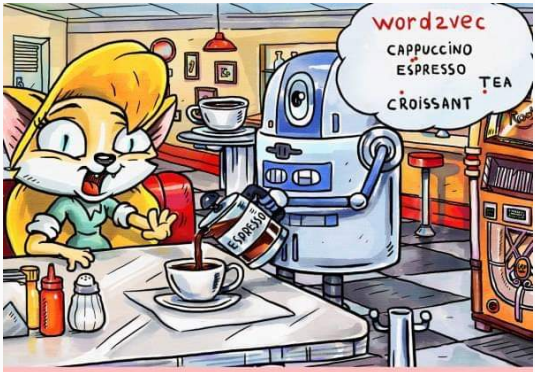
KOINet – New Tools for old Problems

Large Pre-trained Language Models – can social sciences also benefit from recent developments in NLP?

Matthias Aßenmacher, Christian Heumann (LMU München)

09. Januar 2021

Representing words . . .



- Espresso? But I ordered a cappuccino!
- Don't worry, the cosine distance between them is so small that they are almost the same thing.

That's not how it's intended to work ..

Recent advances in NLP

We encounter more and more NLP applications in everyday life:

- Chatbots are on the rise
- Alexa or Siri have become standard tools
- GoogleTranslate or DeepL are commonly used

The world's largest Tech companies are investing heavily:

- fb ai research, google ai, microsoft research have own NLP groups
- Leading researchers like Geoffrey Hinton (Google) or Yann LeCun (Facebook) start working for the industry

How to represent words? – The distributional hypothesis

Zellig S. Harris (1954):

▸ *Distributional Structure*

J.R. Firth (1957):

"*You shall know a word by the company it keeps.*"

Learn something about the meaning of *football* by studying which context it appears in:

.. the score of the	*football*	game was 3:0 ..
.. he shot the	*football*	directly at the goalkeeper ..
.. last night, I was watching	*football*	on tv ..

One-hot vs. context-based encoding

One-hot encoding:

$$\textit{football} = [0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]$$
$$\textit{basketball} = [0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]$$

Two major problems:

- $\textit{similarity}(\textit{football}, \textit{basketball}) = ?$
  The vectors are orthogonal to each other, so $sim(w_i, w_j) = 0 \; \forall \; i, j$
- The dimensionality of these vectors?

One-hot vs. context-based encoding

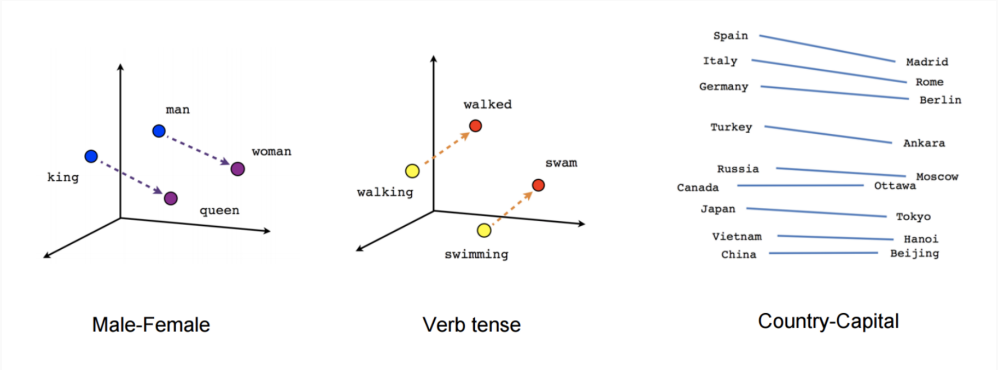
Context-based encoding:

$$\textit{football} = [0, 3, 0, 0, 1, 0, 0, 0, 0, 0, 2, 0, 2, 1, 4]$$
$$\textit{basketball} = [0, 0, 0, 2, 0, 0, 0, 0, 0, 2, 1, 0, 3, 3, 2]$$

Two major problems:

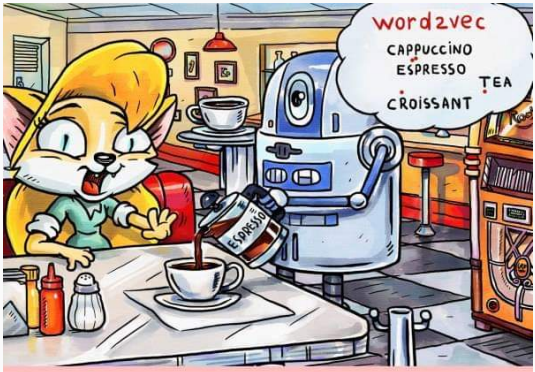
- $\textit{similarity}(\textit{football}, \textit{basketball}) = ?$
- The dimensionality of these vectors?

The breakthrough: *Word embeddings*



Source: towardsdatascience

Maybe now it's funnier?



- Espresso? But I ordered a cappuccino!
- Don't worry, the cosine distance between them is so small that they are almost the same thing.

Context-free Embeddings

1st Generation of neural embeddings are *"context-free"*

- Models learn *one single* embedding for each word
- Why could this possibly be problematic?
  - "She was sitting on a *bank* in the park."
  - "He transferred the money to her *bank* account."
- Would be nice to have different embeddings for these two occurrences

How to become "contextual"?

