Deep Learning for Natural Language Processing

Perspectives on word embeddings

UNIVERSITY OF GOTHENBURG

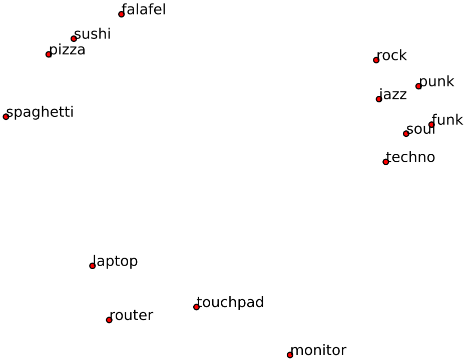
CHALMERS

WASP WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

Richard Johansson

richard.johansson@gu.se

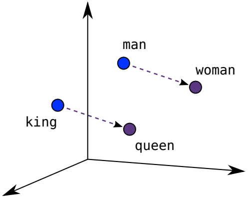
- word embedding models learn a “meaning representation” automatically from raw data



- that sounds really nice, doesn’t it?

bias in pre-trained embeddings

- word embeddings store statistical knowledge about the words
- Bolukbasi et al. (2016) point out that embeddings reproduce gender (and other) stereotypes



Extreme *she*	Extreme *he*
1. homemaker	1. maestro
2. nurse	2. skipper
3. receptionist	3. protege
4. librarian	4. philosopher
5. socialite	5. captain
6. hairdresser	6. architect
7. nanny	7. financier
8. bookkeeper	8. warrior
9. stylist	9. broadcaster
10. housekeeper	10. magician

	Gender stereotype *she*-*he* analogies	
sewing-carpentry	registered nurse-physician	housewife-shopkeeper
nurse-surgeon	interior designer-architect	softball-baseball
blond-burly	feminism-conservatism	cosmetics-pharmaceuticals
giggle-chuckle	vocalist-guitarist	petite-lanky
sassy-snappy	diva-superstar	charming-affable
volleyball-football	cupcakes-pizzas	lovely-brilliant
	Gender appropriate *she*-*he* analogies	
queen-king	sister-brother	mother-father
waitress-waiter	ovarian cancer-prostate cancer	convent-monastery

does this matter?

ENGLISH	FINNISH
He's my brother.	Hän on veljeni.
She's my sister.	Hän on siskoni.
She's a brain surgeon.	Hän on aivokirurgi.
She's a computer programmer.	Hän on tietokoneohjelmoija.
He's a nanny.	Hän on lastenhoitaja.
He's a nurse.	Hän on sairaanhoitaja.
She's a car dealer.	Hän on autokauppias.
She's a philosopher.	Hän on filosofi.
He's a librarian.	Hän on kirjastonhoitaja.

FINNISH	ENGLISH
Hän on veljeni.	He's my brother.
Hän on siskoni.	She's my sister.
Hän on aivokirurgi.	He's a brain surgeon.
Hän on tietokoneohjelmoija.	He's a computer programmer.
Hän on lastenhoitaja.	She's a nanny.
Hän on sairaanhoitaja.	He is a nurse.
Hän on autokauppias.	He's a car dealer.
Hän on filosofi.	He is a philosopher.
Hän on kirjastonhoitaja.	She's a librarian.

stereotypes in NLP models (1)

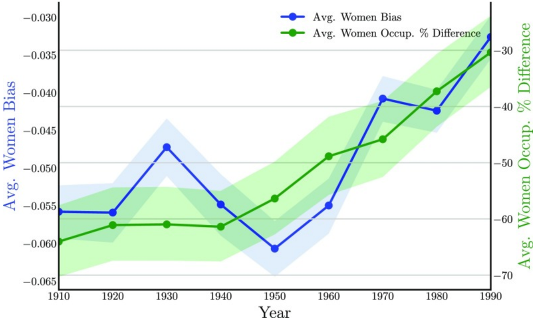
```
In [18]: text_to_sentiment("My name is Emily")
Out[18]: 2.2286179364745311

In [19]: text_to_sentiment("My name is Heather")
Out[19]: 1.3976291151079159

In [20]: text_to_sentiment("My name is Yvette")
Out[20]: 0.9846380213298556

In [21]: text_to_sentiment("My name is Shaniqua")
Out[21]: -0.47048131775890656
```

see https://blog.conceptnet.io/2017/07/13/how-to-make-a-racist-ai-without-really-trying/

see also Bolukbasi et al. (2016) *Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings*
Caliskan et al. (2017) *Semantics derived automatically from language corpora contain human-like biases*
Kiritchenko and Mohammad (2018) *Examining Gender and Race Bias in Two Hundred Sentiment Analysis Systems*

