Natural Language Processing with Deep Learning Class

Paweł Budzianowski, 2022

An over a decade of fascinating progress

- NLP <-> General ML
- ASR, TTS, Computer Vision, RL

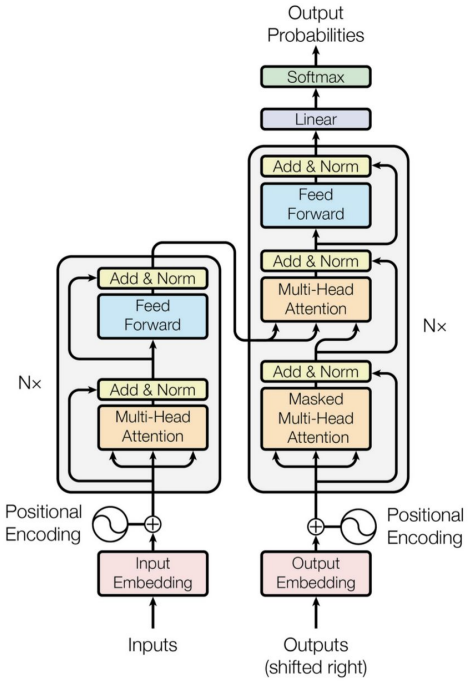


Figure 1: The Transformer - model architecture.

Figure 1: From 'Attention Is All You Need' by Vaswani et al.

MIM UW & IIMK UJ

1. Jagiellonian University, Faculty of Mathematics and Information Technologies

2. University of Warsaw, Faculty of Mathematics, Informatics, and Mechanics,

Lectures plan

1. Introduction to NLP and Meaning (28.02.2022)
2. Word vectors (07.03)
3. Language modeling (14.03)
4. RNNs (21.03)
5. Machine Translation, Attention (28.03)
6. Transformers (04.04)
7. Pre-training (Piotr Rybak, Allegro) (11.04)
8. Conversational AI (25.04)
9. Vision + language (Mateusz Malinowski, Deepmind) (09.05)
10. Vision + language (Mateusz Malinowski, Deepmind) (16.05)
11. Question-Answering (23.05)
12. Multilingual NLP (Ivan Vulić, Cambridge University) (30.05)
13. Data for NLP / Model Analysis (06.06)
14. New frontiers (TBD) (13.06)

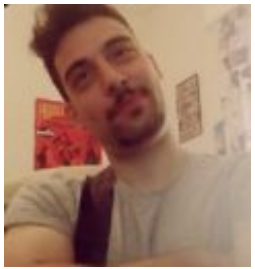
Coursework plan

Practicals (Laboratoria) - 50% of the final grade

Group projects - 50% of the final grade

Submissions on Moodle (UW) / Pegaz (UJ)






Coursework plan

Practical 1 - 5% (due 15.03)

Practical 2 - 10% (due 21.03)

Practical 3 - 15%

Practical 4 - 15%

Practical 5 - 5%

Final proposal - 15%

Final report - 35%

Due date policy

1 day late: -10%

2 days late: -20%

3 days late: -30%

4 days late: -50%

5 days late: -100%

Attendance policy

Attendance will be checked at each lecture and practicals with possible perks.

Course slack

https://join.slack.com/t/dnlpclass/shared_invite/zt-141v925w5-FcX3MZiYPw5VNikA9BKdCQ

Questionnaire

https://docs.google.com/forms/d/e/1FAIpQLSd9A1U06NVL436qvHr9WTI4YtgAX-W9Zfm-j9kCk041Zm3j7w/viewform?usp=sf_link

Language, thoughts, meaning and word vectors

Deep Natural Language Processing, 2022

Paweł Budzianowski

Natural Language

- Language is quite a recent ‘invention’ of evolution
- Human knowledge is mainly translated through language.
- Writing translated mass knowledge.
- It’s only 5000 years old.
- Bible and others as ‘MetaBrain’ transferred over generations.

- Language is a really fast communicator since we compressed a lot assuming the second side understands a lot of things for themselves.

Language acquisition

Genetics inheritance of language knowledge - strong hypothesis.

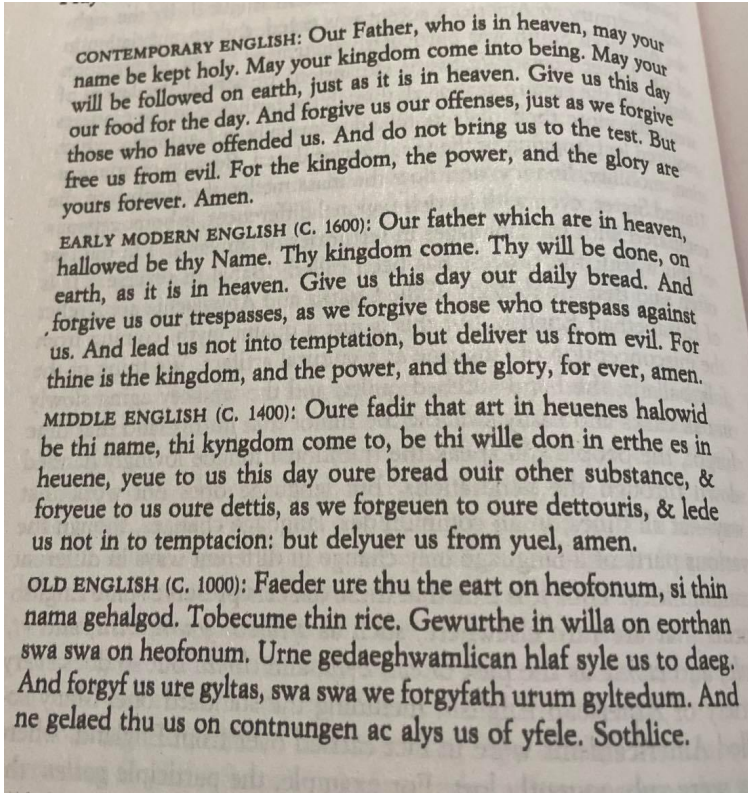
Language instinct - weak hypothesis.

