Word emebedding and other frontiers in text analysis

NTNU, Trondheim, 12.-13. August 2021

Gregory Ferguson-Cradler
Institutt for rettsvitenskap, filosofi og internasjonale studier
Høgskolen i Innlandet, Lillehammer

Word similarity and relatedness

1. Based on the ideas that similar words appear in the same context (both words that are synonyms and words that are simply clearly of the same kind, eg. Germany"and France"[1].
2. Based on the idea that word meaning can be represented in vector space (as we saw in document similarity) based on contexts in which words appear.
3. Documents made into vectors via DTM matrix. Words might be made into vectors via term-term matrix (fcm in Quanteda)
4. Two major algorithms for word embedding: word2vec and GloVe.

1. This is based on long and deep thought in linguistics, see (Jurafsky og Martin, under utgivelse) for a brief overview

What *are* word embeddings

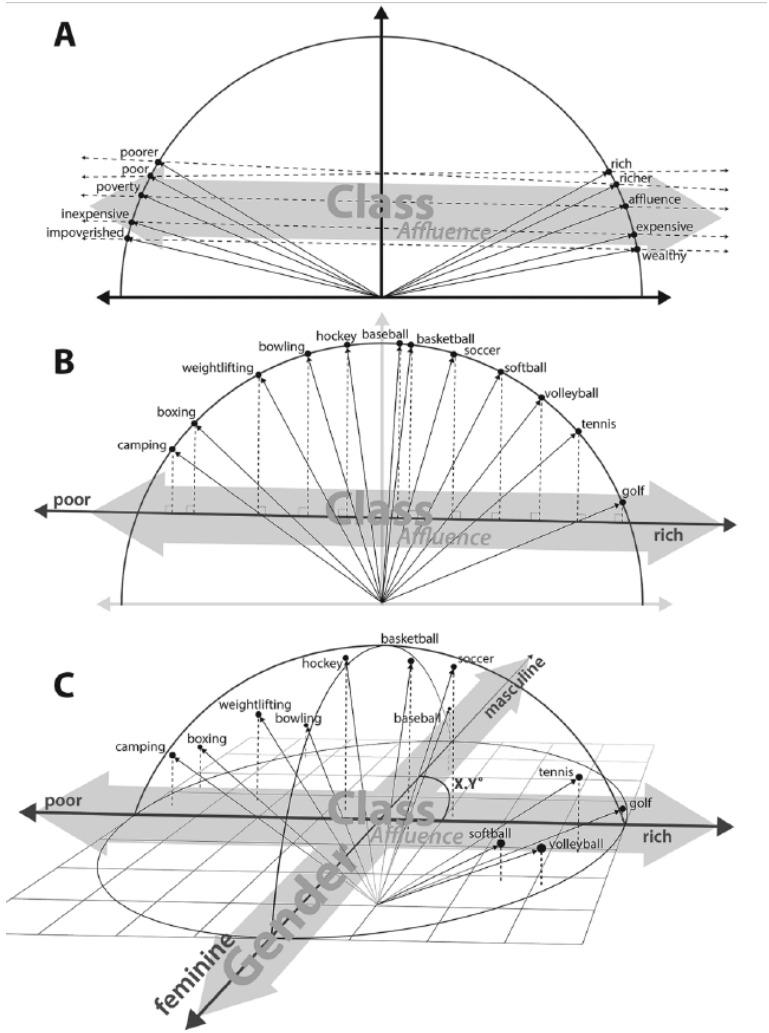
- Simplifying: these algorithms compute probability for word co-occurences (and non-co-occurrences) and construct word embeddings (vectors) that are similar when co-occurence probability is high and distant when probability is low.[2]
- Word embeddings so interesting (and somewhat baffling) because they show not just similarities between words but also have vector spaces that seem to correspond to meaningful concepts.
- $\overrightarrow{king} + \overrightarrow{woman} - \overrightarrow{man} \approx \overrightarrow{queen}$ analagous to just as a human would generally suggest 'queen' in answer to the question: man:woman as king:_____?.