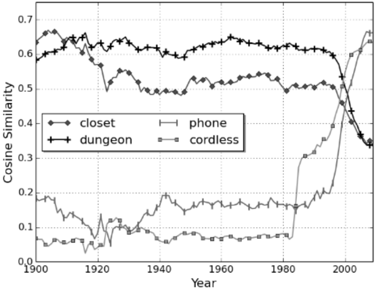
word embeddings in historical investigations (1)

- Garg et al. (2018) investigate gender and ethnic stereotypes over 100 years



word embeddings in historical investigations (2)

- Kim et al. (2014) (and many followers) use word embeddings to investigate semantic shifts over time
- for instance, the following example shows the similarity of ***cell*** to some query words:



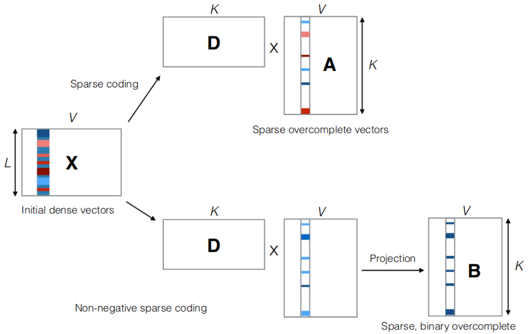
- see also http://languagechange.org

interpretability

- it's hard to **interpret** the numbers in a word embedding

  2739 ("cucumber") → [0.7, -1.2, ..., -0.1]

- traditional lexical semantics (descriptions of word meaning) often use **features**
- a number of approaches have been proposed to convert word embeddings into a more feature-like representation
  - for instance, SPOWV (Faruqui et al., 2015) creates sparse binary vectors

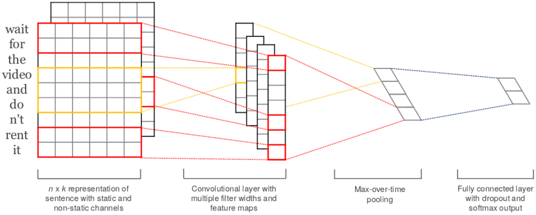


to read

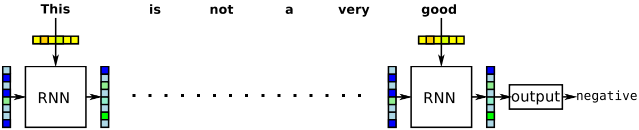
- Goldberg chapters 10 and 11
- evaluation survey: Schnabel et al. (2015)

what happens next?

- **convolutional models**



- **recurrent models**



references

T. Bolukbasi, K.-W. Chang, J. Zou, V. Saligrama, and A. Kalai. 2016. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In *NIPS*.

A. Caliskan, J. Bryson, and A. Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science* 356(6334):183–186.

M. Faruqui, Y. Tsvetkov, D. Yogatama, C. Dyer, and N. A. Smith. 2015. Sparse overcomplete word vector representations. In *ACL*.

N. Garg, L. Schiebinger, D. Jurafsky, and J. Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *PNAS* 115(16).

Y. Kim, Y.-I. Chiu, K. Hanaki, D. Hegde, and S. Petrov. 2014. Temporal analysis of language through neural language models. In *LT and CSS @ ACL*.

S. Kiritchenko and S. Mohammad. 2018. Examining gender and race bias in two hundred sentiment analysis systems. In **SEM*. pages 43–53.

T. Schnabel, I. Labutov, D. Mimno, and T. Joachims. 2015. Evaluation methods for unsupervised word embeddings. In *EMNLP*.