Yet patterns are visible across thousand of them!

- subject-verb-predicate (SVP)

- predicate-verb-subject (PVS)

Language evolves!



CONTEMPORARY ENGLISH: Our Father, who is in heaven, may your name be kept holy. May your kingdom come into being. May your will be followed on earth, just as it is in heaven. Give us this day our food for the day. And forgive us our offenses, just as we forgive those who have offended us. And do not bring us to the test. But free us from evil. For the kingdom, the power, and the glory are yours forever. Amen.

EARLY MODERN ENGLISH (C. 1600): Our father which are in heaven, hallowed be thy Name. Thy kingdom come. Thy will be done, on earth, as it is in heaven. Give us this day our daily bread. And forgive us our trespasses, as we forgive those who trespass against us. And lead us not into temptation, but deliver us from evil. For thine is the kingdom, and the power, and the glory, for ever, amen.

MIDDLE ENGLISH (C. 1400): Oure fadir that art in heuenes halowid be thi name, thi kyngdom come to, be thi wille don in erthe es in heuene, yeue to us this day oure bread ouir other substance, & foryeue to us oure dettis, as we forgeuen to oure dettouris, & lede us not in to temptacion: but delyuer us from yuel, amen.

OLD ENGLISH (C. 1000): Faeder ure thu the eart on heofonum, si thin nama gehalgod. Tobecume thin rice. Gewurthe in willa on eorthan swa swa on heofonum. Urne gedaeghwamlican hlaf syle us to daeg. And forgyf us ure gyltas, swa swa we forgyfath urum gyltedum. And ne gelaed thu us on contnungen ac alys us of yfele. Sothlice.

Civilization vs language

Civilization growth and its advancement does not go hand in hand with the complexity of the language.

90% of languages are doomed to extinct.

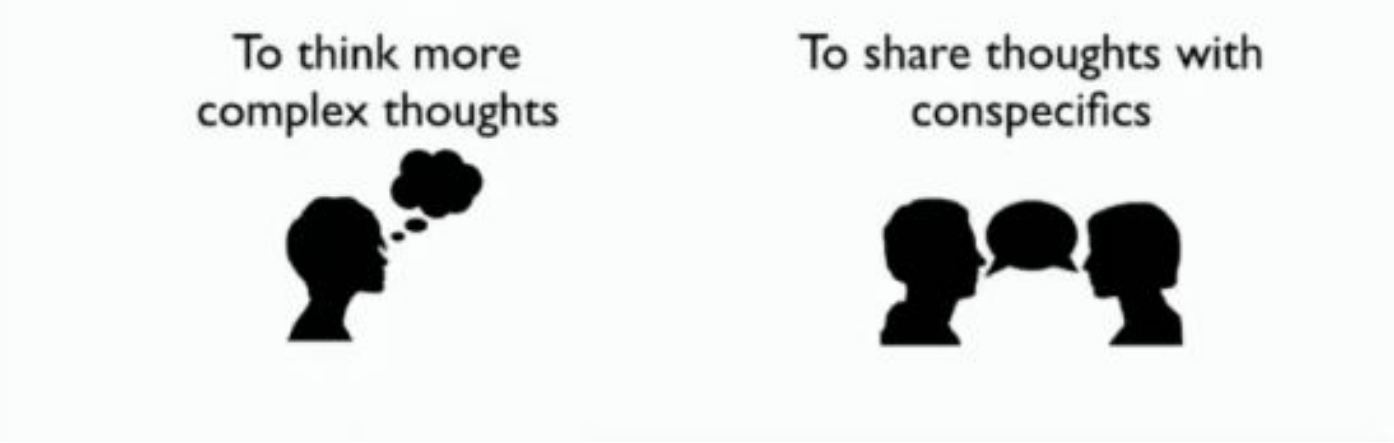
Sapiro-Whorf Hypothesis

1) Strong version - the language we speak determines the way we think and view the real world

2) Weak version - the language does influence to some extent the way we think and view the real world, however, does not fully determine or constraint it.

What is the computational goal of language?

Why did language emerge in our species?



What is the computational goal of language

Why did language emerge in our species?