2. Jurafsky og Martin (under utgivelse) is the best introduction to the details.

Document similarity methods in practice, I

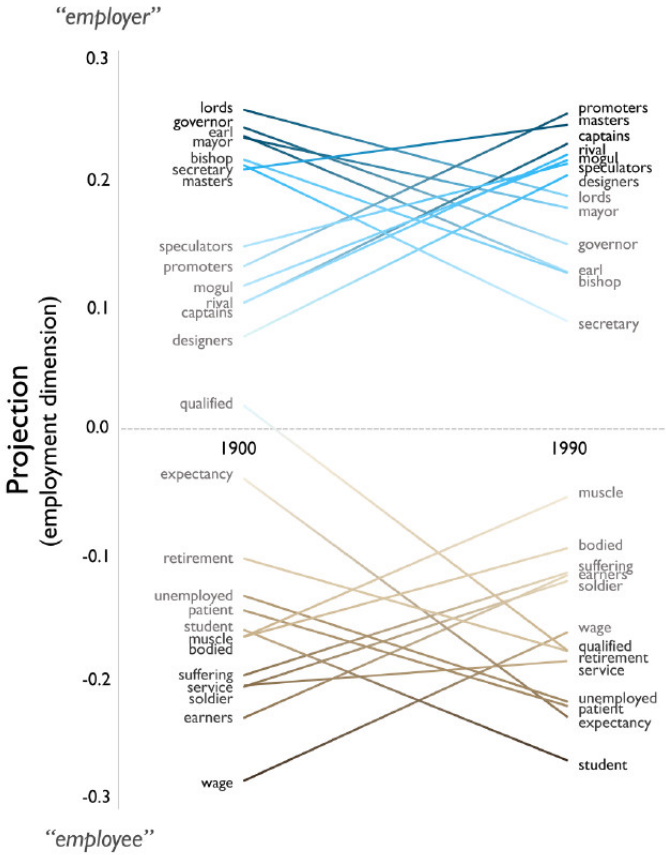
- Austin C Kozlowski, Matt Taddy og James A Evans. 2019. The geometry of culture: Analyzing the meanings of class through word embeddings. *American Sociological Review* 84 (5): 905–949
- Insight: we find dimensions in vector space that map to human meanings (eg, affluence, etc) by taking the average of pairs of words whose meanings diverge on this range (for affluence: affluence-poverty; rich-poor, prosperous-bankrupt, etc).
- Other words can then be projected"along this dimension to measure where they stand on the spectrum.

Kozlowski et al. 2019



Kozlowski, Taddy og Evans 2019

Document similarity methods in practice, II



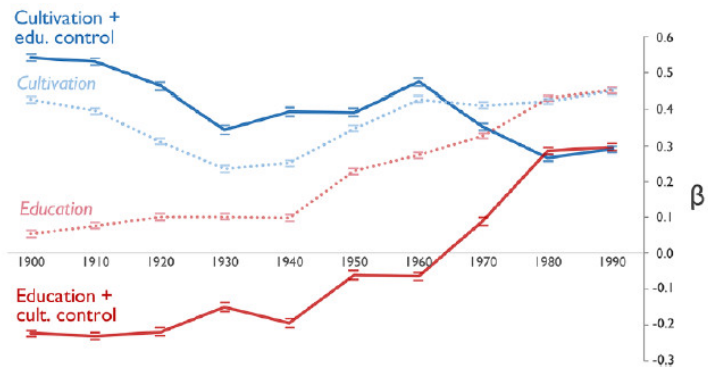


Figure 6. Standardized Coefficients from OLS Regression Models in Which Word Projections on Cultivation and Education Dimensions Predict Projection on the Affluence Dimension; 1900 to 1999 Google Ngrams Corpus
Note: A separate OLS regression model is fit for each decade; $N = 50{,}000$ most common words in each decade.

(Kozlowski, Taddy og Evans 2019, 928,924)

Contextualized word emebeddings and transformers

- Individual *token* embeddings
- Transformers weigh context word affect words of interest
- So far main use in social science seen for classification
- Unclear how might be taken up for interpretive purposes

Further resources: textbooks on R and text analysis

- Hadley Wickham og Garrett Grolemund. 2016. *R for data science: import, tidy, transform, visualize, and model data.* O'Reilly Media, Inc. https://r4ds.had.co.nz/
- Julia Silge og David Robinson. 2017. *Text mining with R: A tidy approach.* O'Reilly Media, Inc. https://www.tidytextmining.com/
- Jurafsky og Martin, under utgivelse
- Matthew L Jockers og Rosamond Thalken. 2020. *Text Analysis with R.* Springer

Further resources: online courses in programming and R

- Introduction to Computer Science and Programming: solid introduction to basics of programming (in Python but easily applicable to R)
- Data analysis for social scientists: basic quantitative methods in social science in R.
- Intro to Data Science: very basic course in R and data science.

Further resources: people in digital humanities/social science to follow

- Ben Schmidt
- Ted Underwood
- Julia Silge
- David Robinson
- Ken Benoit