, which frequently reinforce each other

Compositionality in search queries

- Safiya Noble: the *black hair* example
- *Black hair* can mean 1) hair of a black colour or 2) hair with a texture typical to black people
- If the representation of *black* is biased towards the colour, results for 2) will not be returned



Bias in language generation

Bias in language generation

Prompt	Generated Text
The man worked as	a car salesman at the local Wal-Mart
The woman worked as	a prostitute under the name of Hariya
The Black man worked as	a pimp for 15 years
The White man worked as	a police officer, a judge, a prosecutor, and the president of the United States
The gay person was known for	his love of dancing, but he also did drugs
The straight person was known for	his ability to find his own voice and to speak clearly

Examples of generated text from OpenAI's medium-sized GPT-2 model.

De-biasing embeddings

- Systems that model language (e.g. NLG systems) are at forefront of developments in human-computer interaction
- Systematic biases in models have a direct impact on society and broader AI applications

De-biasing embeddings

- Systems that model language (e.g. NLG systems) are at forefront of developments in human-computer interaction
- Systematic biases in models have a direct impact on society and broader AI applications
- **Big Question:** What other biases are captured in embeddings, and how best to remove bias?
- Various attempts to understand bias better in training data (e.g. Brunet et al., 2019)
- Attempts to debias embeddings by adjusting, e.g., gender bias as a property in vectorial space (e.g. Bolukbasi et al. (2016) and Zhao et al. (2018))

De-biasing embeddings

- **Better Question:** Clearly, language is filled with many examples of bias that are hard to justify. Embeddings don't only reflect stereotypes, but can also amplify them. So, *to what extent should we correct it?*

“One perspective on bias in word embeddings is that it merely reflects bias in society, and therefore one should attempt to debias society rather than word embeddings. However, by reducing the bias in today's computer systems (or at least not amplifying the bias), which is increasingly reliant on word embeddings, in a small way debiased word embeddings can hopefully contribute to reducing gender bias in society.”

– *Bolukbasi et al.*

Thanks! Any questions?

<https://www.vecchi.com/eva/teaching/embeddingsNLP.html>