Noam Chomsky

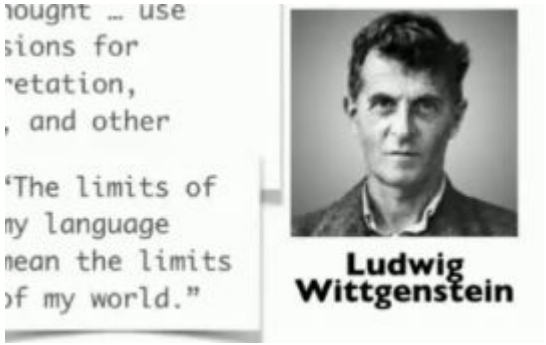
"Almost all of your use of language is internal. Virtually all of the use of language has nothing to do with communication. The idea that language has evolved as a system of communication, or designed for communication, makes no sense."

(presentation at Arizona State University, March 2019)

[https://www.statepress.com/article/2019/03/spcommunity-noam-chomsky-speaks-as-keynote-at-asu-philosophy-club-conference]

What is the computational goal of language?

Why did language emerge in our species?



The language network [Fedorenko et al., 2010]

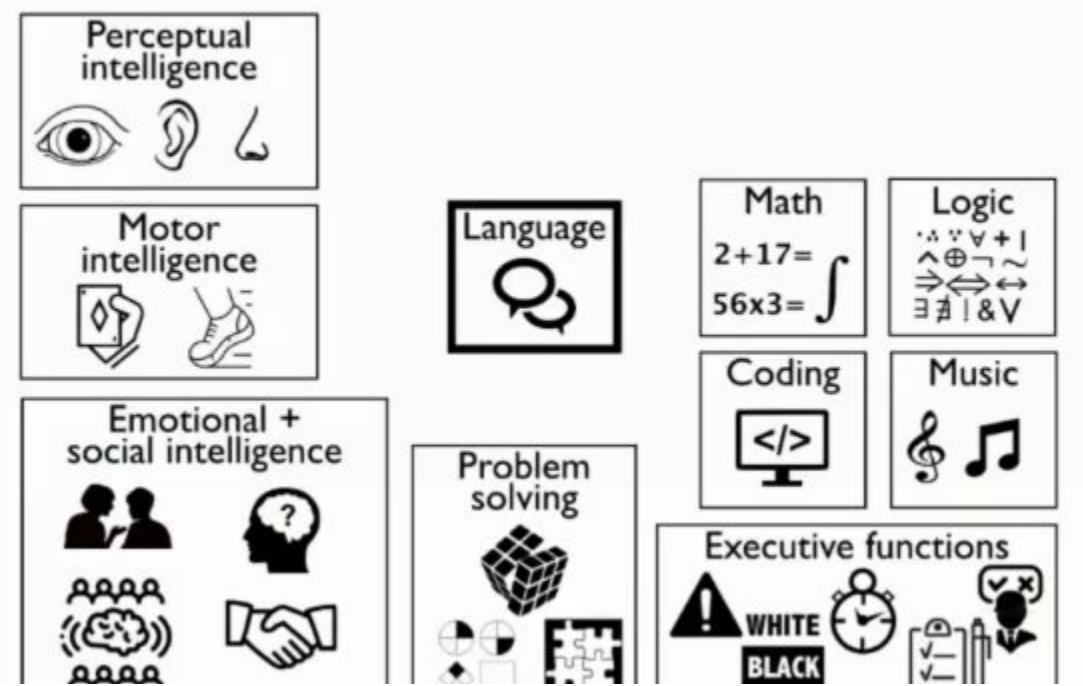
Key functional properties

The language network

Key functional properties:

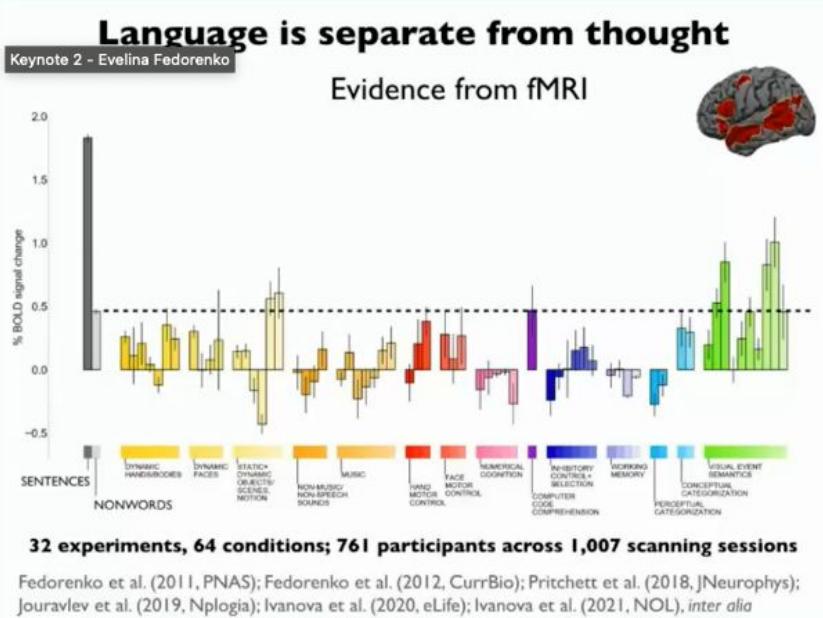
- They respond robustly during comprehension.
- They respond robustly during production.
- Are similar across diverse.
- Form a functionally integrated network.

Language vs other cognitive abilities



Language vs other cognitive abilities

The language system does not support any other cognitive abilities!