- Model makes further use of the context a word appears in
- Embeddings depend on the context around a word
- Distinguish between:
  - Unidirectional
  - Bidirectional

Advancing Word Embeddings

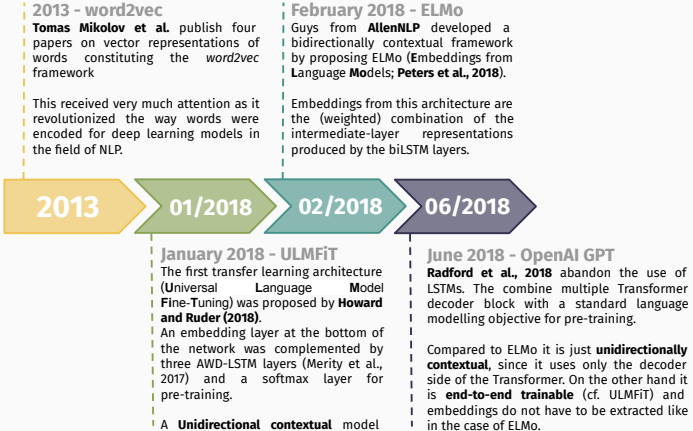
2013 - word2vec

Tomas Mikolov et al. publish four papers on vector representations of words constituting the *word2vec* framework

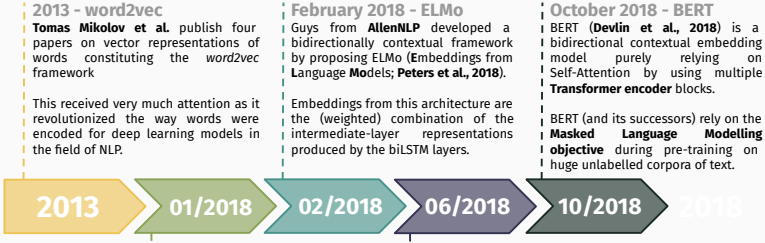
This received very much attention as it revolutionized the way words were encoded for deep learning models in the field of NLP.

2013

Advancing Word Embeddings



Advancing Word Embeddings



For now only in the US, but:

2019-10-25

0

AI / INDUSTRY

Milestone: BERT Boosts Google Search

In what the company calls “the biggest leap forward in the past five years, and one of the biggest leaps forward in the history of Search,” Google today announced that it has leveraged its pretrained language model BERT to dramatically improve the understanding of search queries.

Source: Synced

Corresponding blog post by Google:

https://www.blog.google/products/search/search-language-understanding-bert/

Now what does this all mean?

Key facts:

- *Super large* models applicable to a wide range of tasks
- Compute and data hungry, but:
  - pre-trained versions available
  - for a wide range of languages

Two exemplary use cases for applications in Social Science:

- Fake News Detection ▸ Guderlei & Aßenmacher (2020)
- Automated coding of Open-ended survey responses ▸ Meidinger & Aßenmacher (2021)
  (To appear at: *13th International Conference on Agents and Artificial Intelligence 2021*)

Task description: Stance detection of article body towards headline

Headline: Hundreds of Palestinians flee floods in Gaza as Israel opens dams	
Agree (AGR)	Hundreds of Palestinians were evacuated from their homes Sunday morning after Israeli authorities opened a number of dams near the border, flooding the Gaza Valley in the wake of a recent severe winter storm. [...]
Disagree (DSG)	Israel has rejected allegations by government officials in the Gaza strip that authorities were responsible for released storm waters flooding parts of the besieged area. "The claim is entirely false, and [...]" [...]
Discuss (DSC)	Palestinian officials say hundreds of Gazans were forced to evacuate after Israel opened the gates of several dams on the border with the Gaza Strip, and flooded at least 80 households. Israel has denied the claim as “entirely false”. [...]
Unrelated (UNR)	A Catholic priest from Massachusetts had been dead for 48 minutes before he was miraculously resuscitated. However, it is his description about God that is bound to spark a hot debate about the almighty. [...]

Results:

	BERT		RoBERTa		DistilBERT		ALBERT		XLNet	
Metric	FNC-1	+ ARC	FNC-1	+ ARC	FNC-1	+ ARC	FNC-1	+ ARC	FNC-1	+ ARC
F_1-m	**70.18**	**72.20**	**78.18**	**78.19**	**72.11**	**73.59**	**59.80**	**65.01**	**75.00**	**75.57**
F_1-AGR	60.31	63.48	70.69	70.57	61.95	65.29	53.19	53.97	68.00	68.57
F_1-DSG	41.76	48.28	56.15	58.92	45.09	50.46	13.21	34.07	49.47	53.69
F_1-DSC	80.36	78.82	86.78	84.16	82.83	80.22	76.16	75.18	83.73	81.43
F_1-UNR	98.28	98.22	99.10	99.09	98.58	98.38	96.65	96.83	98.80	98.60

Table 4: Model performances with respect to class-wise F_1 as well as F_1-m in comparison for FNC-1 and FNC-1 ARC. For better readability we indicate the columns for FNC-1 ARC just with "+ ARC".

Automated coding ▸ Meidinger & Aßenmacher (2021)

Setup:

- **Task:** Assign survey responses to pre-defined classes a.k.a. "*Codes*"
- ANES 2008 survey on political opinion and voting behavior
- 10 different data sets for different (groups) of questions with 9 – 72 Codes
- Data set size varies between around 200 up to over 8000 observations

→ Partly *extremely* challenging data sets

Results:

- Mixed performances (especially not good for small data sets)
- But tentatively promising results for some data sets
- **Main contribution:** Reproducible preparation of a new benchmark data set

Can social sciences also benefit from recent developments in NLP?

Happy to discuss this with you! 😊