Properties of language

- Language is not suitable for complex thought.

- Language is well-suited for communication.

Language does not rely on abstract syntax

- No syntactic hubs: syntactic processing is distributed across the language network.
- No syntax selectivity: every syntax-responsive cell population or brain area is robustly sensitive to word meanings.

Language is well-suited for communication

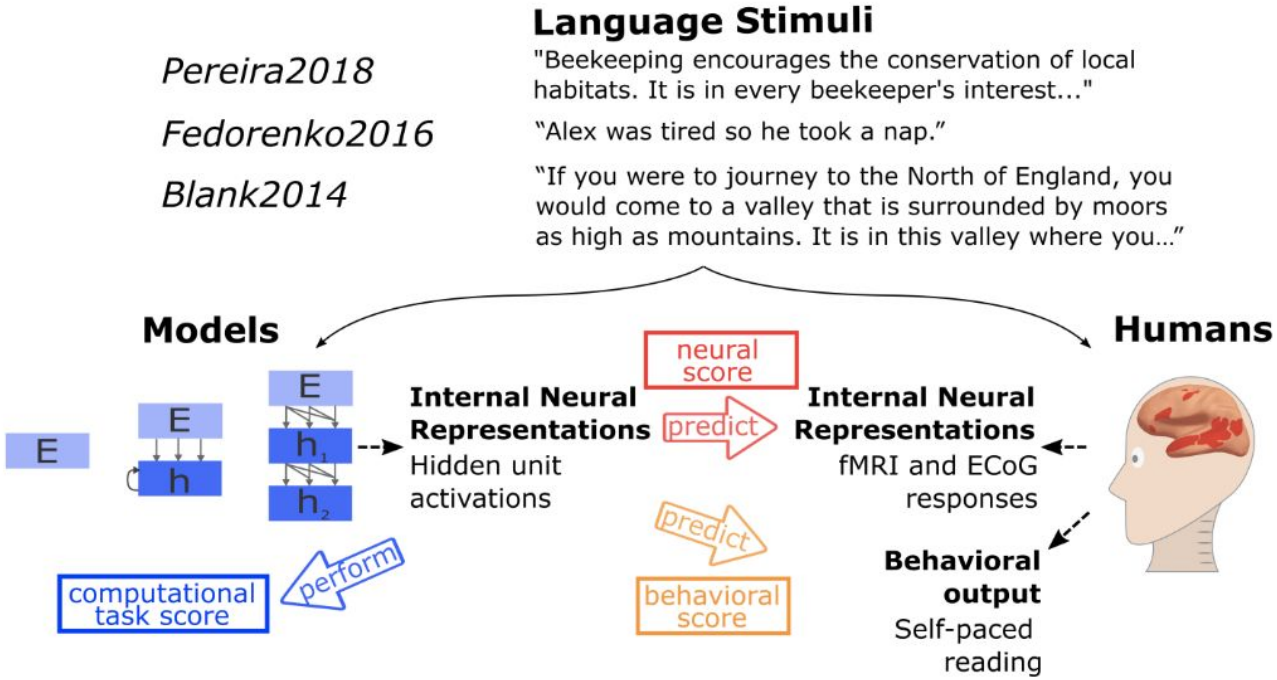
- Natural languages are **efficient** for information transfer.
- Linguistic code is:
  - short
  - contextually disambiguated
  - redundant
- Language processing is fundamentally **predictive**.
- Training on **word prediction** leads to representations that approximate meaning well enough to perform well on diverse language tasks.

Two linguists worlds

Chomsky’s theses in Norvig’s eyes:

1. Statistical language models have had engineering success, but that is irrelevant to science.
2. Accurately modeling linguistic facts is just butterfly collecting; what matters in science (and specifically linguistics) is the underlying principles.
3. Statistical models are incomprehensible; they provide no insight.
4. Statistical models may provide an accurate simulation of some phenomena, but the simulation is done completely the wrong way; people don't decide what the third word of a sentence should be by consulting a probability table keyed on the previous two words, rather they map from an internal semantic form to a syntactic tree-structure, which is then linearized into words. This is done without any probability or statistics.
5. Statistical models have been proven incapable of learning language; therefore language must be innate, so why are these statistical modelers wasting their time on the wrong enterprise?

Language processing is fundamentally predictive



Meaning and Language

How do we actually represent the meaning of a word?

What is meaning? What is life?

- Definition of a meaning:
  - the idea that is represented by a word, phrase
  - the idea that a person wants to express by words or signs
  - the idea that is represented through art

- Linguist puts some sort of equivalence between:
  idea and **symbol** (or thing)

The meaning of concept and category [Brown, 1958]

A *concept* is the mental representation of a *category* (bird) where instances of objects or events are grouped together.

Category members do not have equal status: some are more *prototypical* than others (penguins vs robins in *birds* category) and boundaries of peripheral members are fuzzy.

They also display a degree of variation across cultures.

Categories form a hierarchy, from the general to specific. The basic level is not universal though.

Meaning through human definitions

Common solution: use **Wordnet**, a thesaurus containing lists of synonym lists and hypernyms (“is a” relationships)

Meaning through human definitions

Noun

- S: (n) **cool** (the quality of being at a refreshingly low temperature) *"the cool of early morning"*
- S: (n) aplomb, assuredness, **cool**, poise, sang-froid (great coolness and composure under strain) *"keep your cool"*

Verb

- S: (v) **cool**, chill, cool down (make cool or cooler) *"Chill the food"*
- S: (v) **cool**, chill, cool down (loose heat) *"The air cooled considerably after the thunderstorm"*
- S: (v) **cool**, cool off, cool down (lose intensity) *"His enthusiasm cooled considerably"*

Adjective

- S: (adj) **cool** (neither warm nor very cold; giving relief from heat) *"a cool autumn day"; "a cool room"; "cool summer dresses"; "cool drinks"; "a cool breeze"*
- S: (adj) **cool**, coolheaded, nerveless (marked by calm self-control (especially in trying circumstances); unemotional) *"play it cool"; "keep cool"; "stayed coolheaded in the crisis"; "the most nerveless winner in the history of the tournament"*
- S: (adj) **cool** (inducing the impression of coolness; used especially of greens and blues and violets when referring to color) *"cool greens and blues and violets"; "the cool sound of rushing water"*
- S: (adj) **cool** (psychologically cool and unenthusiastic; unfriendly or unresponsive or showing dislike) *"relations were cool and polite"; "a cool reception"; "cool to the idea of higher taxes"*

Problems with resources like WordNet

- Great resources but missing nuances:
  popular is not a synonym to cool

- Missing new meanings of words cause it’s impossible to keep it up to date (think of dzban, preppers, mem)

- Highly subjective.

- Requires human labour to create and adapt.

- Can’t compute accurate word similarity.

Representing words as discrete numbers

- In traditional NLP, we regard words as discrete symbols:
  restaurant, bistro, coffee shop

- We called this localist representation - for any concept there is a one place with a definition of it (a categorical variable)

- Words can be represented as one-hot vectors:

  bistro = [0, 0, 1,0,0,0,0,0,0,0,0,0]
  cafe = [0, 1, 0,0,0,0,0,0,0,0,0,0]

- Vector dimension == number of words in the vocabulary. We end up with very big vectors

Problem with words as discrete numbers

- Looking for similarity between similar concepts like bistro and cafe is hard

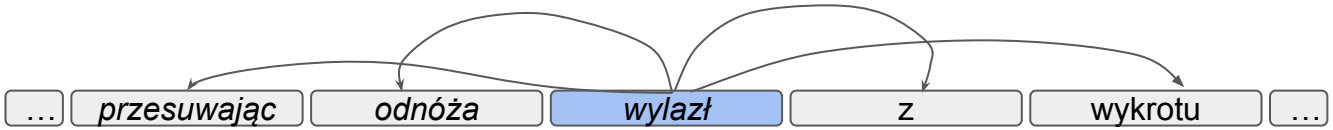
  bistro = [0, 0, 1,0,0,0,0,0,0,0,0,0]
  cafe = [0, 1, 0,0,0,0,0,0,0,0,0,0]

- Vectors are orthogonal and thus there is no natural notion of similarity.

- We could use WordNet similarity table - however it’s incomplete and it’s hard to operate with.

Distributional semantics

- **A word's meaning is given by the words that frequently appear close-by**
- Firth - "***You shall know a word by the company it keeps***"
- Wittgenstein - ***"In most cases, the meaning of a word is its use"***
- One of the most successful ideas of modern statistical NLP
- When a word *w* appears in a text, its context is the set of words that appear nearby (within a fixed size window)
- Use the many contexts of *w* to build up a representation of *w*



Idea: Word vectors

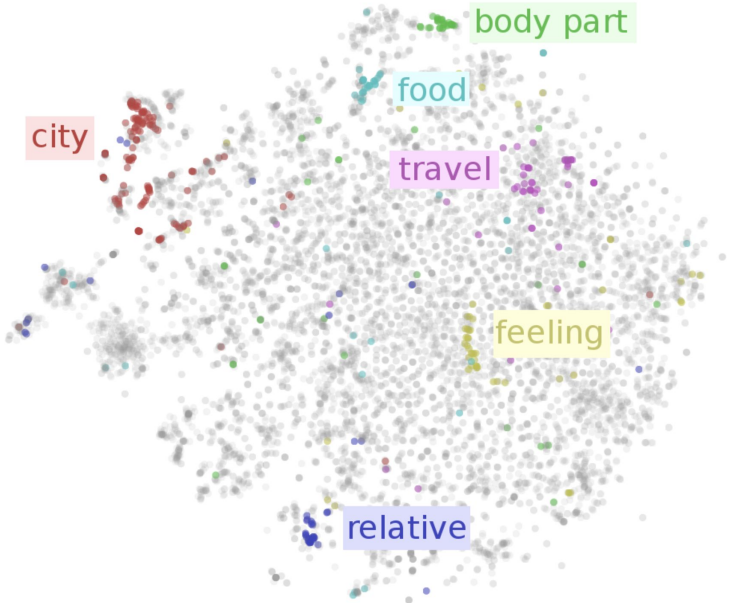
We will build a dense vector for each word, chosen so that it is similar to vectors of words that appear in similar contexts.

For example let’s put our word in n-dimensional space?

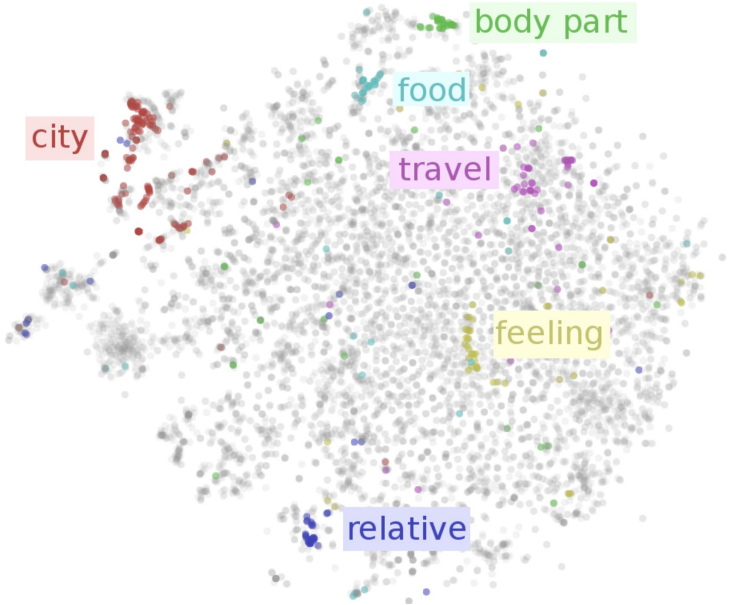
motel = (2, 3,45, 6, 7.323, -12.32, 0.2123, …,0.2323,0.123,0.312)

Word vectors are interchangeably called **word embeddings** or word representations. They are a **distributed** representation.

Word meaning as a neural word vector - embedded in space



Word meaning as a neural word vector - embedded in space



This was sort of Imagenet moment when playing with this worked out.

Efficient Estimation of Word Representations in Vector Space

Tomas Mikolov
Google Inc., Mountain View, CA
tmikolov@google.com

Kai Chen
Google Inc., Mountain View, CA
kaichen@google.com

Greg Corrado
Google Inc., Mountain View, CA
gcorrado@google.com

Jeffrey Dean
Google Inc., Mountain View, CA
jeff@google.com

Word2vec (Skip-gram) [Mikolov et al., 2013]

Word2vec is a framework for learning word vectors.

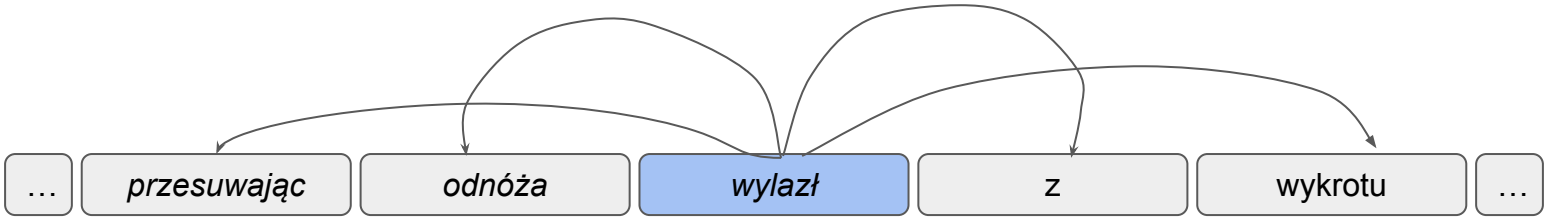
Idea:

- We have a large corpus of text,
- Every word in a fixed vocabulary is represented by a **vector,**
- Go through each position *t* in the text, which has a center word *c* and context, (“outside”) words *o,*
- Use the **similarity of the word vectors** for *c* and *o* to **calculate the probability** of *o* given *c* (or vice versa),
- **Keep adjusting the word vectors** to maximize this probability.

Main idea of the word2vec

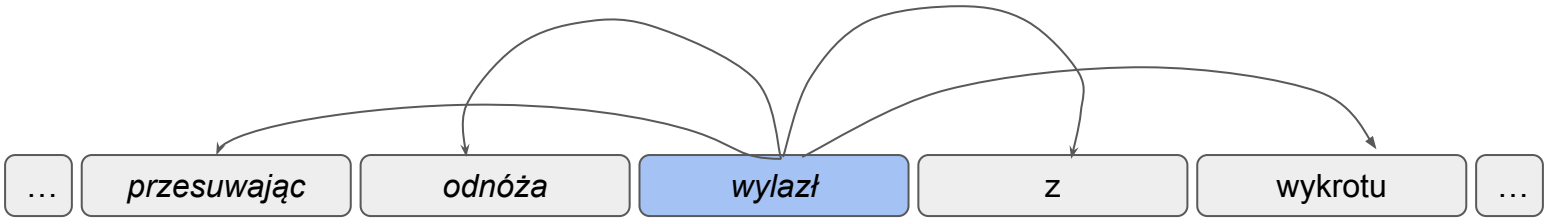
Iterate through each word of the whole corpus.

Predict the words around the center word.



Word2vec: Overview

$$P(w_{t+k}|w_t)$$



$$P(w_{t-1}|w_t) \qquad P(w_{t+1}|w_t)$$

Word2vec: objective function

For each position *t = 1,..., T*, predict context words within a window of fixed size m given center word *w_j.*

Likelihood: $$L(\theta) = \prod_{t=1}^{T} \prod_{-m \leq j \leq m, j \neq 0} P(w_{t+j} | w_t; \theta)$$

The objective function is the average negative log likelihood.

$$J(\theta) = -\frac{1}{T} \sum_{t=1}^{T} \sum_{-m \leq j \leq m, j \neq 0} \log(P(w_{t+j} | w_t; \theta))$$

Minimizing objective function ⇔ Maximizing predictive accuracy.

Word2vec: objective function

We are about to minimize the objective:

$$J(\theta) = -\frac{1}{T}\sum_{t=1}^{T}\sum_{-m\leq j\leq m, j\neq 0}\log(P(w_{t+j}|w_t;\theta))$$

Question: How to calculate $P(w_{t+k}|w_t)$?

Answer: we will use two vectors per word w:

- *v_w* when *w* is a center word
- *u_w* when *w* is a context word

Then for a center word *c* and a context word *o*:

$$P(O=o|C=c) = \frac{\exp(\mathbf{u}_o^T\mathbf{v}_c)}{\sum_{w\in Vocab}\exp(\mathbf{u}_w^T\mathbf{v}_c)}.$$

Word2vec: probability function

$$P(O = o|C = c) = \frac{\exp(\mathbf{u}_o^T \mathbf{v}_c)}{\sum_{w \in Vocab} \exp(\mathbf{u}_w^T \mathbf{v}_c)}.$$

- The exponential makes all positive.
- The dot product compares the similarity of words *o* and *c*. The larger the dot-product the larger the similarity between the words.
- The denominator normalizes over entire vocabulary to give probability distribution.

- This loss function is an example of the *softmax* function that you will see all over the place throughout your life in ML universe.
- The exponential function pushes the bigger number even more apart creating the max approach.

Training?

To train a model, we adjust parameters to minimize a loss.
In our model all parameters \theta are our word vectors.

In our case with *d*-dimensional vectors and *V*-many words we get

$$\theta = [v_{abecadło}, v_{arkany}, \ldots, v_{zeta}, u_{abecadło}, u_{arkany}, \ldots, u_{zeta}] \in \mathbb{R}^{2dV}$$

Remember: every word has two vectors.
We optimize these parameters by walking down the gradient.

Training?

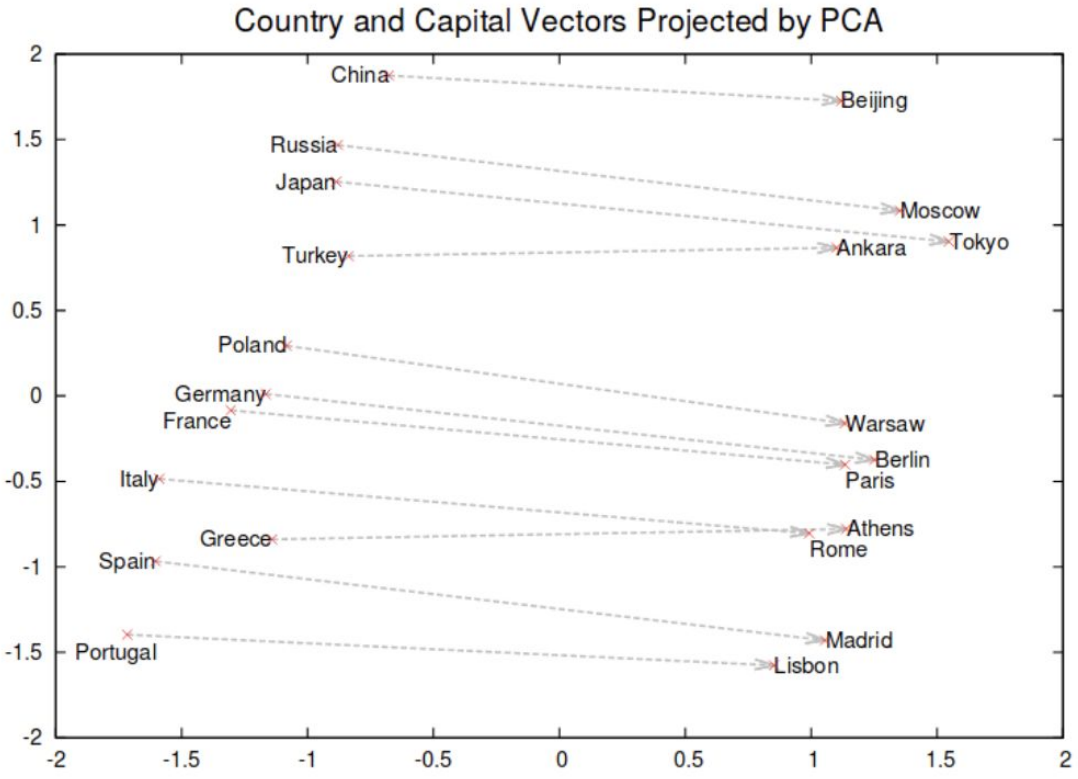
$$\min J(\theta) = -\frac{1}{T}\sum_{t=1}^{T}\sum_{-m\leq j\leq m, j\neq 0}\log(P(w'_{t+j}|w_t;\theta))$$

$$P(o|c) = \frac{\exp(\mathbf{u}_o^T\mathbf{v}_c)}{\sum_{w\in Vocab}\exp(\mathbf{u}_w^T\mathbf{v}_c)}.$$

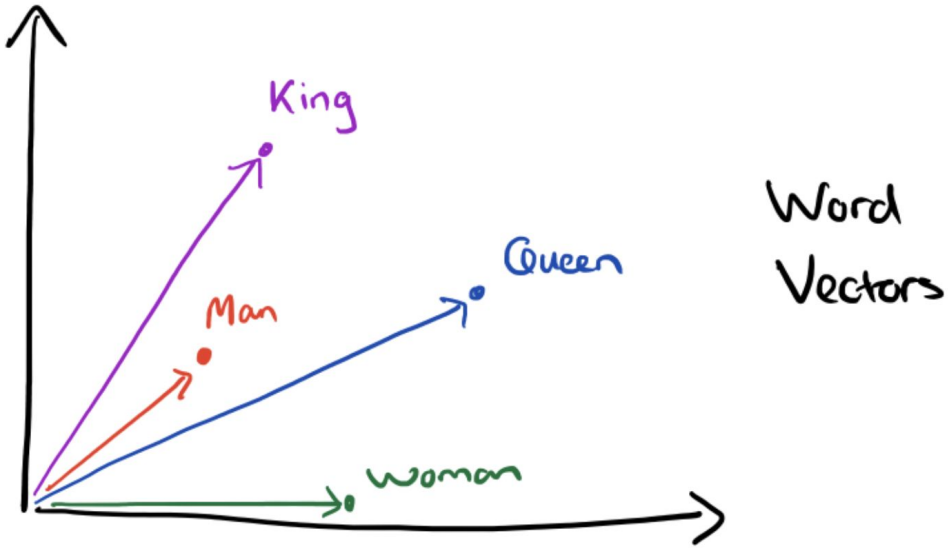
Finding a meaning?

$$\frac{\partial}{\partial \mathbf{v}_c} P(O = o|C = c) = u_o - \sum_{x \in Vocab} p(x|c) u_x$$

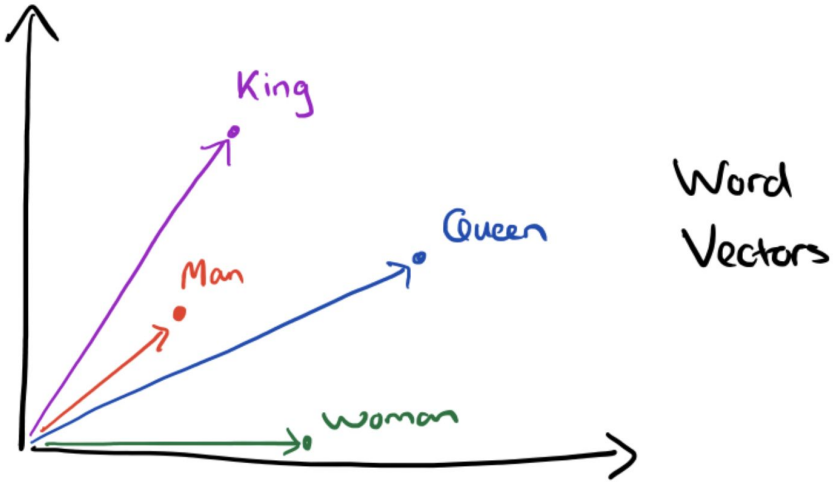
TSNE/PCA plots



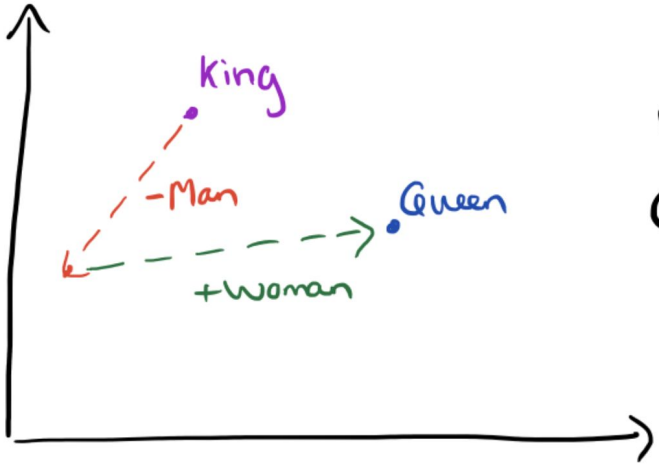
Unexpected miracles (Colyer 2022)



Unexpected miracles (Colyer 2022)



Word Vectors



Vector Composition

Where is the catch?

$$P(O = o|C = c) = \frac{\exp(\mathbf{u}_o^T \mathbf{v}_c)}{\sum_{w \in Vocab} \exp(\mathbf{u}_w^T \mathbf{v}_c)}$$

Literature:

1. Mikolov et al., 2012 - https://arxiv.org/pdf/1310.4546.pdf
2. Norvig, 2001 - https://norvig.com/chomsky.html
3. Fedorenko et al., 2010 - https://journals.physiology.org/doi/full/10.1152/jn.00032.2010